ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE SISTEMAS

UNIDAD DE TITULACIÓN

TRANSFER AND ENSEMBLE LEARNING MODELS IN BREAST MAMMOGRAM PATHOLOGY CLASSIFICATION

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGÍSTER EN COMPUTACIÓN

LENIN GONZALO FALCONÍ ESTRADA

lenin.falconi@epn.edu.ec

Director: MARÍA GABRIELA PÉREZ HERNÁNDEZ

maria.perez@epn.edu.ec

Co-Director: WILBERT GEOVANNY AGUILAR CASTILLO

wilbert.aguilar@epn.edu.ec

2019

DIRECTOR'S APPROVAL

As director of the TRANSFER AND ENSEMBLE LEARNING MODELS IN BREAST MAMMOGRAM PATHOLOGY CLASSIFICATION thesis developed by Lenin G. Falconí, student of the Master's degree in Computing, I certify that I have supervised the completion of this work and made the corresponding corrections, and approve the final drafting of the written document to continue with the corresponding procedures to support oral defense.

María Gabriela Pérez Hernández
DIRECTOR

CO-DIRECTOR'S APPROVAL

As co-director of the TRANSFER AND ENSEMBLE LEARNING MODELS IN BREAST MAMMOGRAM PATHOLOGY CLASSIFICATION thesis developed by Lenin G. Falconí, student of the Master's degree in Computing, I certify that I have supervised the completion of this work and made the corresponding corrections, and approve the final drafting of the written document to continue with the corresponding procedures to support oral defense.

Wilbert Geovanny Aguilar Castillo
CO-DIRECTOR

AUTHORSHIP DECLARATION

I, Lenin Gonzalo Falconí Estrada, declare under oath that the work described here is my responsibility; that has not been previously submitted for any degree or professional qualification; and, that I have consulted the bibliographic references included in this document.

Escuela Politécnica Nacional (EPN) and the participating universities in the project REDU (PREDU-2016-013) may use the rights corresponding to this work, as established by the Intellectual Property Law, by its Regulations and by current institutional regulations.

Lenin Gonzalo Falconí Estrada

DEDICATED TO

Silvana A. and the the fabulous 14 gang.

ACKNOWLEDGMENT

A research work is a demanding task that is not possible without the aid of many different people that contribute both emotionally and intellectually to the author. I want to express my thankfulness to both María G. Pérez and Wilber G. Aguilar for guiding me in this process. Also I want to thank to Silvana Arévalo C. and her family that made the fulfillment of this step in my career as researcher possible despite the difficulties that total dedication to study imply. Finally, I want to thank to the Faculty of Systems at E.P.N for creating this new master of science and Dr. Jose F. Lucio Naranjo for facilitating additional computing resources at LASINAC. This research was carried out using the research computing facilities offered by Scientific Computing Laboratory of the Research Center on Mathematical Modeling: MODEMAT, Escuela Politécnica Nacional - Quito and the Red Nacional de Investigación y Educación Ecuatoriana CEDIA.

Contents

List of Figures

List of Tables

Abstract

Breast cancer is a global concern disease that specially affects women. Early detection of the disease, through mammography, increases life expectancy and reduces serious consequences of it. However, the sensitivity of mammography is variable; especially in dense breast. Therefore, it is important to develop cost effective tools that help to reduce the false positive and negative rates in the radiologist's diagnosis by providing a second opinion. Since year 2012, Deep Learning has achieved optimistic results in image classification, object detection, and image segmentation through convolutional neural networks. This master's thesis explores transfer learning, fine tuning, whole retrain, and ensemble learning techniques to classify mammogram masses as benign or malignant. For this purpose, 20 pre-trained convolutional neural networks on the ImageNet Dataset were compared. Furthermore, this work's proposed Automatic Soft Voting Ensemble (ASVE), which used a Perceptron to automatically tune the soft voting weights of the fine tuned convolutional neural networks ensemble, has achieved promising results on Inbreast ($AUC = 98.2\%$), MIAS ($AUC = 97.8\%$), and in a merged dataset of both ($AUC = 91.8\%$). This suggests that this work's approach is suitable to be implemented in a computer aided diagnostic tool.

Keywords: transfer learning, fine tuning, convolutional neural networks, ensemble learning, classification, digital mammography, breast cancer

Resumen

El cáncer de mama es una enfermedad que afecta especialmente a la mujer. La detección temprana de la enfermedad, a través de la mamografía, aumenta la esperanza de vida y reduce las graves consecuencias de la misma. Sin embargo, la sensitividad de la mamografía es variable; especialmente en senos densos. Por ello, es importante desarrollar herramientas rentables que ayuden al diagnóstico del radiólogo al proporcionar una segunda opinión que reduzca la tasa de falsos positivos y negativos. Desde el año 2012, Deep Learning ha logrado resultados optimistas en clasificación, segmentación de imágenes y detección de objetos, a través de redes neuronales convolucionales. Esta tesis explora *transfer learning*, *fine tuning*, re-entrenamiento completo y técnicas de *ensemble learning* para la clasificar las masas mamográficas como benignas o malignas. Un total de 20 redes pre-entrenadas en el conjunto de datos ImageNet fueron comparadas. El modelo propuesto por esta tesis de Ensamble Automático de Voto Ponderado ha logrado resultados prometedores en Inbreast ($AUC = 98.2\%$), MIAS ($AUC = 97.8\%$) y en un conjunto de datos combinado de ambas ($AUC = 91.8\%$). Esto sugiere que el enfoque propuesto es adecuado para implementarse en una herramienta de diagnóstico asistida por computadora.

Palabras clave: transfer learning, fine tuning, redes neuronales convolucionales, ensemble learning, clasificación, mamografía, cáncer de mama

Chapter 1

Introduction

1.1 Background

Cancer is one of the noncommunicable diseases affecting humankind (NCDs[1]). This term refers to a large group of diseases that can affect any part of the body. The main feature of this disease is the rapid creation of abnormal cells that grow beyond their usual boundaries, invading adjoined parts of the body and spreading to other organs. The latter mentioned process is referred as metastasizing which is a major cause of death from cancer (Reboux, 2018).

According to B. W. Stewart, Wild, International Agency for Research on Cancer, and The World Health Organization (2014), it is predicted that the cancer burden will exceed the 20 million new cases by 2025. In fact, in the “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”, Bray et al. (2018a) summarize the estimated numbers of new cases and deaths between 36 types of cancer. The two most deadly types found are: lung (2,093,876 new cases and 1,761,007 deaths) and breast (2,088,849 new cases and 626,679 deaths) (see Figure 1.1). The latter is most commonly diagnosed in women, and also constitutes their main cause of death (Bray et al., 2018a) (see Figure 1.2). In the opinion of Lukong (2017), even after 50 years of advances in medical research, breast cancer is still a major public health problem worldwide.

[1] Noncommunicable diseases are also known as chronic diseases because of their long duration

In order to reduce mortality, early detection of breast cancer has proven to be of most importance. Screening methods like mammography increase significantly life expectancy and reduce serious consequences of the disease (B. Stewart, Wild, et al., 2019). Because of these reasons, it is of interest for science, as a mean to increase the quality of life of humankind, to develop technologies that encourage screening methods and even automate the prediction of pathologies.

Technological advances have been decisive to lower the death toll risk in breast cancer. In fact, as mentioned by Stoitsis et al. (2006), advances in image technology and computer science have contributed in an enhanced interpretation of medical images. Thus, the discovery of the X-Rays by Röentgen (1896) set the technological basis to obtain the mammogram image. Years later, in the 1960s, this technology was further explored by Robert Egan who published a work titled: “Experience with mammography in a tumor institution: Evaluation of 1000 cases”, where he set the procedures to conduct a mammogram (Kalaf, 2014; Lukong, 2017). In fact, Egan was able to report 53 cases of occult breast cancer by using mammography images (Kalaf, 2014). Today, high quality mammography along with clinical examinations remain the gold standard in breast cancer screening (Jalalian et al., 2017; Thomassin-Naggara, Tardivon, & Chopier, 2014) because mammography aids to detect cancers before they are clinically evident and in women who have no symptoms (see Baum & Henderson, 2004, Chapter 7). Moreover, mammography is able to detect small tumors, which are hard to detect by manual exam, and deposits of calcium that might indicate the presence of cancer.

A weakness in mammography is the variability of the exam’s sensitivity specially in dense breast. In such case, the contrast between cancer and background is low and affects the diagnosis outcome. Furthermore, it has been shown that mammography sensitivity is inversely proportional to breast density (Drukteinis, Mooney, Flowers, & Gatenby, 2013). In addition, Jalalian et al. (2017) report that radiologists fail to detect between 10% to 30% of breast cancers. This means that mammography can yet be improved to lower false positives (non cancerous lesions wrongly identified as malignant) and false negatives (missed malignant lesions). One way to aid radiologists in detecting abnormalities in mammogram images is through the development of Computer Aided Diagnostic (CADx) and Detection (CADe) software tools (Dromain et al., 2013).

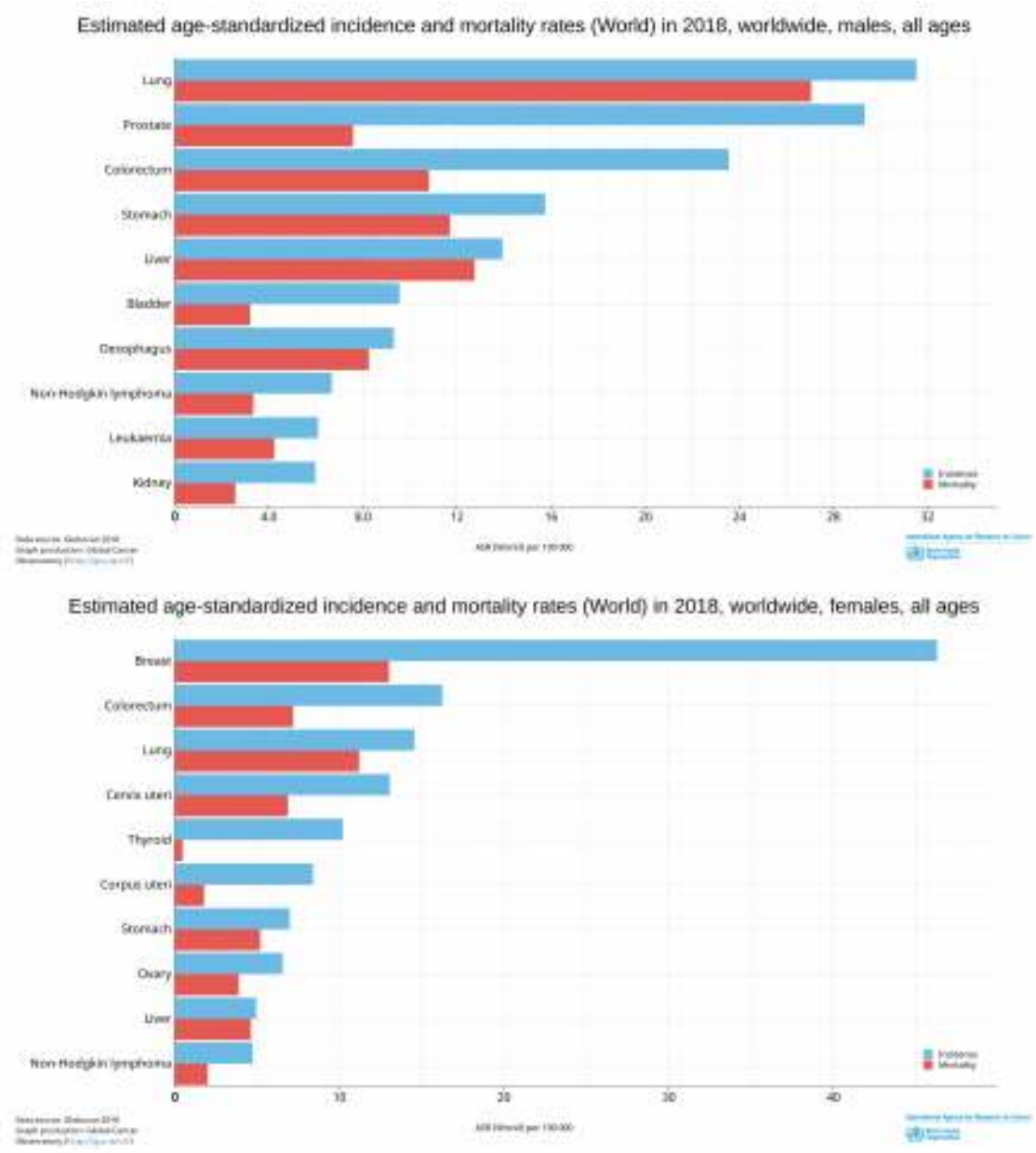

Figure 1.1: Estimated age-standardized incidence and mortality rates worlwide 2018.

SOURCE: Bray et al., 2018b

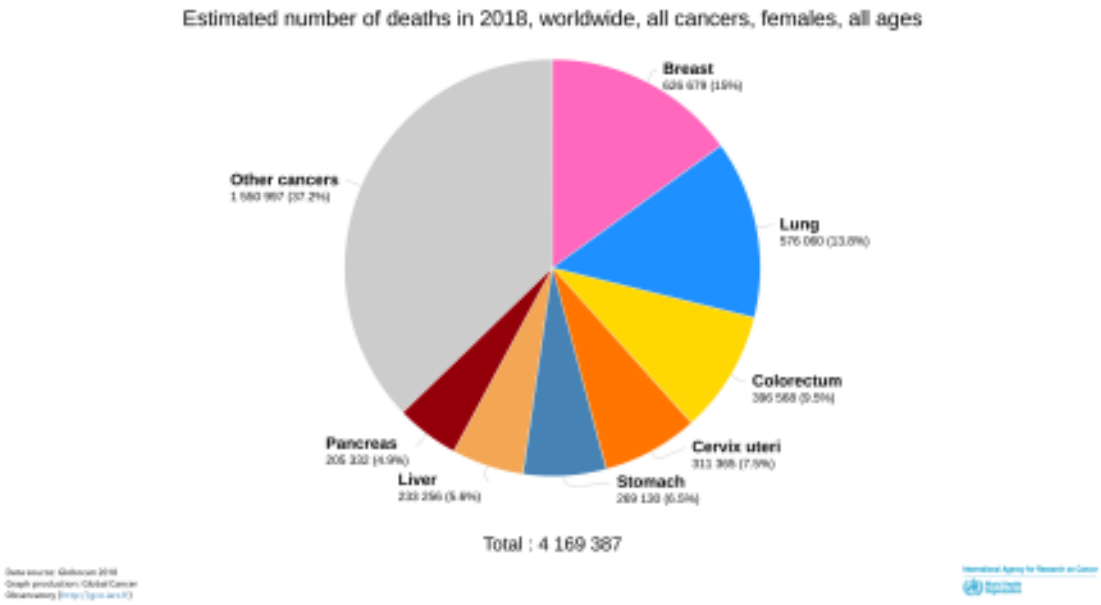

Figure 1.2: Number of deaths in 2018 for female population.
SOURCE: Bray et al., 2018b

A CAD system uses computer vision techniques together with artificial intelligence and machine learning algorithms to extract useful information from a digital image. This information is processed to ultimately make useful decisions; which agrees with the goal of computer vision (Shapiro, 1992). Common problems studied by computer vision, artificial intelligence, and machine learning are image classification, object detection and image segmentation. In this thesis, the classification problem of mammogram images according to its pathology is addressed.

Image classification consists in assigning a label to an image or picture automatically by a machine or computer program (Rasche, 2019). This task is very easy for most humans because our visual system can discriminate between thousands of categories and find objects in images accurately in a matter of milliseconds (Rasche, 2019). However, machines have a hard time solving this problem. Previous research has focused in the development of hand engineered feature descriptors that could be used with trainable classifiers. This approach is implemented in the work by Perez, Benalcazar, Tusa, Rivas, and Conci (2017); where Haralick's texture descriptors are used to characterize the mammogram image to later train an Artificial Neural Network (ANN).

Recently, thanks to the development of Convolutional Neural Networks (ConvNets), machines have even been able to surpass human's natural ability to classify natural images. ConvNets were firstly proposed in 1990s, when LeCun et al. (1990) used them for hand written digit recognition. Years later, in 2012, Krizhevsky, Sutskever, and Hinton (2012) developed the concept of *Deep Learning* through *Deep Convolutional Neural Networks* and tested it successfully in the classification of 1 000 categories of natural images. However, their approach relies heavily on the use of large amounts of data and regularization techniques to prevent overfitting [2]. This leads to a problem in mammogram image classification because public datasets of mammography are not large enough for Deep Learning. One way to deal with this problem is through *Transfer Learning*.

Today, numerous investigations focus on the application of *Deep Learning* models in mammogram classification because of its performance in natural image classification. Some of the works have designed specific ConvNet architectures and trained them from scratch (Guan & Loew, 2017), whereas others have implemented Transfer Learning (TL) or Fine Tuning (FT) (Jiang, Liu, Yu, & Xie, 2017) by using pre-trianed ConvNets. However, little attention has been payed to formally defining TL and FT to describe the training processed followed. The present study provides a formalization of TL and FT, and compares the performance of different state of the art pre-trained ConvNets tested on ImageNet (Deng et al., 2009) when fine tuned on mammography datasets to facilitate a benchmark for future works.

1.2 Research Focus and Objectives

Mass classification in mammograms is not a new topic in research but it is critically important because breast cancer is a public health problem. In fact, Dias Pedro, Machado-Lima, and Nunes (2018), in their systematic literature review (SLR), ask if mass classification in mammograms is already a solved problem. Their findings are very interesting. For instance, their SLR shows that Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are the two most common machine learning techniques used to address the problem. Little is discussed about Convolutional Neural Networks (ConvNets) because in their findings ConvNets are mostly used to form the feature vector for another trainable classifier.

[2] As a matter of fact, the ImageNet dataset has around 3.2 million images (Deng et al., 2009)

It is difficult to establish a concrete state of the art performance value due to the lack of uniformity in the metrics and datasets used (Dias Pedro et al., 2018, page 100). However, the studies reviewed by the authors show a classification performance above 90%. Moreover, just a few works validate their results with radiologists.

Considering the variety of metrics in the state of the art and the observations about binary classification made by Canbek, Sagiroglu, Temizel, and Baykal (2017), this thesis uses first and second level classification metrics. These metrics are: Accuracy (ACC), Balanced Accuracy ($BACC$), the Area Under the Receiving Operating Characteristics Curve (AUC), The F_1SCORE and the Mathews Correlation Coefficient (MCC).

Besides the different metrics used by authors in the state of the art and the lack of uniformity of the available datasets, it is important to notice that some authors rely on whole mammogram classification (Perez et al., 2017), whereas others classify region of interest (ROI) mammogram images (Chougrad, Zouaki, & Alheyane, 2018; Xiaoyong Zhang et al., 2017). This is a distinction absent in the SLR of Dias Pedro et al. (2018). This research focuses on ROI mammogram images.

Since 2012, different architectures of ConvNets have been designed to solve natural image classification on the ImageNet dataset. Some of these models are: VGG (Simonyan & Zisserman, 2015), Inception-v3 (Szegedy, Vanhoucke, Ioffe, & Shlens, 2016), Mobilenet (Howard et al., 2017a), NasNet (Zoph, Vasudevan, Shlens, & Le, 2018). Due to the fact that these models have achieved great results, similar to the human level, this thesis studies their performance in mammogram classification by using *Transfer Learning* (TL) and *Fine Tuning* (FT). In this work, a total of 20 different pre-trained convolutional neural network models are compared. Furthermore, this research provides a formal definition of TL and FT for the specific case of ConvNets.

Generalization is the main purpose of a classification system. The ability to address to new not previously seen data successfully is highly desirable for a model that will integrate a CAD system. Since most mammograms can be either of digi-

tal or film nature, a ConvNet trained in the former may have difficulties in predicting a mammogram image from the latter. Because of this, it is of importance to review how an ensemble of models could help to improve classifications.

1.2.1 Research question and Hypothesis

At present, the classification of natural images is being solved by the use of *Deep Learning* models. These models rely on special configurations of a basic convolutional neural network, forming convolutional cells, blocks or modules to increase performance without increasing the number of trainable parameters [3]. In all cases, their success depends on training on a large labeled dataset. This is a problem in medical imaging because manufacturing a labeled dataset can be expensive. However, the natural question that arises in this context is whether these models that have proven to classify 1 000 categories of natural images can classify mammogram images according to their pathology. If that is true, Is it possible to re-use the model using *Transfer Learning* to achieve an improved classification performance? Since there are some models already developed for natural images classification, Is there a specific pre-trained ConvNet that outperforms others in the task of mammogram classification? This thesis aims to solve these questions.

Increasing the depth of the ConvNet model has proven to improve classification results. For instance, Alexnet (Krizhevsky et al., 2012) is at most an 8 layer deep convolutional neural network whereas NASNet (Zoph et al., 2018) is around 1000 Layers. Resnet-152 is another example that is 8 times deeper than VGG (He, Zhang, Ren, & Sun, 2016). But, how deep must the pre-trained ConvNet be to address mammogram classification? Also, what happens if different pre-trained convolutional neural networks are fine tuned and configured to form an ensemble of predictors? This research hypothesizes that an ensemble of fine tuned pre-trained ConvNets will improve average classification accuracy. This is formally stated as follows:

[3] That is the case of GoogleNet (Szegedy et al., 2015) and NASNet (Zoph et al., 2018)

Hypothesis

If different pre-trained convolutional neural networks are fine-tuned on a mammogram dataset then, an ensemble of those models will improve the average performance of classification in the literature review and generalize for both film and digital mammograms.

The overall aim of this research is to find a pre-trained convolutional neural network model that can be re-trained through transfer learning or fine tuning to classify mammogram masses as either benign or malignant. This is stated as follows:

Research Goal

Design a classification model for breast lesions in mammograms using transfer and ensemble learning techniques to increase literature's average performance.

1.2.2 Specific goals of the Research

In order to achieve the main goal, the present research proposes the following specific goals.

1. Identify how transfer and ensemble learning are being used in breast mass abnormality classification by the research community.
2. Experiment with different state of the art pre-trained convolutional neural networks and measure their performance in classification of breast pathology in 2 classes: malign, and benign.
3. Propose a model that improves average breast pathologies classification from mammogram images.
4. Evaluate the performance of the proposed classification system

To full fill objective 1, it is necessary to get to know the state of the art of breast mass abnormality classification trough a literature review in the matter of the models used and configurations tested. The key contributions of this work are placed in objectives 2 and 3 because they are the experimental research part of this thesis.

1.3 Value of this Research

Breast pathologies are one of the main causes of death among women. Early detection of breast cancer has proven to be successful in lowering the death risk rate (Charan, Khan, & Khurshid, 2018). Different image modalities have been developed such as mammography, ultrasound, magnetic resonance, infrared thermography (Selvathi & Aarthypoornila, 2017; Yassin, Omran, El Houby, & Allam, 2018). However, the process for detection is done manually and it is related to radiologist's experience. Moreover, current practice has shown that 10% of all women screened for cancer are called back for additional testing and just as little as 0.5% of them are diagnosed with breast cancer (X. Zhang, Zhang, Han, et al., 2017). This background shows that it is important to design CAD systems to aid specialists and train new ones in breast lesions detection. In fact, in the last years, such systems have been developed to assist the specialist to have a second opinion for diagnostic. Mammographic images are mainly used in diagnostic because they have proved to be an effective tool to reduce breast cancer mortality and aid in visual detection of abnormalities (Jalalian et al., 2017).

However, detection and classification by a CAD system is a challenging problem that involves both machine learning and computer vision because it is difficult to find a unique representation of the characteristics of mammographic images. Deep learning through convolutional neural networks (ConvNet) architecture helps to solve the problem of image feature representation because a ConvNet could be said to synthetize its own feature extractor (Bengio, 1997).

This thesis aims to experiment with ConvNets that have not previously been used in the state of the art. To achieve this specific objective, a literature review is carried out. Then, experimentation is carried out with different pre-trained neural networks on the ImageNet dataset adapting them to mammogram classification by using Transfer Learning and Fine Tuning. Whole Retrain of the network, with randomly initialized weights, is used as a control group. In Chapter 2, a formal definition of these three concepts, besides the literature review, is presented. Moreover, because there is not a common performance metric in the state of the art, as indicated by Dias Pedro et al. (2018), the present thesis evaluates each model using first and second level metrics as proposed by Canbek et al. (2017). Finally, since public datasets are of two classes (film and digital), this thesis uses an ensemble of fine tuned models to achieve a better generalization score in

mammogram classification. Consequently, this work has scientific relevance because it provides a systematic experimental comparison of 20 models in different training modalities for mammogram classification. Furthermore, successful results of this thesis are of interest to build cost effective tools that aid the physician as a second opinion device for early diagnose of the pathology.

This work's approach involves medical image processing, which is related to computational perception, and a classification model, which is related to artificial intelligence and machine learning research lines in Escuela Politécnica Nacional (EPN). By improving medical image pre-processing and classification, this thesis is also improving the development of accurate CAD systems for medical application in breast pathologies detection.

The remaining of this thesis is organized as follows:

Chapter 1: Introduction
This chapter provides the reader with background information about breast cancer and CAD systems that aid radiologists in early detection of the disease by processing the mammogram image. Deep Learning (2012) is introduced as an effective approach for image classification in the natural domain and Transfer Learning is discussed as a means to adapt pre-trained models.

Chapter 2: Theoretical Framework
This chapter has two main parts. First part provides a deep learning background to familiarize the reader with concepts such as convolutional neural networks (ConvNets) and thus form a theoretical base to propose the formal definition of transfer learning and fine tuning presented in this thesis. The second part consists of a literature review and discussion of related works found in scientific databases.

Chapter3: Research Methods
This chapter illustrates the research approach and methodology used in the present study to address the research problem. The problem formulation, research strategy and experimental design are presented here to describe all the steps carried out in this research. Moreover, limitations and potential problems are discussed.

Chapter4: Results

This chapter presents the results obtained through the application of the experimental research design. All experimental evidence is provided through comparative Tables and Figures for transfer learning, fine tuning, whole retrain modalities, and ensemble methods on both datasets (Inbreast, MIAS). Finally a comparison of this thesis results wrt. literature is provided

Chapter 5: Conclusion

This chapter revisits the overall aim and specific objectives of this research providing a summary of our main findings and discussing the results.

Chapter 6: References

This chapter contains an alphabetical listing of the sources referred in this work. The APA (author - year) system is used.

Chapter 2

Theoretical Framework and Literature Review

2.1 Introduction

The main objective of this research is to classify region of interest images from mammograms as either benign or malignant by using Convolutional Neural Networks (ConvNets). Traditional CAD systems use a feature extraction stage where a pattern is extracted or handcrafted to classify the image. On the other hand, ConvNets automatically find a feature vector by learning from data through a training process. However, ConvNets require large labeled datasets. On account of the limited size of mammogram datasets, this research focuses on *Transfer Learning (TL)* to improve classification results because it is commonly used when there is a limited supply of training data (Weiss, Khoshgoftaar, & Wang, 2016).

This Literature Review discusses how Transfer Learning, Fine Tuning and Ensemble Learning are being used in mammogram pathology classification. In addition, it provides a basic theoretical background of Deep Learning and a formal definition of the concepts: *Transfer Learning* (TL), *Fine Tuning*(FT) and *Whole Retrain* (WR) as they are used in this Thesis. Furthermore, the proposed formal definition of the aforementioned concepts adapts the traditional definition found in literature to the the specific application on ConvNets and Computer Vision.

At the end of this chapter, it is expected that a critical understanding of TL and its application in mammogram classification is exhibited. Also, it is expected to provide enough evidence to justify the need for empirical research in this thesis.

2.2 Deep Learning Background

This section reviews the concept of Deep Learning (DL) that is a new field of research in Machine Learning (ML) and Pattern Recognition. Specifically, the main characteristics of Convolutional Neural Networks (ConvNets) are presented to provide the reader with a technical background to evaluate this research work.

When reviewing DL from an academic perspective, there are different definitions of the concept in literature and on the internet. It certainly started as a new research trend that has proved to be successful in image and speech recognition tasks (Bengio, 1997). In 2012, Krizhevsky et al. (2012) used Deep Learning in image classification, outperforming other techniques. These advances have been possible due to the development of Graphical Processing Units (GPU) that reduce the time invested in training ConvNets (LeCun, Bengio, & Hinton, 2015). These reasons have turned DL into a concept that is understood by what it does through its applications.

Research in Artificial Neural Networks (ANN) dates back to the 1940s and 50s (McCulloch & Pitts, 1943; Rosenblatt, 1957). Pattanayak (2017) defines DL as the evolution of ANN whereas Guo et al. (2016) define it as a sub-field of ML that attempts to learn high level abstractions from data by using hierarchical architectures. Today, DL is said to be a sub-field of Representation Learning [1] that aims to automatically find features to learn complex functions by using a general learning procedure (Goodfellow, Bengio, & Courville, 2016; LeCun et al., 2015). Consequently, the following definition is proposed:

Definition 2.2.1. Deep Learning is a Representation Learning method that uses a general purpose learning procedure to adjust internal parameters (weights) in order to automatically discover features or representations from data to solve an artificial intelligence task.

[1] Representation Learning is considered field in itself according to Yoshua Bengio, Courville, and Vincent (2013). The reader can find additional information in the aforementioned paper

2.2.1 Convolutional Neural Network Structure

A basic ConvNet is a form of Neural Network organized hierarchically to process data in a series of stages. Each stage contains layers that perform operations like the convolution of data with a Kernel (or filter) and its pooling (LeCun et al., 2015).

Figure 2.1 depicts a simple convolutional neural network used in digit recognition. A sample image of the number 5 with a size of 28×28 is convolved with a bank of 10 kernels with a size of 24×24. After that, the pooling operation is executed. This is repeated until reaching a Full Connecting Layer (FC) with 100 units (or neurons). At the end of the architecture is the output layer with 10 neurons for each digit class. In fact, the two FC Layers at the end of the model are a common ANN. Although there are variations to this basic set up, the three basic types of layers found in a ConvNet are the convolutional layer, the pooling layer and the full connecting layer.

The architecture portrayed in Figure 2.1 has evolved since it was proposed in 1998 (LeCun, Bottou, Bengio, Haffner, et al., 1998). Some improvements have been achieved and new layers have been proposed like: the ReLU layer, Batch Normalization. Even the concept of the convolution has been revisited in Mobilenets to create models that are more friendly with mobile and embedded devices. However, it is not the purpose of this research to study the convolutional architecture itself but to understand it because all the models used rely on it. The main types of layers of a ConvNet are described next. If needed, the reader will be pointed to additional information sources.

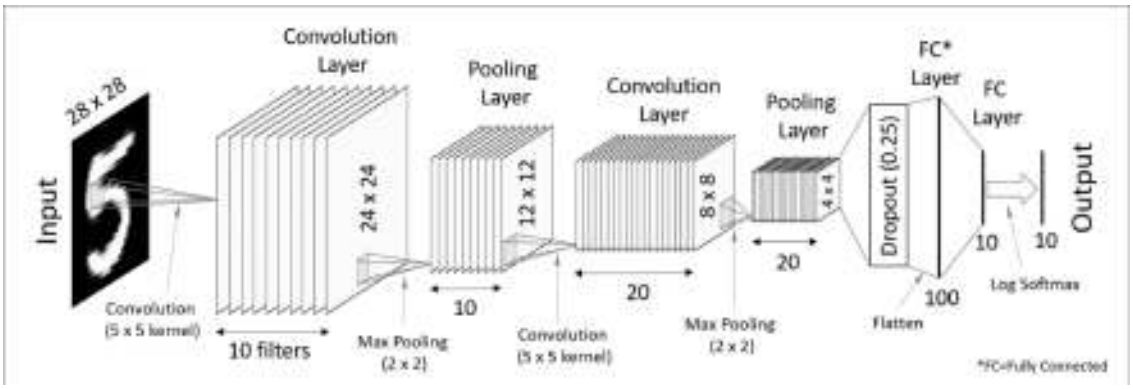

Figure 2.1: Convolutional Neural Network

A convolutional neural network to classify handwritten digits.Source: (Theart, 2017)

Convolutional Layer

The convolution operation of two real value functions $x(t)$ and $v(t)$, denoted by $x * w$, is defined by

$$z(t) = (x * v)(t) = \int_{-\infty}^{+\infty} x(a)v(t-a)da \tag{2.1}$$

In the context of ConvNets, x will be renamed as **input** whereas v will be known as **kernel**. The result $z(t)$ is known as the **feature map**. For discrete data, (2.1) is rewritten as follows:

$$z(t) = (x * v)(t) = \sum_{-\infty}^{+\infty} x(a)v(t-a) \tag{2.2}$$

The convolution operation defined in (2.1) and (2.2) consists in sliding a filter w over the input x. However, when dealing with monochromatic images, it is required to expand the convolution operation by using a two dimensional kernel. Because images are a form of multi-array data, the convolution operation for a 2D-array I of size $w \times h$ (i.e. $I \in \mathbb{R}^{w \times h}$), with a Kernel K of size f (i.e. $K \in \mathbb{R}^{f \times f}$) is defined in (2.3); where i and j are the coordinates of each new pixel of the feature map. If the Image or mathematical object has more dimensions, the convolution operation is extended accordingly. Hence, the use of tensors in Deep Learning.

$$(I * K)(i,j) = \sum_{w} \sum_{h} I(w,h)K(i-w, j-h) \tag{2.3}$$

In order to automatically learn representations from data, the convolutional layer applies the convolution operation between a given input and a bank of kernels of size f whose values are randomly initialized (Khan, Rahmani, Shah, & Bennamoun, 2018). The weights of each filter are learned during the training of the ConvNet. The purpose of this design is to detect local conjunctions of features from a previous layer (LeCun et al., 2015).

The convolutional layer uses four hyperparameters to produce the Activation Maps (AM) (or Feature Maps) $\mathbf{X}^{(L)}$ at a layer L, over the operation of convolution with the bank of filters. These hyperparameters are:

- Depth of Filters (k): the number of filters to be used to learn features.
- Size of Filter (f): each filter has a size of $f \times f$ to convolve with the previous layer.

- Stride (s): it is the amount of space that the kernel skips to apply the convolution between the input layer and the filters bank. This is a sliding operation.
- Zero Padding (p): it consists in fill with zeros around the border of the input in order to preserve the spatial size of the input.

Thus, the convolutional layer operator is defined in (2.4), which receives an input tensor (feature map) $\mathbf{X}^{(L-1)}$ and produces an output tensor ($\mathbf{X}^{(L)}$).

$$\mathbf{X}^{(L)} = Conv_{2D}(\mathbf{X}^{(L-1)}, f, k, s, p) \tag{2.4}$$

The dimensions of the resulting volume of Activation Maps are: $w^{(L)} \times h^{(L)} \times d^{(L)}$. This means that the Activation Maps at layer L are tensors of order 3: $\mathbf{X}^{(L)} \in \mathbb{R}^{w^{(L)} \times h^{(L)} \times d^{(L)}}$.

The relation between the size of the input layer tensor $\mathbf{X}^{(L-1)}$ at $L-1$ and the output tensor $\mathbf{X}^{(L)}$ at layer L is described in (2.5) (Murphy, 2016; O'Shea & Nash, 2015; Wu, 2017).

$$\begin{aligned} w^{(L)} &= \frac{w^{(L-1)} - f + 2p}{s} + 1 \\ h^{(L)} &= \frac{h^{(L-1)} - f + 2p}{s} + 1 \\ d^{(L)} &= k \end{aligned} \tag{2.5}$$

For instance, in Figure 2.1, the input image is $w^{(1)} = h^{(1)} = 28$, $d^{(1)} = 1$. This image is operated through convolution with $k = 10$ filters of size $f = 5$. Because the dimensions of the Activation Map (AM) are not preserved at the next layer, $p = 0$. According to this, and assuming a stride of 1, the dimensions of the output volume of AM are expected to be:

$$w^{(2)} = \frac{28 - 5}{1} + 1 = 24$$

The convolution layer benefits in three main aspects:

- **Local connections:** for instance, in an image, local groups of values are highly correlated and thus become a local motif (LeCun et al., 2015). The convolution operation is a local operation, which means that all spatial locations share the same convolution kernel (Wu, 2017).
- **Reduction of parameters by Shared weights:** This benefits both the performance and the complexity of the neural network. If a feature is useful in

a set spatial region, then it is likely to be useful somewhere else(O'Shea & Nash, 2015).

- **Object Location invariance:** Since an object can appear in any part of the image, the use of units of filters with the same tuned weight values after training will detect the same pattern independently of location (Guo et al., 2016; LeCun et al., 2015).

Pooling Layers

The pool operation aims to perform a semantic merge of similar features by taking neighboring pixels into account (LeCun et al., 2015). Its main effect is the reduction of dimensionality of Feature Maps and parameters (Guo et al., 2016), because it reduces the maps of a subregion into a single number (Wu, 2017). The two most common form of Pooling operations used are *Average Pooling* and *Max Pooling*. According to LeCun et al. (2015), pooling has the effect of creating an invariance to small shifts and distortions.

Similarly to the convolutional layer, this layer uses two hyperparameters:

- Size of Pooled area (f)
- Stride (s)

Like the convolution operation, Pooling is also of the local type. For instance, a max pooling operation extracts the maximum value of the region. Let $\mathbf{X}^{(L-1)}$ be an input tensor at layer $L-1$, where $\mathbf{X}^{(L-1)} \in \mathbb{R}^{w^{(L-1)} \times h^{(L-1)} \times d^{(L-1)}}$, and $\mathbf{X}^{(L)}$ the tensor obtained by Pooling layer with $\mathbf{X}^{(L)} \in \mathbb{R}^{w^{(L)} \times h^{(L)} \times d^{(L)}}$, then it can be defined Max Pooling as indicated in (2.6). Similarly, the relation of the dimensions of the input and output tensors are defined in (2.7):

$$\mathbf{X}^{(L)} = Max\_Pool(\mathbf{X}^{(L-1)}, f, s) \tag{2.6}$$

$$\begin{aligned} w^{(L)} &= \frac{w^{(L-1)} - f}{s} + 1 \\ h^{(L)} &= \frac{h^{(L-1)} - f}{s} + 1 \\ d^{(L)} &= d^{(L-1)} \end{aligned} \tag{2.7}$$

For instance, in Figure 2.1, if the first convolution layer is defined as the input layer $L-1=2$, the dimensions of the output tensor considering $s=2$ and $f=2$ are:

$$w^{(3)} = \frac{24-2}{2} + 1 = 12$$

The max and average pooling operations for a $x_{i^{(L)},j^{(L)},d^{(L)}} \in \mathbf{X}^{(L)}$ are defined in (2.8) and (2.9), respectively.

$$x_{i^{(L)},j^{(L)},d^{(L)}} = max(x_{i^{(L-1)},j^{(L-1)},d^{(L-1)}}) \tag{2.8}$$

$$x_{i^{(L)},j^{(L)},d^{(L)}} = \frac{1}{f^2} \sum_{i^{(L-1)},j^{(L-1)},d^{(L-1)}} (x_{i^{(L-1)},j^{(L-1)},d^{(L-1)}}) \tag{2.9}$$

Where,

$$\begin{aligned} a_n \leq i^{(L-1)} < a_n + f, \qquad & a_n = ns, \qquad & 0 \leq a_n \leq w^{(L-1)} - s \\ b_n, \leq j^{(L-1)} < b_n + f, \qquad & b_n = ns, \qquad & 0 \leq b_n \leq h^{(L-1)} - s \end{aligned}$$

The arithmetic progressions a_n and b_n, start with 0 value and increase their value according to the stride s. They define the start point of the sub-region to be pooled. The size of the region is clearly defined by f. Notice that (2.9) is written in short, since it goes for 3 dimensions w, h and d.

Fully connected layers (FC)

A fully connected layer is a convolutional layer with a filter of size $f=1$. This produces a flatten effect of the previous layer reducing it to a number d of neurons. This means that each unit or neuron in this layer is densely connected to all the units in the two adjacent layers (Khan et al., 2018; O'Shea & Nash, 2015); just like a hidden layer in a Multi-Layer Perceptron. The purpose of this Layer is to map the automatically extracted features to the desired outputs. However, it can also be used as a feature vector representation. Usually this layer is located at the end of the architecture.

As a matter of fact, a FC can be converted to a convolutional layer and vice versa. Because of this fact, it is possible to use the ConvNet not only for classification

and regression problems but to output a high dimensional structured object. For instance, an application of this characteristic is the pixel wise labeling of images that draw a precise mask for an object (image segmentation) (Goodfellow et al., 2016).

In Figure 2.1, there are two FC at the end of the design. The first one has 100 neurons that are densely linked to the previous and next layers. In order to control overfitting, the dropout regularization is used. Finally at the end, there is a FC with just 10 neurons for classification.

Loss Layer

In a supervised learning procedure, where there is labeled data, an objective function to evaluate the error of the predictions made by the ConvNet on the training data respect to the ground truth is required. This function is used during the training process by the loss layer. The type of loss function to use depends on the problem to solve. According to Khan et al. (2018), there are the following main loss functions:

- Binary Classification: Hinge Loss, Binary Cross-Entropy
- Multi Class Classification: Categorical cross entropy, Expectation loss
- Identity Verification: Contrastive Loss.
- Regression: Euclidean loss, $\mathcal{L}^1$ error, Structural Similarity Measure (SSIM).

On account of the binary classification (benign, malignant) of the mammogram images proposed in this work, the loss function to be used is the *binary cross entropy*. In a supervised learning procedure, where $\mathcal{X}$ is the input space containing N samples of features vectors, such that $\mathbf{x} \in \mathcal{X}$, and $\mathcal{Y}$ being the output space, formed with the labels y, the probability for a sample i to belong to a label y is determined by $P(y_i|\mathbf{x}_i)$. Therefore, the binary cross entropy is defined by

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-y_i)) \tag{2.10}$$

2.2.2 ConvNet Learning

Li, Soltanolkotabi, and Oymak (2019) affirm that modern neural networks are typically over-parameterized. This means that the parameters to train far exceed the size of the training data, leading easily to overfitting. When a ConvNet architecture overfits, it happens that the ConvNet has memorized the training data and is unable to generalize and perform well on unseen data (test set). In order to control overfitting, basically there are two main options[2]:

- Use of Regularization Techniques: Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), early stopping, $\mathcal{L}^1$, $\mathcal{L}^2$, etc.
- Use data augmentation: Artificially create additional training examples, boosting the size of the training set in order to prevent the memorization of the training set by the model (Cadène, Thome, & Cord, 2016).

Dropout was proposed by Srivastava et al. (2014) and consists in reducing the co-adaptation between the neurons by randomly disconnecting a percentage of them. This prevents overfitting and aids in generalization. Early stopping is another regularization procedure aimed to prevent overfitting. Its main idea consists in evaluate the performance of the ConvNet on a held-out validation set at different iterations over a performance metric (usually the cross entropy loss).When the generalization performance of the model decreases, training is stopped to avoid overfitting. In fact, if training is not interrupted, the model overfits inevitably after many iterations (Li et al., 2019). $\mathcal{L}^1$ and $\mathcal{L}^2$ are regularization methods that constrain the values of the weights in the network.

Previously, it has been pointed that Deep Learning automatically learns features from data by a general learning procedure. At the heart of Deep Learning are the ConvNets which are over parameterized models that try to map the input space $\mathcal{X}$ to the output space (or ground truth or label space) $\mathcal{Y}$. The learning process of a ConvNet consists in tuning the parameters (θ) of the network. Khan et al. (2018) defines the training process of a ConvNet as the optimization of its parameters such that a loss function is minimized. Gradient descent based methods are used to tune these parameters by considering the minimization of the loss function. However, these methods have some problems like the vanishing

[2]Khan et al. (2018) present a better classification of regularization approaches that is recommended to be reviewed by the reader.

and exploding gradient. Several algorithms have been proposed to optimize the learning procedure through gradient descent like: RMSprop (Tieleman & Hinton, 2012), Adagrad (Duchi, Hazan, & Singer, 2011), Adadelta (Zeiler, 2012), Adam (Kingma & Ba, 2014), etc. The update of the parameters θ of the ConvNet's objective function $\mathcal{F}$ with a learning rate η wrt. the iteration time t is defined in

$$\theta_t = \theta_{t-1} - \eta \nabla_\theta \mathcal{F}(\theta_t) \tag{2.11}$$

2.2.3 ConvNets Architectures

The ImageNet Dataset (Deng et al., 2009) and the Large- Scale Visual Recognition Challenge (ILSVRC) have become a popular benchmark to test the performance of Deep Convolutional Neural Networks since 2010 (Russakovsky et al., 2015). Some of the tasks that have been evaluated in the challenge consist of: image classification (2010-2014), single object detection (2011-2014) and object detection (2013-2014). As a result of the challenge, several models have been developed. Consequently, year 2012 can be considered as a turning point in the challenge when ConvNets entered the scene and outperformed other architectures at image classification (Krizhevsky et al., 2012). As stated by Russakovsky et al. (2015), there is a steady reduction of error every year in object classification. It has been reduced from 28.2% to 6.7% (GoogleNet in 2014). Because these models are the base to be used in *Transfer Learning*, some of their architectures are reviewed briefly in this section.

Historically, the development of ConvNets dates back to 1990s when LeCun et al. (1998) proposed the LeNet-5 architecture to classify the handwritten digits. This was a ConvNet of 7 layers deep that worked on images of 32×32 size. Its structure is similar to that of Figure 2.1 with some differences: LeCun et al. (1998) uses a bank of $k = 6$ kernels of size $f = 5$, followed by subsampling (pooling). Next, another convolutional layer operation is carried out with $k = 16$, $f = 5$, which is again subsampled to obtain 16 feature maps of 5×5. Another convolutional layer is applied with $k = 120$ and $f = 5$. These feature maps are connected to a FC of 84 neurons and finally to the last FC of 10 neurons for each digit classification output. The second milestone is the work of Krizhevsky et al. (2012) where a thousand category classification of images was achieved by using an 8 layer deep ConvNet known as AlexNet. Also, this work introduced the use of dropout (Srivastava et al., 2014) and ReLu layer after each convolutional and full

connected layers. Different sizes of kernel are used: starting with $f = 11$ in the first $Conv_{2D}(I, 11, 96, 4)$, then using $f = 5$, and finally $f = 3$. AlexNet had around 60 million parameters to train.

VGG

The VGG Net architecture was proposed by Simonyan and Zisserman (2014a). The two most known models are the configurations D and E, commonly refered as Vgg16 and Vgg19. Their improvement relies in the use of small kernels ($f = 3$) that allow for deeper configurations without compromising the number of parameters and therefore the efficiency in training. As pointed out by Cadène et al. (2016), the VGG setup shows that multiple kernels of size 3×3 in sequence can emulate the effect of larger receptive fields. Another important point in Vgg16 design is the use of a ReLU layer after each convolution. A ReLU layer (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Nair & Hinton, 2010), defined by $g(z) = max\{0, z\}$, is used to increase the non linearity of the model and also contribute to aid the vanishing gradient problem. The Vgg16 has around 138 million parameters to train. If the previously introduced notation is used, Vgg16 can be described as a series of functions applied on tensors. This architecture is also one of the first to use blocks of Convolutional Layers; where the size of the tensors are withheld and changed only in the pooling layer. This is indicated by the symbol $\times$. For padding the value *same* is used. Hence, the Vgg16 architecture is shown as follows:

$$
\begin{aligned}
\mathcal{X}^{(1)} &= Conv_{2D}(\mathcal{I}, 3, 64, 1, same) \times 2 \\
\mathcal{X}^{(3)} &= Max\_Pool(\mathcal{X}^{(2)}, 2, 2) \\
\mathcal{X}^{(5)} &= Conv_{2D}(\mathcal{X}^{(3)}, 3, 128, 1, same) \times 2 \\
\mathcal{X}^{(6)} &= Max\_Pool(\mathcal{X}^{(5)}, 2, 2) \\
\mathcal{X}^{(9)} &= Conv_{2D}(\mathcal{X}^{(4)}, 3, 256, 1, same) \times 3 \\
\mathcal{X}^{(10)} &= Max\_Pool(\mathcal{X}^{(9)}, 2, 2) \\
\mathcal{X}^{(13)} &= Conv_{2D}(\mathcal{X}^{(6)}, 3, 512, 1, same) \times 3 \\
\mathcal{X}^{(14)} &= Max\_Pool(\mathcal{X}^{(13)}, 2, 2) \\
\mathcal{X}^{(15)} &= FC(\mathcal{X}^{(14)}, 4096) \times 2 \\
y &= SoftMax(\mathcal{X}^{(15)}, 1000)
\end{aligned}
$$

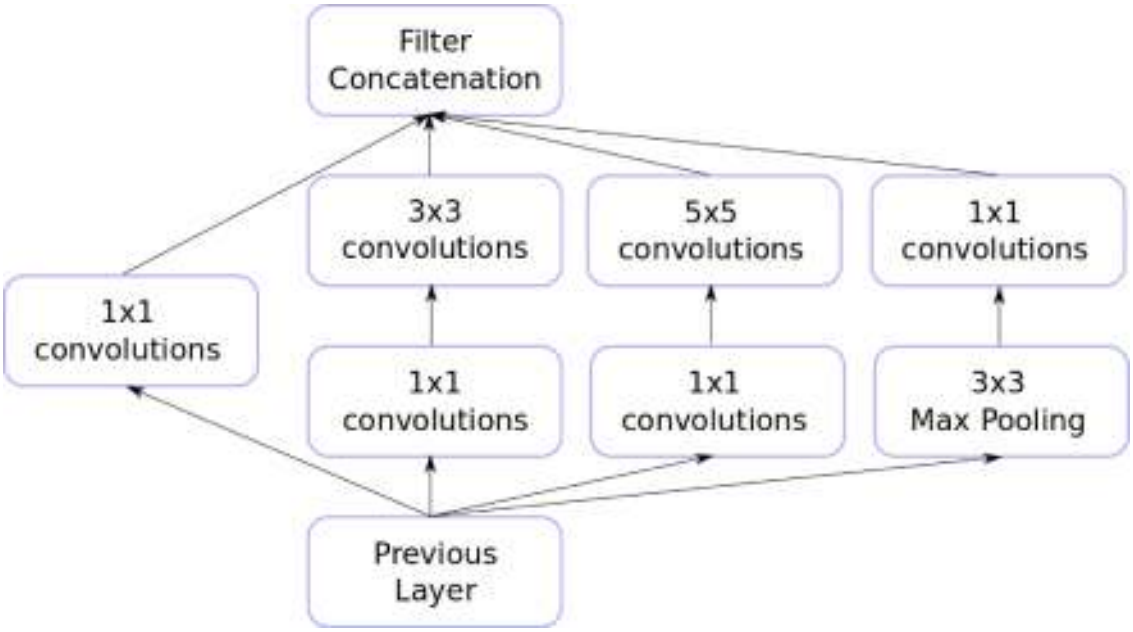

Figure 2.2: Inception Module

GoogleNet-Inception

This architecture was proposed by Szegedy et al. (2015). Their main contribution is the development of the inception module as indicated in Figure 2.2. Altogether with the use of average pooling instead of fully connected layers, the inception architecture permitted to reduce the number of trainable parameters. The inception module is characterized by its parallel design that is different from the sequential architecture proposed in previous models. Also, in this design, batch normalization layers (Ioffe & Szegedy, 2015) are used. The Inception-v3 uses 9 inception modules with a total of 100 layers with 23 million parameters.

ResNet and ResNext

The residual network from Microsoft (He et al., 2016) introduced what is known as the *residual block*, which permitted to train deeper networks (as deep as 152 layers) overcoming the problem of the accuracy saturation due to the inability to propagate gradients. The residual block uses shortcuts connections (or skip connections) where the gradient is passed directly to the deeper layers. In fact, the skip connection adds the outputs of the stacked layers to an identity mapping of the previous one or two layers. In the same fashion as GoogleNet, this network uses the residual blocks as building blocks stacked on top of each other. This design has reached 3.57% top 5 error on the ImageNet, winning the competition

in 2015. For the ResNet 50 model, a total of 25 million parameters is required to train. A model somehow based on both the Inception and Resnet architectures is the ResNeXt (Xie, Girshick, Dollár, Tu, & He, 2016), where the use of the parallel design (up to 32 branches) and the residual from a previous stage are combined. The residual block from ResNet is modified in ResNeXt to use considerably more branches than the inception design while containing the skip connections. Additional information can be consulted in the original paper (He et al., 2016) or in the Guide to Convolutional Neural Networks for Computer Vision by Khan et al. (2018).

DenseNet

Proposed by Huang, Liu, Van Der Maaten, and Weinberger (2017), this model propagates the output of each layer to all subsequent layers. In order to accomplish this, it is necessary to concatenate the feature maps from a layer with all the feature maps that preceded. This permits the pass of information without any loss of information. As stated by the authors of the model, the information is washed out through the processing of each convolutional and pooling layer, hence the need to use identity mappings to pass the original information. To improve model compactness, Huang et al. (2017) introduced transition layers; where a combination of 1×1 convolutions, batch normalization and a 2×2 average pooling is implemented. Also, the concatenated feature maps from preceding layers are not tunable. Only the global information from each stage is used. Because of that, each layer learns its own representation of the problem. Despite the dense architecture of DenseNet, the 121 layer version needs only 8 million parameters to train; which is an improvement compared to other models.

Mobilenet

Mobilenets were designed to be used in mobile and embedded vision applications by optimizing computation and model size (Howard et al., 2017b). The architecture of Mobilenets rely on *depthwise separable convolutions*. This operation splits the traditional convolution operation in two layers: depth wise convolution, pointwise convolution. The former carries the filtering operation applying a single filter to each input channel. The latter makes the combination by applying a 1×1 convolution to the outputs of the depthwise convolution. The advantage of this factorized convolution is that most of the computation is carried in the combination

stage. Because the pointwise uses a 1×1 kernel, the operation is implemented directly with General Matrix Multiply functions without reordering memory. Mobilenet is 28 layers deep with a total of 4.2 million parameters for version 1; being the smallest of all available models. Furthermore, the authors use two other hyper-parameters to create even smaller models. These are: width multiplier and resolution multiplier. The former, reduces the number of input channels by a factor α. The latter ρ is applied to the input image which reduces the internal representation of every layer.

Summary

In this section, some basic concepts of Deep Learning together with some pre-trained models that have been tested in the ImageNet contest have been reviewed. Deep Convolutional Neural Networks have become a milestone in computer vision by presenting a hierarchical architecture which is able to extract features automatically from a image. If the dataset is ”deep enough”, a ConvNet is more likely to meet excellent results. However, when training data is scarce, the models are not able to produce good results. The next section addresses the concept of transfer learning as it has been defined in scientific literature and how it is being reshaped by the advances in Deep Learning and computer vision.

2.3 Transfer Learning in Machine Learning: Concept Review

In *“Transfer learning for visual categorization: A survey”*, Shao, Zhu, and Li (2014) affirm that Transfer Learning (TL) in visual categorization tries to mimic the behavior of the human vision system where the previous knowledge gathered from experience and accumulated about a domain is used to learn about new tasks in a different but related domain. For instance, imagine about learning to recognize a new object that you have never seen before. All the knowledge that you already have comes in use and with some training you are able to categorize or distinguish this new object from others.

In literature, there are plenty of works about TL. However, the two most important about the definition of TL are the surveys by Weiss et al. (2016) and Pan and Yang (2010). Some important concepts from these works are reviewed in this

section. Also, their notation is used as a basis for the notation presented in this work.

The basic notion of transfer learning is that in order to perform a new *Task* (i.e. image classification, play and instrument, etc), the previous knowledge from another *Source Domain* is used to improve the performance in the new *Target Task*. In fact, the main purpose of TL is to create high performance learners in a target domain ($\mathcal{D}^T$) by training them with data from other domain known as the source domain ($\mathcal{D}^S$) (Pan & Yang, 2010; Weiss et al., 2016). Shanmugamani (2018) gives a descriptive definition of the concept TL by defining it as the process of learning from a pretrained model that was trained in a larger dataset. His definition is interesting in the fact that points that the source domain is bigger than the target domain.

Deep Learning is an inductive learning approach, because it tries to infer a mapping from a set of training examples that contain input features and class labels (Sarkar, Bali, & Ghosh, 2018). In traditional supervised machine learning, the training and testing samples are under the same distribution (Shao et al., 2014). A machine learning problem has both a domain and a learning task. The domain is formed by the feature space $\mathcal{X}$ and the marginal probability $P(X)$, where X is a sample of the feature space $\mathcal{X}$ formed by n feature vectors; thus $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathcal{X}$. The learning Task $\mathcal{T}$ is formed by the label space $\mathcal{Y}$ and the predictive function $\phi(\cdot)$ such that $y_i = \phi(\mathbf{x}_i)$ and $y_i \in \mathcal{Y}$. The predictive function can be rewritten in a probabilistic fashion as: $\phi(\mathbf{x}_i) = P(y_i|\mathbf{x}_i)$. The domain and task are expressed in (2.12) and (2.13) respectively.

$$\mathcal{D} = \{\mathcal{X}, P(X)\} \tag{2.12}$$

$$\mathcal{T} = \{\mathcal{Y}, \phi(\cdot)\} \tag{2.13}$$

Notice that a domain $\mathcal{D}$ is formed by tuples $(\mathbf{x}_i, y_i)$, such that: $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Therefore, the definition of *Transfer Learning* as given by Pan and Yang (2010) can be stated as follows:

Definition 2.3.1. Given a source domain $\mathcal{D}^S$ and a learning task $\mathcal{T}^S$, a target domain $\mathcal{D}^T$ and learning task $\mathcal{T}^T$, transfer learning aims to help improve the learning of the target predictive function $\phi^{\mathsf{T}}(\cdot)$ in $\mathcal{D}^T$ using the knowledge in $\mathcal{D}^S$ and $\mathcal{T}^S$, where $\mathcal{D}^S \neq \mathcal{D}^T$ or $\mathcal{T}^S \neq \mathcal{T}^T$.

In literature, there is a current discussion about the types of transfer learning and its categorization. In this work, the proposal presented by Shao et al. (2014), where for computer vision problems, transfer learning techniques could be categorized in the feature representation category, is adopted. Because, as the author points out, computer vision tasks have some particular characteristics due to the visual properties of objects (edges, textures, shapes, etc) that make them different from data mining tasks.

According to Pan and Yang (2010), when using transfer learning, it is important to define: *what to transfer*, *when to transfer* and *how to transfer*. In this work, by using pre-trained convolutional neural networks, what is desired to transfer are some relevant image features that could be reused to find patterns that aid in mammogram classification. It must be noticed that image problems are different from data mining applications. In fact, a radiologist has already learned to find mammogram abnormalities by having previously seen a lot of natural images. Therefore, despite the difference and the lack of similitude between a mammogram image and a natural image, it may be possible to transfer general descriptors to aid in the classification problem. For instance, in the article *How transferable are features in deep neural networks* the authors show that the first layers in a ConvNet have general features that are not specific to the problem at hand, but as the network turns deep, it specializes. Also, they show that generalization performance is improved when weight initialization occurs with transferred features rather than random initialization. Regarding the case of when to transfer, it can be stated that mammogram classification is suitable for transfer learning since there are not deep enough datasets and according to Weiss et al. (2016) TL is used when there is a limited supply of target training data. Finally, the how to transfer is addressed in the Experimental setup of this work. In the next section a formal definition of TL in the context of convolutional neural networks and computer vision is presented.

2.4 Transfer Learning in Convolutional Neural Networks

In the previous section, the definition of TL and the notation commonly used was introduced. In this section, the TL concept proposed by Pan and Yang (2010) is adapted to its application in convolutional neural networks, which are widely used in computer vision tasks.

Consider a ConvNet as a black box with L layers that given an image $\mathcal{I}$ as input is able to predict a class label y by its predicting function $\phi(\cdot)$. This is represented in

$$\mathcal{Y} = \phi(\mathcal{I}) \tag{2.14}$$

In this fashion, the target ConvNet and the pre-trained ConvNet are described in (2.15) and (2.16), respectively.

$$\mathcal{Y}^S = \phi^{\mathsf{S}}(\mathcal{I}) \tag{2.15}$$

$$\mathcal{Y}^T = \phi^{\mathsf{T}}(\mathcal{I}) \tag{2.16}$$

Now, the definition of transfer learning for the case of convolutional neural networks and computer vision is adapted as follows:

Definition 2.4.1. In the context of convolutional neural networks, Transfer Learning $\mathbb{T}_{\mathbb{L}}$ is the process that improves the learning of a target predictive function $\mathcal{Y}^T = \phi^{\mathsf{T}}(\mathcal{I}^T)$ in a target domain $\mathcal{D}^T$ wrt. a target task $\mathcal{T}^T$ by re-using a pre-trained ConvNet $\mathcal{Y}^S = \phi^{\mathsf{S}}(\mathcal{I}^S)$ of L layers previously trained on a larger dataset from a given source domain $\mathcal{D}^S$ and task $\mathcal{T}^S$ as described in (2.17); where $\mathcal{A}$ represents the transfer learning process over the source ConvNet of L layers.

$$\mathbb{T}_{\mathbb{L}} \left\langle \phi^{\mathsf{S}}(\mathcal{I}^S), \mathcal{I}^T \right\rangle = \mathcal{A} \left(\phi^{\mathsf{S}}(\mathcal{I}^T) \Big|_0^L \right) \tag{2.17}$$

The definition previously given states that TL trains a new model (or target model) by using the pre-trained source model, as indicated in (2.18).

$$\mathcal{Y}^T = \mathbb{T}_{\mathbb{L}} \left\langle \phi^{\mathsf{S}}(I^S), \mathcal{I}^T \right\rangle \tag{2.18}$$

$\mathcal{A}$ represents a series of operations (additional layers) that are applied to the source ConvNet to adapt it to the new learning task (hence, the representation as a function). For instance, in this research, the ImageNet dataset is used as the source domain. The source task is the classification of 1 000 natural objects and the target learning task is the classification of mammogram abnormalities which is just two classes (benign or malignant). Hence, the softmax layer of the source ConvNet is substituted to adapt it to a binary classification that uses a sigmoid function. Moreover, notice that operator $\mathcal{A}$, as described in (2.17), indicates the layers that are used for the transfer learning process. In the case of TL, all the learned weights from layer 0 to L are used. Only the additional layers that are represented by $\mathcal{A}$ are trained. All weights from layer 0 to L remain the same and are unchanged. This is why, transfer learning is used as a feature extractor that lets to build any classifier on top of it. Remind that a Deep Learning system learns different features at different layers (Sarkar et al., 2018). In the case of not using an ANN as the top trainable classifier, the $\mathcal{A}$ operator will not contain any additional layers. Because of this, the proposed definition is defined for the ConvNet context, where a trainable Deep Learning classifier is used.

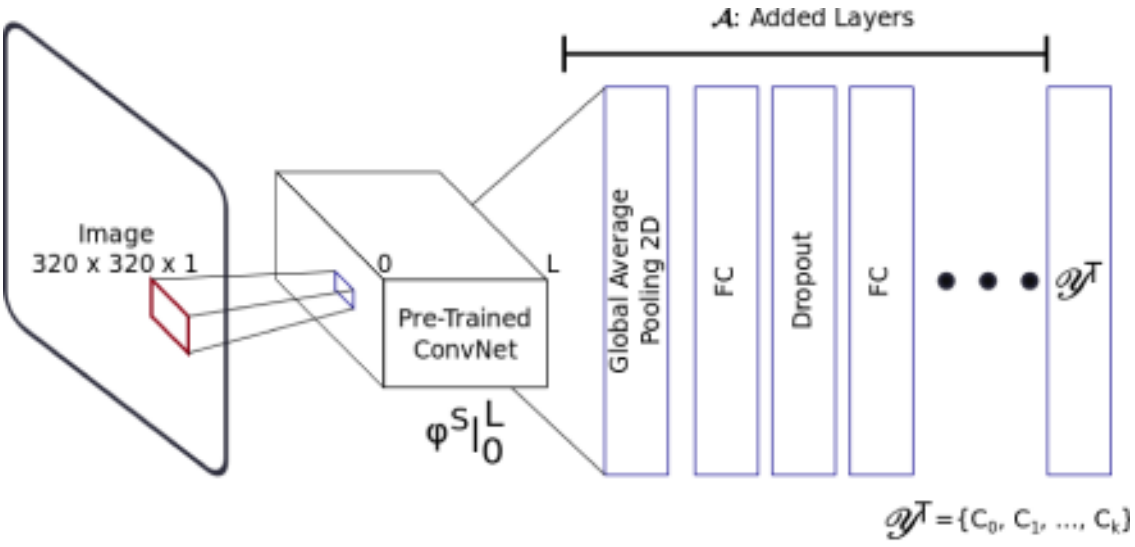

Figure 2.3: Transfer Learning Model

The figure represents the process of transfer learning, where the pre-trained weights of the original L layer ConvNet are used. Additional Layers ($\mathcal{A}$) could be used to obtain the target classification

Figure 2.3 shows the ideas upon (2.17) graphically. The 3 Dimensional box named Pre-Trained ConvNet represents the convolutional neural network pre-trained on a source dataset (i.e. ImageNet). As can be seen, all layers are used as they were originally trained. Some additional layers are used modifying the original output so that the new model classifies in the target task (e.g. mammo-

gram pathology). Therefore, the additional layers to be used in this thesis are presented in

$$\mathcal{A} = SoftMax \circ \cdots \circ Dro \circ Fc \circ GAvg_{2D} \tag{2.19}$$

2.5 Fine Tuning in Convolutional Neural Networks

Fine tuning is recommended when the target domain is very different from the source domain (Shanmugamani, 2018). This technique splits the pre-trained model in two parts. The layers contained in the first section of the ConvNet are frozen and their weights remain unchanged, whereas the layers in the second section are retrained. Similar to TL, the final layer is replaced (Sarkar et al., 2018). Next, a definition for Fine Tuning (FT) is proposed

Definition 2.5.1. In the context of convolutional neural networks, Fine Tuning $\mathbb{F}_{\mathbb{T}}$ is the process that improves the learning of a target predictive function $\mathcal{Y}^T = \phi^{\mathsf{T}}(\mathcal{I}^T)$ in a target domain $\mathcal{D}^T$ wrt. a target task $\mathcal{T}^T$ by re-training only the last r layers of a pre-trained ConvNet $\mathcal{Y}^S = \phi^{\mathsf{S}}(\mathcal{I}^S)$ of L layers previously trained on a larger dataset from a given source domain $\mathcal{D}^S$ and task $\mathcal{T}^S$, as described in (2.20); where $\mathcal{A}$ represents the fine tuning process over the source ConvNet of L layers and $\gamma = L - r$ is the layer from where fine tuning occurs.

$$\mathbb{F}_{\mathbb{T}}\left\langle \phi^{\mathsf{S}}(\mathcal{I}^S), \mathcal{I}^T \right\rangle = \mathcal{A}\left(\phi^{\mathsf{S}}\left(\phi^{\mathsf{S}}(\mathcal{I}^T)\Big|_0^{\gamma-1}\right)\Big|_\gamma^L\right) \tag{2.20}$$

Fine Tuning, in contrast with Transfer Learning, splits the pre-trained ConvNet in two parts at layer γ. From layer 0 to $\gamma - 1$, the original weights are frozen (i.e. they are not updated or re-trained). From layer γ until the end, the weights will be re-trained or "fine tuned" to adapt the source task to the target task. An interesting fact from (2.20) is that it permits to define Transfer Learning as a special case of Fine Tuning, when $\gamma = L$.

Figure 2.4 represents (2.20) graphically. As can be seen, the Pre-Trained ConvNet is split at layer γ. From 0 to $\gamma - 1$, the weights are frozen. From γ until L, the weights will be re-trained or "fine tuned" together with the additional layers $\mathcal{A}$. These additional layers are described by $\mathcal{A}$ in a similar way as in (2.19).

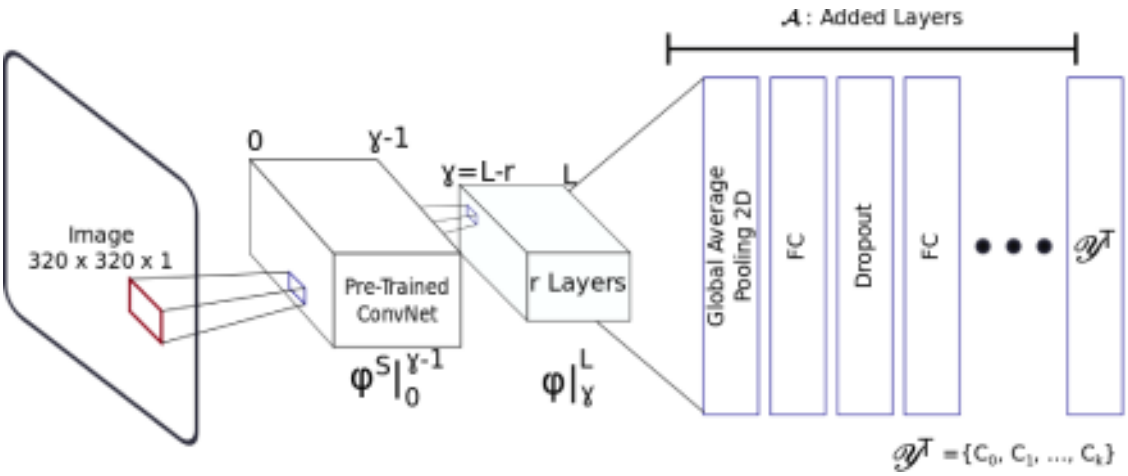

Figure 2.4: Fine Tuning Model

The figure represents the process of fine tuning, where the original ConvNet is split in two sections: The first one conserves the pre-trained weights of the original L layer model, while the following section that connects with the final output layer is random initialized. Additional Layers ($\mathcal{A}$) could be used to obtain the target classification

2.6 Whole Retrain in Convolutional Neural Networks

The case of Whole Retrain (WR) refers to using the architecture of the convolutional neural network and retrain it completely with the data from the target domain. Here, there is no transfer of knowledge, except for the structure of the model. Retraining a ConvNet from randomly initialized weights is the hardest of all cases and the solution converges when there is enough data; contrary to TL and FT that aid in preventing overfitting when data is not abundant. This case in fact is the traditional machine learning set up where: $\mathcal{D}^S = \mathcal{D}^T$ and $\mathcal{T}^S = \mathcal{T}^T$.

Regarding to (2.20), Whole Retrain (WR) is the special case of Fine Tuning (FT), when $\gamma = 0$, which implies $L = r$. This means that the whole source model is retrained. Notice that this is different from using the pre-trained model as a feature extractor. Here, all weights in each layer are initialized randomly and trained. Furthermore, Whole Retrain is the most time consuming of the three. In fact, the fastest of all is TL, because it only retrains $\mathcal{A}$. Fine Tuning is at second place, because it retrains the output layers determined in $\mathcal{A}$ and the r layers from the original model.

2.7 Ensemble Learning: Concept Review

An ensemble classifier consists in combining different classifiers to get a better generalization performance than each individual classifier when addressing novel instances (Hill & Kanagaratnam, 2016; Opitz & Maclin, 1999). According to Rasche (2013), an ensemble classifier makes multiple classification estimates (probability score, class label) using different classifiers and combines them into a single decision. One way of combining multiple classifiers is through *voting*. However, in order to design an ensemble classifier, it is necessary to have accurate and diverse individual classifiers (Dietterich, 2000). The diversity is the key in an ensemble method because an ensemble will be useful only if their individual classifiers or predictors disagree (Opitz & Maclin, 1999). This diversity includes: data diversity, parameter diversity, and structural diversity (Ren, Zhang, & Suganthan, 2016). The two most common methods to create ensemble predictors are Bagging and Boosting (Opitz & Maclin, 1999).

According to Dietterich (2000), a learning algorithm can be interpreted as a method to search a space of hypotheses. Notice that a training dataset cannot reflect the universe formed by all the examples. Therefore, each classifier predicts well in a determine subspace of the universe. By combining different predictors, it is expected to increase the probabilities of an accurate final decision. In this work, training different models in different techniques has been proposed to account for structural diversity. However, despite the fact that two different datasets are used, bagging is not used to create a new randomized training set. The latter could be a future work.

The two ensemble methods that are to be used in this research work are: *hard voting* and *soft voting*.

2.7.1 Hard Voting

Consider a standard supervised learning problem in classification, where it is required to find a function ϕ that maps an input space $\mathcal{X}$ (features) to an output space $\mathcal{Y}$. Trough the training of a classification algorithm (such as a ConvNet), it is obtained: $y_i = \phi(\mathbf{x}_i)$, for a particular test sample i. In this context, the hard voting ensemble is defined by the majority vote or mode of the predicted classes by each individual classifier. If m classifiers are trained, then hard voting is defined

in

$$y_i = mode\{\phi_1(\mathbf{x}_i), \phi_2(\mathbf{x}_i), \ldots, \phi_m(\mathbf{x}_i)\} \tag{2.21}$$

2.7.2 Soft Voting

Soft voting is similar to hard voting but has an essential difference. It depends on the probability score given by each individual classifier. Again, consider the supervised learning problem from the probabilistic point of view where the prediction can be written as $p_i = P(y_i|\mathbf{x}_i)$. In this case, the classifier returns a probability score p_i of sample i pertaining to class y_i. Hence, soft voting is defined by (2.22) for m individual classifiers.

$$y_i = \underset{i}{argmax} \sum_{j=1}^{m} w_j p_{ij} \tag{2.22}$$

Equation (2.22) indicates a weight sum of the probabilities scores. The corresponding class selected by the ensemble is the maximum argument of the weight sum of each probability score given by each individual classifier. Remind that in a binary classifications the probabilities for each class are complementary. This means for a sample i the probability of belonging to the *benign* class is p_i, and the probability of belonging to the *malignant* class is $1 - p_i$.

An interesting question that arises from (2.22) is regarding the selection of the values for w_j. These values can be selected to have the same value or can be manually input. In this work, a Perceptron is used in order to search for recommended values for w_j. This is suggested, because the soft voting equation resembles a similarity to the equation of a Perceptron.

2.8 Transfer Learning and Ensemble Learning in Mammogram mass classification

Breast Cancer is a public health problem due to the rates of death and incidence. Because of that, it is one of the most deadly types of cancer. Furthermore, the advances in medical imaging and computer science have aid in lowering the death toll risk by enhancing the interpretation of medical images and aiding the work of radiologists through the development of CAD systems (Stoitsis et al., 2006). As

reported by Jalalian et al. (2017), radiologists fail to detect from 10% to 30% of breast cancers. In fact, it can be suggested that the development of ConvNets is modifying the traditional 4 stages CAD model (image pre-processing, image segmentation, feature extraction and classification), because ConvNets are able to automatically extract features from data. Additionally, considering the capacity of ConvNets to perform pixel wise labeling, it could be stated that a CAD system based on ConvNets may perform image segmentation, feature extraction and classification as a one stage; needing only some image pre-processing.

This literature review aims to identify how transfer and ensemble learning are currently being used in the state of the art of breast mass abnormality classification. Some questions that naturally arise are:

- What pre-trained and ensamble models are most used to classify mammogram abnormality?
- How do researchers improve breast mass abnormality classification through transfer learning or ensemble learning?
- Is there a top performance or benchmark work?
- What are the possible future research topics in this area (research gaps)?

In order to try to figure the big picture of transfer and ensemble learning in the case of mass abnormality classification, some of the steps in the methodology proposed by Kitchenham and Charters (2007) for Systematic Reviews were used. Yet not all the steps of a systematic literature review were developed in this section.

2.8.1 Search Process

In order to retrieve relevant information of current state of the art regarding our research question (as indicated in Seccion 1.2.1), a search string was designed considering population, intervention, context and outcome and used in several scientific electronic databases such as: Springer Link (http://www.springerlink.com), Science Direct (Elsevier) (http://www.sciencedirect.com), IEEE Xplore (http://www.ieeexplore.ieee.org), Scopus (https://www.scopus.com), Web of Science (https://login.webofknowledge.com), ACM digital library (https://dl.acm.

org/), and PubMed (https://www.ncbi.nlm.nih.gov/pubmed/).

As second step, the search string was used in the aforementioned scientific digital repositories. Last iteration was on November 22 2018. Next, inclusion and exclusion criteria were designed to classify documents found by the use of the search string. To facilitate the SLR research, the software Start (State of the Art Through Systematic Review is used. It helps to identify duplicated documents automatically and aids in the classification of them as well.

Search String Design

A search string is built considering population, intervention, context and outcome viewpoints:

- Intervention: It is related to the means or technologies being investigated to answer the research question (RQ). These are: *ensemble learning* and *transfer learning*.
- Outcome: This is the desired result of the research. In other words, this research looks for an improvement on the *accuracy* and/or *precision* of *classification* or *detection* or *prediction* of breast cancer in a mammogram image.
- Context: this research is looking for information related to mammography, because it is considered the goal standard for early detection of breast cancer Drukteinis et al., 2013.
- Population: medical image analysis by computer aided systems

Consequently, the proposed search string is shown in Table 2.1

Table 2.1: Search String

”breast cancer” **AND** (”classification” **OR** ”detection” **OR** ”prediction”) **AND** (”ensemble learning” **OR** ”transfer learning”) **AND** mammo*

Inclusion and Exclusion criteria

To identify the most relevant documents in the search phase, the following Inclusion and Exclusion criteria were applied.

Inclusion:

- Digital mammography images from public datasets are used.
- Document uses transfer learning or ensemble learning.
- Document solves classification problem.

Exclusion:

- Tomosynthesis, histological, ultrasound or magnetic resonance images are used
- Studies prior year 2014 are excluded
- papers without results are excluded
- segmentation and detection are not of main interest

A special case is when the search process finds systematic literature reviews or surveys. This documentation has been reviewed, but it was not included in extraction because those documents do not perform any experimentation. This is the case of the work in Yassin et al. (2018) and Dias Pedro et al. (2018).

2.8.2 Study Selection

For the proposed research questions, documents must be related to the use of transfer and ensemble learning in the classification of cancerous masses in mammography. It is advisable for the document to be written in English and published in a journal. Also, proceeding works are included. The procedure of selecting primary studies is indicated as follows and summarized in Figure 2.5:

- Studies Pre-classification using *Start*
- Manual identification and removal of duplicated studies among the databases used in Search Process.
- Application of selection criteria to title and abstract of each publication. In cases were abstract and title were not clear enough, complete review of the publication was carried out.

- Accepted works are then classified for data extraction

2.8.3 Search String Results

The search string designed in Section 2.8.1 produced a total of 173 studies from July 2018 to November 22, 2018. This result is shown in Table 2.2. After that, Start tool was used to identify duplicated studies. Later on, studies were further filtered with the aid of the proposed inclusion and exclusion criteria. In Figure 2.6 a pie chart with the number of studies found in each database is shown. Finally, studies were classified in four categories: *accepted*, *rejected*, *duplicated*, *unclassified*.

Table 2.2: Search Results

Database	**#Publications: 2014-2018**
Scopus	38
IEEE	11
Science Direct	51
PubMed	10
ACM	25
Web of Science	23
Springer	15
Total	**173**

Additionally to the 32 accepted studies, a previous work by the author of this thesis was added. The aforementioned work explores mammogram classification and transfer learning on the CBIS-DDSM dataset (Lee et al., 2017). This work was published after the application of the search string (Falconí, Pérez, & Aguilar, 2019). Because of that, the number of accepted papers is increased to 33.

2.8.4 Study Selection Results

By using the proposed selection process and reviewing title, abstract, keywords and content of the articles, the initial 173 results were pruned to 31 studies. A previous work titled *"Transfer Learning in Breast Mammogram Abnormalities Clas-*

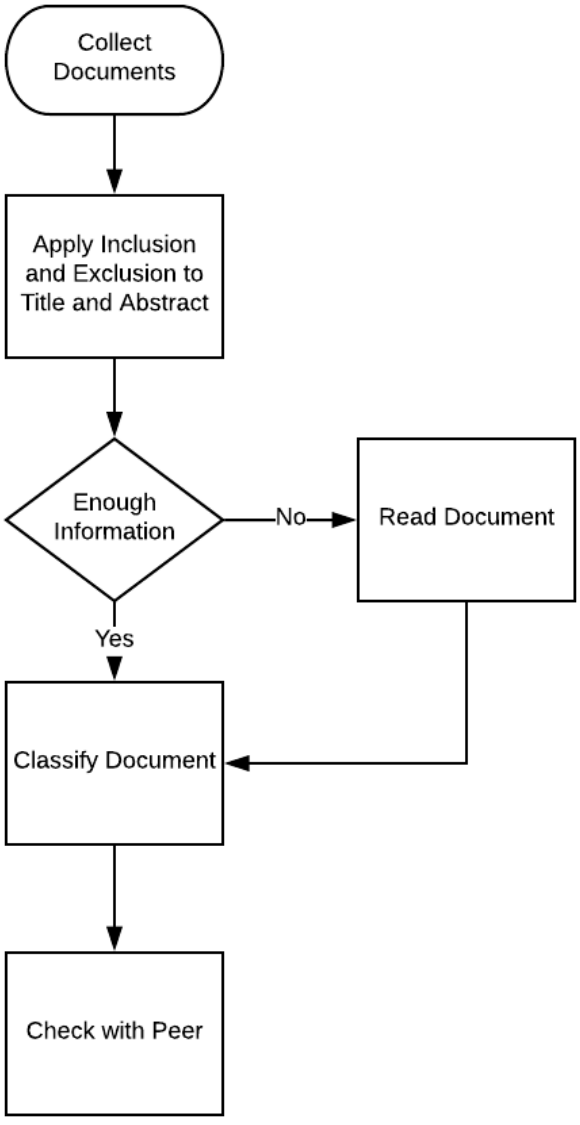

Figure 2.5: Selection Process

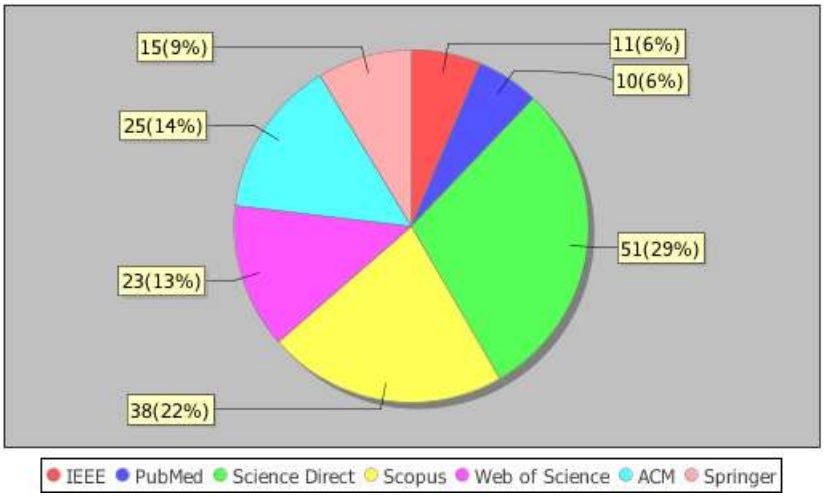

Figure 2.6: Databases Primary Documents

Table 2.3: Selection Results

Document Status	Number of Documents	%
ACCEPTED	32	18%
REJECTED	93	53%
DUPLICATED	45	26%
UNCLASSIFIED	4	2%
TOTAL	174	100%

sification With Mobilenet and Nasnet" (Falconí et al., 2019) was additionally included. Consequently, Table 2.3 shows a total of 32 studies accepted to review, including the aforementioned article.

2.8.5 Literature Review

The search string found a total of 174 research documents related to transfer and ensemble learning in mammogram abnormalities classification. There were two systematic literature (SLR) works found that were reviewed in order to have a global picture of the stated of the art. However, since these studies portray a general overview of machine learning and breast cancer, they were separated from the specific articles that proposed solutions about mammogram abnormalities classification.

This literature review is organized in five sub-sections. The first sub section section reviews the SLRs found in the search process. Next, the models that have

been used in transfer learning in the literature review are addressed. The third subsection reviews the mammography datasets used in literature. The last subsection discusses the performance of the selected studies in mammogram classification.

Machine Learning and Breast Cancer Detection

The SLR by Yassin et al. (2018), titled *"Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review"*, aims to investigate the state of the art of computer aided diagnostic (CADx) and detection (CADe) systems for breast cancer. The authors gathered a total of 154 final studies. Their study reflects the difficulties to establish a benchmark in the area due to the diversity of datasets, metrics and methodologies involved. As pointed out by Dias Pedro et al. (2018), it is not a common practice to use the radiologist assistance to validate the results obtained by the machine learning technique. Consequently, research of machine learning in mammogram classification should be integrated with the review of physicians and radiologists when possible. Furthermore, Dias Pedro et al. (2018) questions if mammogram classification is already a problem solved due to the classification results obtained by Deep Learning. In the opinion of the former author, research in this field will continue mainly because of the importance of breast cancer in public health. It is also correct to say that research in the field of classification of mammograms will continue linked to the development of computer vision techniques. On account of researchers keep improving the convolutional neural network model to create lighter and stronger networks, research areas like medicine benefit of such progress.

Regarding the most used techniques, both authors show that Artificial Neural Networks (ANN), Support Vector Machine and K-Nearest Neighbors (KNN) are frequently used in literature(Dias Pedro et al., 2018; Yassin et al., 2018). However, in the articles reviewed by the authors ConvNets are mainly used as a feature extraction option. In fact, Deep Learning models, such as auto-encoders, regional convolutional neural networks, single shot detection algorithms, are not mentioned in their reviews. The use of transfer learning is also not included in their works. Therefore, it could be of interest to submit a systematic review of transfer learning in mammogram classification.

About metrics, the two most commonly used are *accuracy* (ACC) and the *area under the ROC curve* (AUC) (Dias Pedro et al., 2018). This shows that the use of metrics like *balanced accuracy* (BACC) or the *Mathews correlation coefficient* are little used in literature. Hence, it is of interest to extend research in mammogram classification by using new classification metrics. Also, Dias Pedro et al. (2018) indicates that the hold out approach is the most used validation and test technique, followed by k-fold cross validation.

Regarding ensemble learning, the work of Yassin et al. (2018) indicates that they are not so frequently used. This technique involves decision trees, adaptive boosting algorithms and random forest. The performance of the ensemble classifier as well as the performance of ANN, SVM and KNN varies among the works reviewed.

In the following sections, the analysis of the collected studies by the process detailed earlier is carried out. The works reviewed in the next sections exclude systematic literature reviews and focus on research that aims to solve the classification problem by providing experimental evidence.

Pre-Trained ConvNets used in Mammogram Classification

Convolutional Neural Networks have become essential in image classification (Krizhevsky et al., 2012), image recognition (Simonyan & Zisserman, 2014b), sentiment analysis (Dos Santos & Gatti, 2014), and natural language processing (Conneau, Schwenk, Barrault, & Lecun, 2016). In those fields, ConvNets have achieved successful results. Researchers in the field of Deep Learning present upgrades or new versions of some previous models and even new architectures.

In this literature review, several pre-trained convolutional neural networks have been used in mammogram classification. Some of the pre-trained models have been used as a basis for new designs. In some cases, there are various versions of the same original architecture with some modifications. This is the case of AlexNet, Vgg-F, Vgg-M that were modified by Chatfield, Simonyan, Vedaldi, and Zisserman (2014) . In the case of AlexNet, both versions, the original by Krizhevsky et al. (2012) and the one proposed by Chatfield et al. (2014), will be referenced as AlexNet. In Figure 2.7, the models found in this literature review

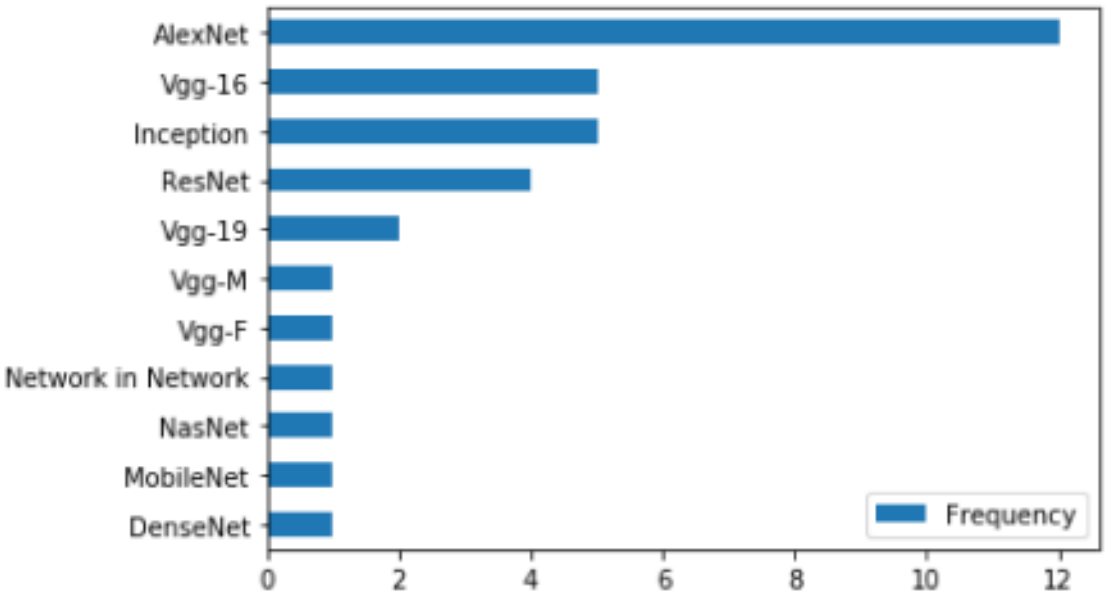

Figure 2.7: Frequency of Pre-Trained ConvNets found in Literature

are shown. The most frequent ConvNet used in the collection of studies corresponds to AlexNet (12 cases), followed by Vgg-16 (5 cases), inception (5 cases) and ResNet (4 cases).

The fact that AlexNet is the most frequent architecture found in the studies reviewed may be twofold: AlexNet is not as deep as other networks which makes it not so hungry of computation resources and fast to train compared with ConvNets as deep as ResNet, DenseNet or NasNet. The second reason might be because it is historically the first ConvNet to solve the problem of 1000 image classification. This surely might have called the attention of the research community in applying it in medical images.

Mammogram Databases

Datasets are the main source in a machine learning study, whether they are labeled or not. In the case of breast cancer, there are other imaging techniques besides mammography like: ultrasound, magnetic resonance imaging (MRI), (Yassin et al., 2018), and tomosynthesis (Niklason et al., 1997), among the principals. The mamogram image used in this research comprises 4 images in total, 2 for each breast and from two different perspectives: Cranial-Caudal (CC) and Mediolateral-Oblique (MLO). There are also two types of mammogram images: full digital mammography and film mammography.

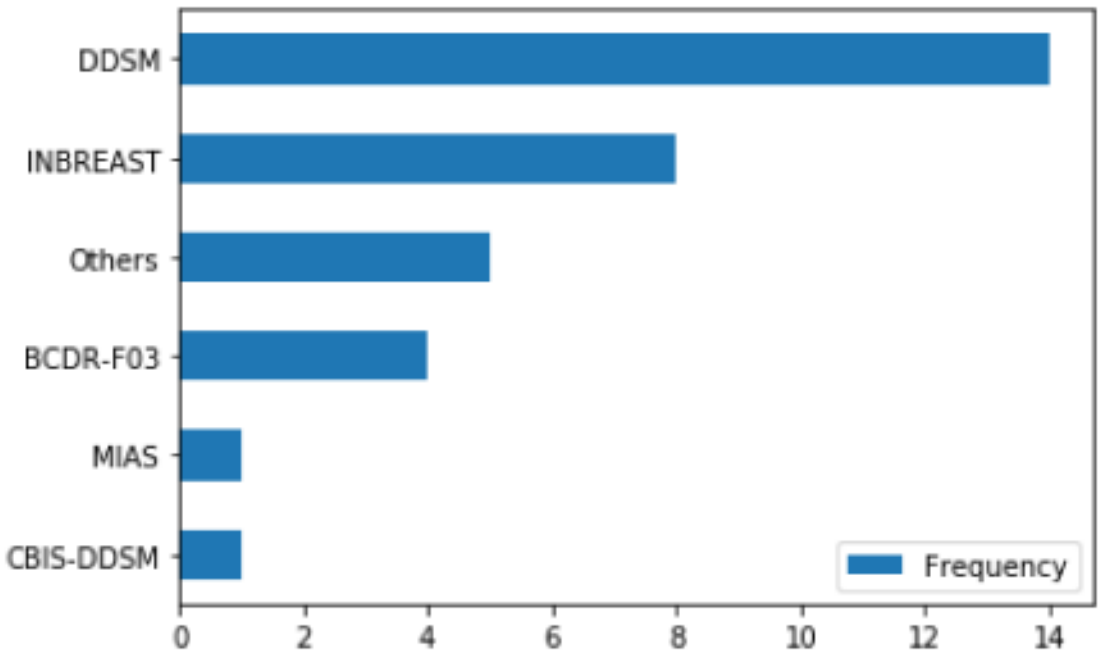

Figure 2.8: Datasets used in mammogram classification

Figure 2.8 shows the most common databases found in this review. DDSM (Heath, Bowyer, Kopans, Moore, & Kegelmeyer, 2000) is the most used dataset which appears in 14 documents. At a second place is the Inbreast dataset(Moreira et al., 2012), which is used to experiment in this work. Inbreast contains full digital mammogram images which are of better quality and more recent than DDSM. However, a drawback of this dataset is that it contains very few samples.

Table 2.4: Inbreast Literature Review Performance

Article Title	Authors	AUC
"Deep Convolutional Neural Networks for breast cancer screening"	Chougrad, Zouaki, and Alheyane (2018)	0.97
"Automated Analysis of Unregistered Multi-View Mammograms with Deep Learning"	Carneiro, Nascimento, and Bradley (2017a)	0.94
"Deep generative breast cancer screening and diagnosis"	Shams, Platania, Zhang, Kim, and Park (2018)	0.925
"Unregistered multiview mammogram analysis with pre-trained deep learning models"	Carneiro, Nascimento, and Bradley (2015)	0.91
"A deep learning approach for the analysis of masses in mammograms with minimal user intervention"	Dhungel, Carneiro, and Bradley (2017)	0.91
Deep Learning Models for Classifying Mammogram Exams Containing Unregistered Multi-View Images and Segmentation Maps of Lesions	Carneiro, Nascimento, and Bradley (2017b)	0.91
"Deep multi-instance networks with sparse label assignment for whole mammogram classification"	Zhu, Lou, Vang, and Xie (2017)	0.89
"Comparing the performance of various deep networks for binary classification of breast tumours"	Hamidinekoo, Suhail, Denton, and Zwiggelaar (2018)	0.87

Table 2.5: MIAS Literature Review Performance

Article Title	Authors	ACC	AUC
"Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks"	Guan and Loew (2017)	0.91	
"Deep Convolutional Neural Networks for breast cancer screening"	Chougrad, Zouaki, and Alheyane (2018)		0.99

Related works performance

In this research work, the main databases used were: Inbreast (Moreira et al., 2012) and MIAS (SUCKLING J, 1994). In Figure 2.8, there are 8 works related to Inbreast dataset and 1 work for MIAS. The best result achieved for Inbreast is the work of Chougrad et al. (2018) with $AUC = 0.97$. The average of results for Inbreast for the AUC metric is 0.915. For MIAS there are two works with results of $ACC = 0.91$ (Guan & Loew, 2017) and $AUC = 0.99$ (Chougrad et al., 2018). Assuming that with high values of ACC, this metric is similar to AUC, the average result for MIAS is around 0.95. These results are shown in Tables 2.4 and 2.5 for Inbreast and MIAS, respectively.

Discussion

Medical images are a challenge for Deep Learning because they are scarce and different from natural images. All authors in the reviewed studies agree in using transfer learning to overcome the overfitting problem that is prone to happen when training convolutional neural networks with little data. The deeper the network, the more unstable it can be; specially if enough data is not provided. Because of this reason, even though transfer learning is not a regularization technique, it certainly helps to reduce overfitting.

The works in the field of mammogram classification can be classified in two main groups: whole mammogram classification and region of interest (ROI) mammogram classification. The former aims to process the whole mammogram image while the latter extracts patches of only the mass images from the mammography. This work is fitted in the second category.

In whole mammogram classification, Al-masni et al. (2018) and Al-antari, Al-masni, Choi, Han, and Kim (2018) have based their design on the YOLO model (Redmon, Divvala, Girshick, & Farhadi, 2016) to recognize the mass in the mammogram and classify it as benign or malignant. Transfer learning is used to initialize the weights of the ConvNets proposed by Al-antari et al. (2018). In fact, this author provides a full mammogram solution that includes mass detection, mass segmentation and mass classification. Yolo is used in the mass detection stage as it divides the image in a grid and predicts if that grid is a mass or not. Full resolution Convolutional Network (FrCN) is used to segment the detected mass. Finally, for classification they use a simplified version of AlexNet. Even though they work with whole mammogram images, the classification is in fact of a proposed region of interest previously defined by their methodology. This is understandable since working with the whole mammogram image is difficult due to its size and pixel value. Usually, the size of a mammogram image could be bigger than 1024 pixels, and the color value is in the 16 bit resolution; which ranges from 0 to 65535. However, the pixel value is not a problem since it is usually normalized to a real value of from 0.0 to 1.0; this practice helps the ConvNet to converge. The overall accuracy for classification in this approach is 97% (Al-masni et al., 2018). A similar approach is followed by Hu, Li, and Jiao (2016) with some differences. First, they use the saliency maps approach to detect possible regions containing masses. Second, they use AlexNet as a feature extractor by

training a SVM classifier with the 4096 FC layer. The 4096 feature vector is obtained by feed forward of the mammgram image through the pre-trained ConvNet.

The work of Carneiro, Nascimento, and Bradley (2017b) is another approach for whole mammogram classification. They base their design on the AlexNet architecture where they use transfer learning by removing the last FC output layer, which originally has 1000 categories, with a 3 neuron output (Benign, Malignant and Negative). Their work is different from others in literature in providing a 3 classification category of the mammogram. In their design they use up to 6 convolutional neural networks trained in transfer learning mode. Each ConvNet is fed with a different view of the Mammogram image, which includes both the micro calcification mask and the mass mask. This is an interesting contribution, because the real classification of a mammogram exam depends on masses and calcifications. In this work, only mass images have been considered.

Regarding the ROI mammogram classification, there are different approaches. In the work by Perre, Alexandre, and Freire (2018), 3 ConvNets are fine tuned in a mammogram database by randomly initializing the last FC layer. The ConvNets used are Vgg-like. Once they have fine tuned the ConvNets, they use the sixteen layer as a feature vector to train a SVM for the final classification. An interesting contribution of the authors is that they shows that global contrast image normalization does not improve the classification results. In fact, they find that results are deteriorated by using global contrast normalization. Because of that, in the present work, only image rescaling and pixel normalization were used as pre-processing steps. Chougrad et al. (2018) compare the classification performance of Vgg16, Resnet-50 and Inception-v3 by adding 5 dense layers at the end of each pre-trained model. The latter authors use fine tuning and investigate different depths to optimize the classification result in two categories. It is reported by them that fine tuning from the last 2 layers of the model improves classification performance.

In the case of X. Zhang, Zhang, Han, et al. (2017) and Hamidinekoo, Suhail, Denton, and Zwiggelaar (2018), the concepts of transfer learning and fine tuning are understood a little different from the definitions proposed in Sections 2.4 and 2.5. In their case, even though they substitute the last FC from 1000 output neurons to the two output neurons, they understand fine tuning as the initialization of the

weights values across the whole model to retrain it with the new data. This motivated to define TL and FT as previously stated. Consequently, this thesis is more similar to Chougrad et al. (2018) and Guan and Loew (2017).

Regarding the pre-processing of the images, all authors use image rescaling to adapt the image's original size to that required by the ConvNet model. This step alters the information contained in the image. Some authors control the aspect ratio of the image to lessen this effect (Falconí et al., 2019; Perre et al., 2018). Also, segmentation through Otsu and morphological operations are carried out in the work of Jiang et al. (2017), but they do not report the parameters used clearly. Because of that, and based on our previous work (Falconí et al., 2019), where the Otsu segmentation step was tested with not a big improvement in the overall result, this work did not implement Otsu.

Data augmentation is widely use by researches in literature. However, there is some discussion regarding the rotation of the original images. Most authors use angles of $\{\pi/2, \pi, 3\pi/2\}$ which clearly do not affect the quality of the image (Chougrad et al., 2018; Jiang et al., 2017; Perre et al., 2018). In this work, in order to avoid distorting the image substantially, the Augmentor (Bloice, Roth, & Holzinger, 2019) library has been selected to perform data augmentation.

2.9 Emerging issues and need for empirical research

The research community is clearly enthusiastic in the field of mammogram classification and in the development of CAD software to improve early detection. Transfer learning is clearly becoming an important technique in order to improve both image detection and classification in Deep Learning. Due to the fact that medical imaging datasets have few samples compared with datasets like ImageNet, where Deep Learning through ConvNets has achieved great results, researchers are using it to improve the performance of predictive models.

The literature review here presented has shown that some authors use Transfer Learning in different ways. One of the uses is to obtain a highly reliable feature vector that is used to train other types of classifiers like SVM. Others, retrain the whole network by using the original weights as starting values and just substitute the last FC according to the classification Task. Also, it has been found in authors

like Chougrad et al. (2018) a similar approach to this work's understanding of the concepts of FT and TL; where additional layers are used previous the final FC with the two categories.

There are two main approaches to mammogram classification in the works here studied. The first is to process the whole mammogram image (which is considered more challenging). The second, consists to process ROI images. In either case, both approaches use a resize of the original image. Consequently, this means that there has not yet been found a model to process a mammogram image in its original size.

Since most authors compare at most 3 pre-trained ConvNets, and being AlexNet the most studied model in literature, it is concluded that it is of scientific interest to make a systematic study of Transfer Learning, Fine Tuning and Whole Retrain (with random initialization) in new architectures like ResNeXt, Xception, Nasnet, Mobilenet, by experimenting with them and comparing its performance in mammogram pathology classification. As a result, this work experiments with 20 pre-trained ConvNets.

Chapter 3

Research Methods

3.1 Introduction

This research study focuses on the performance of classification of breast pathologies as benign and malignant in mammogram images, using fine tuned pre-trained convolutional neural networks. The hypothesis proposes the study of Transfer and Ensemble learning. According to the literature review, presented in the previous section, Transfer Learning is useful to train pre-trained ConvNets with small target data (which is the case of mammogram), but just some models have been explored; being ALexNet the most studied and leaving models such as ResNext or Xception behind. Because of this, it is of interest of this work to perform a systematic study that tests the effect of Transfer Learning (TL), Fine Tuning (FT) and Whole Retrain (WR) of 20 pre-trained natural models in the classification of mammogram images. This corresponds to the second specific objective, as described in Chapter 1.2.2, where it was proposed to explore state of the art pre-trained ConvNets. How these experiments are designed and performed is discuss in this chapter.

Ensemble learning is useful to increase the generalization capabilities of a model and it is usually implemented once other models have been tested (Géron, 2017). Because of that, it is important to investigate if an ensemble of the explored ConvNets is able to perform better in mammogram classification. According to the presented literature review, an ensemble of ConvNets has not yet been tested. This could be because the training of each ConvNet takes a lot of time. Moreover, once a ConvNet has been trained, it performs accurately in the given dataset but poorly in a different one. On account of this work's research question that aims to

increase classification performance and system's generalization, ensemble techniques are explored. This is important so that the final model can work with the two main types of mammogram images: digital (Inbreast dataset) and film (MIAS dataset).

According to the literature review presented in Chapter 2, there are different metrics used by different researchers in the field. Because of that, and considering the observations in Canbek et al. (2017), regarding the binary classification metrics, this work considers of interest to use several metrics to evaluate classification performance. The behavior of each metric is also compared and analyzed. This is a contribution that hopes to standardize performance metrics in mammogram binary classification.

This section -Research Methods- provides the details of the research strategy adopted as well as the methodology proposed to address the research issues previously described regarding mammogram classification. First, a formalization of the problem at hand -mammogram abnormalities classification- is introduced. Then, the research strategy is discussed regarding the means to dataset generation (or data collection), experiment design and analysis framework. Finally, threats to validity and reliability of the proposed method are discussed.

3.2 Problem Formulation

A mammogram ROI patch image is represented as a tensor $\mathcal{I} \in \mathbb{R}^{w \times h \times d}$, where w, h and d are the width, height and depth of the image, respectively. Let $\mathcal{X}^{\mathcal{T}}$ to be the target set of all mammogram ROI patch images, such that $\mathcal{I} \in \mathcal{X}^{\mathcal{T}}$, and $\mathcal{Y}^{\mathcal{T}}$ the target label space, such that $\mathcal{Y}^{\mathcal{T}} = \{benign, malignant\}$ where $\mathcal{Y}^{\mathcal{T}} \in \mathbb{R}$. Then, there exists a target function $\phi^{\mathsf{T}}(\cdot)$ that maps the input target space to the label output space as indicated in (3.1) and (3.2).

$$\phi^{\mathsf{T}} : \mathcal{X}^{\mathcal{T}} \rightarrow \mathcal{Y}^{\mathcal{T}} \tag{3.1}$$

$$\mathcal{Y}^{\mathcal{T}} = \phi^{\mathsf{T}}(\mathcal{I}) \tag{3.2}$$

Therefore, the target Domain $\mathcal{D}^{T}$ is defined by the tuple $(\mathcal{I}_i, y_i^T)$ as indicated in (3.3), where i is a sample and n is the total number of samples. Each y_i^T is

calculated as indicated in (3.4).

$$\mathcal{D}^T = \{(\mathcal{I}_1, y_1^T), (\mathcal{I}_2, y_2^T), \ldots, (\mathcal{I}_i, y_i^T), \ldots, (\mathcal{I}_n, y_n^T)\} \tag{3.3}$$

$$y_i^T = \phi^{\mathsf{T}}(\mathcal{I}_i) \tag{3.4}$$

In order to obtain $\phi^{\mathsf{T}}(\cdot)$, Transfer Learning $\mathbb{T}_{\mathbb{L}}$ and Fine Tuning $\mathbb{F}_{\mathbb{T}}$ are to be used from source models $\phi^{\mathsf{S}}(\cdot)$ trained on the ImageNet source domain of natural images $\mathcal{X}^S$. This is indicated in (3.5) and (3.6).

$$\phi^{\mathsf{T}}(\mathcal{X}^{\mathcal{T}}) = \mathbb{T}_{\mathbb{L}} \left\langle \phi^{\mathsf{S}}(\mathcal{X}^{\mathcal{S}}), \mathcal{X}^T \right\rangle \tag{3.5}$$

$$\phi^{\mathsf{T}}(\mathcal{X}^{\mathcal{T}}) = \mathbb{F}_{\mathbb{T}} \left\langle \phi^{\mathsf{S}}(\mathcal{X}^{\mathcal{S}}), \mathcal{X}^T \right\rangle \tag{3.6}$$

In Chapter 2, Sections 2.4 and 2.5, TL and FT were defined in (2.17) and (2.20), respectively. In the aforementioned relations, the operator $\mathcal{A}$ is used. Such operator defines the changes that are made to the original architecture to adapt it to the target task $\mathcal{T}^T$. In this work the target task corresponds to the mammogram pathology classification (benign or malignant). Because TL and FT permit to transfer knowledge from the source domain and adapt it to the the target task, the relations (3.7) and (3.8) represent such process where $\mathcal{I} \in \mathcal{X}^T$:

$$\mathbb{T}_{\mathbb{L}} \left\langle \phi^{\mathsf{S}}(\mathcal{I}^S), \mathcal{I} \right\rangle = \mathcal{A}\left(\phi^{\mathsf{S}}(\mathcal{I}) \Big|_0^L \right) \tag{3.7}$$

$$\mathbb{F}_{\mathbb{T}} \left\langle \phi^{\mathsf{S}}(\mathcal{I}^S), \mathcal{I} \right\rangle = \mathcal{A}\left(\phi^{\mathsf{S}}\left(\phi^{\mathsf{S}}(\mathcal{I}) \Big|_0^{\gamma-1} \right) \Big|_\gamma^L \right) \tag{3.8}$$

As stated in Chapter 2, Section 2.4, the $\mathcal{A}$ operator consists of a series of layers that adapt the pre-trained ConvNet to the target task $\mathcal{T}^T$. The $\mathcal{A}$ operator defines the models for TL and FT Experiments. This is discussed in the second phase of the proposed strategy in the next section.

3.3 Research Strategy

Computer Science (CS) is a relative new field of scientific research. According to Tucker (2004), the field of CS has undergone a dramatic evolution since the 1930s, when its fundamental mathematical principles were developed. Across the years, new research areas and applications have emerged. The problem with CS is that in some cases its object of study is a program (which is a human made object). This has called the attention of scientist in order to determine if CS is Science. Denning (2005) has evaluated this question in his research article titled: *"Is Computer Science Science?"*. The researcher concludes that CS meets every criterion to be considered as such because a main characteristic of science relies in proposing hypotheses and testing them empirically. The difference in CS wrt. other sciences, specifically in the field of Artificial Intelligence (AI), is that CS studies computer programs that perform tasks in environments (Cohen, 1995).

Scientific research could be defined as a careful and systematic study or investigation of a problem of interest in a field of knowledge (Ayash, 2014). In order to perform this work in a systematic manner, research methods are required. Due to the empirical nature of the concept given in Cohen (1995), AI research implies the use of empirical methods that combine exploratory and experimental techniques to run the program and record its behavior for analysis. According to Cohen (1995), studying an AI program is not different from studying a moderately intelligent animal. In his article *"Research Methodologies in Computer Science and Information Systems"*, Ayash (2014) enumerates three methods that are used in CS: experimental method, simulation method, and theoretical method. The former involves setting an hypothesis which is verified or falsified through a designed experiment that collects data (Berndtsson, Hansson, Olsson, & Lundell, 2007).

The second proposed specific objective in this work states the need for experimentation of different pre-trained convolutional neural networks in mammogram classification. This request is fulfilled using the experimental method because the study of programs that perform a task in an environment is considered an experiment (Cohen, 1995). Also, this work's hypothesis states a causal relation: Transfer Learning from Pre-Trained ConvNets in the Natural Domain improves mammogram classification. This sentence itself states that is possible to use TL with enough accuracy (quantitative measure) to classify mammogram images. The literature review has already confirmed that it is possible to use TL and that

there are some ConvNets that haven't been explored yet. However, what happens if retraining with random initialized weights performs better than or at least as well as Fine Tuning or Transfer Learning in mammogram images? The latter question implies the condition for the control group. In experimental methodology, in order to assess if an intervention causes an effect on the object of study is often required to have both control and test groups. The control group does not receive the treatment. The second part of the hypothesis explores the ensemble of pre-trained ConvNets. This again is verified by contrasting the performance results obtained in individual models: are they better than an ensemble?

On account of the reasons exposed, it is concluded that an experimental strategy fits the research question and objectives. Due to the nature of the classification problem, binary classification metrics to evaluate the performance are needed. All of this implies that this research is of quantitative nature, primarily.

3.4 Experimental Research Design

In the previous section, it was showed that an experimental strategy is useful to address the research problem. To achieve this research's goals and objectives, this section presents the experimental design. Usually, an experimental method is comprised of main 5 stages which are: participants, variable analysis, materials and instrumentation, procedures, and analysis. These stages are discussed next for this work's research problem.

3.4.1 Participants

This work aims to find the best model to classify mammogram ROI images by using pre-trained ConvNets on a natural domain. Therefore, the participants of this research are formed by the universe of pre-trained ConvNets. From that universe, 20 models that are included in the *Keras* API (Chollet et al., 2015) of the *Tensorflow version 1.13.1* machine learning library (Martín Abadi et al., 2015) were considered for this study. Because the library contains different types of pre-trained ConvNets, using all of them accounts for providing diversity of models for this work.

Table 3.1 shows the pre-trained ConvNets that were used for experimentation.

They all have been trained on the ImageNet Dataset (Deng et al., 2009) and their weights are available. This work was not limited by the type of ConvNet because changes in the versions of the ConvNets also carry modifications in their inner architecture and number of layers. Consequently, it was chosen to include all pre-trained ConvNets in the *Keras* library available in Tensorflow v1.13.1.

Table 3.1: Pre trained ConvNets used

Vgg	**Densenet**	**Inception**	**Nasnet**
16, 19	121, 169, 201	v3, resnet-v2	large, mobile
Mobilenet	**Resnet**	**Xception**	**Resnext**
v1, v2	50, 50v2, 101, 101v2, 152, 152v2	v1	50, 101

AlexNet is not included in this study because the literature review carried out shows that this ConvNet has already been studied. Also, models that were developed after AlexNet are deeper and have an increased performance on the ImageNet Dataset. Because these ConvNets are not randomly chosen but constitute a *convenience* sample of them, this work's experiments are of the quasy-experimental design kind. The control and experimental group are built according to the research hypothesis. This situation is clarified in the next section where the hypothesis is split in independent and dependent variables.

3.4.2 Experiment Variables and groups

The hypothesis of this research work, presented in Chapter 1, Section 1.2.1, states that an ensemble of different fine tuned ConvNets on the mammogram dataset improves the average classification performance. The independent variable is related to the intervention, which corresponds to what is used to increase classification performance; that is the ensemble itself. The dependent variable is related to the outcome, which is the classification performance. Therefore, the null hypothesis is stated as:

H_0: an ensemble of fine tuned ConvNets on the mammogram dataset does not improve classification performance.

In order to assess the hypothesis, two experimental groups were formed. The control group was formed by the transferred learning and fine tuned single models, whereas the experimental group was formed by the ensemble of the best models that we may find in this research.

As described, in Chapter 2, Fine Tuning of a ConvNet can be achieved by means of Transfer Learning (TL) or Fine Tuning (FT). Equations (3.7) and (3.8) were stated to clarify this situation. A natural question that arises is that of TL and FT effectively achieving better classification performance than the same ConvNets trained on random initialization. Because of that, it was necessary to consider a related hypothesis that evaluates if Fine Tuning and Transfer Learning do improve classification of mammogram pathology images when used. Consequently, two experimental groups were formed: control group formed by models trained with random initialization of weights (named as whole retrain) and the experimental group formed by models trained with Transfer Learning and Fine Tuning.

3.4.3 Materials and Instrumentation

An empirical method to study an AI program involves running the program and recording its behavior (Cohen, 1995). The instrumentation involves a program that trains each model from Table 3.1 in TL, FT, and WR, and records its classification performance (Test) for later analysis. The program uses Tensorflow version 1.13.1 (Martín Abadi et al., 2015) as the main machine learning library to train each model and has 3 main parts:

1. tensor generation: images from the dataset are organized in order to train and evaluate the model.

2. model training: the pre-trained model is trained by using TL, FT or WR as selected by user

3. model evaluation: Once the model has been trained, it is evaluated on a test set.

In order to fine tune each model in the target task of mammogram classification, a dataset with mammogram images is required. For this research work, the Inbreast (Moreira et al., 2012) and MIAS (SUCKLING J, 1994) datasets were used to extract the Region of Interest Images (ROI) of both benign or malignant masses.

Inbreast Database

The Inbreast database (Moreira et al., 2012) has a total of 410 Full Field Digital Mammogram (FFDM) images stored in DICOM format that correspond to a total of 115 cases. The database includes images in both types of view (MLO and CC) for both breast in 90 cases. Regarding the mammogram abnormality class, the database has examples of: normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries and multiple findings. The Calcifications class is prominent in the database. This database uses the Breast Imaging Reporting and Data System (BI-RADS) scale to categorize the level of suspicion of the abnormality. There are six categories in the BI-RADS scale: category 0, not conclusive; category 1, no findings; category 2, benign findings; category 3, probably benign; category 4, suspicious; category 5, highly probably malignant; category 6, proved cancer.

The database also provides annotations of the abnormalities in an XML file. This file contains the locations of the masses, which is of interest in this study, and the locations of calcifications. Furthermore, the database includes mass mask segmentation images in jpg format. These masks were used to extract the ROI from the mammogram images. In this research, it was preferred to use the binary mask rather than the coordinates in the XML file. This database can be accessed online at: http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database.

MIAS Database

The Mammo- graphic Image Analysis Society (MIAS) database (SUCKLING J, 1994) is a public database which contains a total of 322 mediolateral oblique (MLO) film mammography images with examples of both benign and malignant forms of abnormalities commonly encountered in breast cancer altogether with normal cases. Images in the database are in PGM format. The database ground truth is contained in a Readme File. This file has both the pathology of the abnormality found in the image and also the location of the mass. The pathologies registered in the database are shown in Table 3.2. This database can be accessed online at: http://peipa.essex.ac.uk/pix/mias/.

Table 3.2: MIAS Database

Abnormality Class	Total Benign	Total Malignant	Total Images
Calcifications	12	13	25
Circumscribed Masses	16	4	20
Spiculated Masses	12	9	21
Architectural Distorsions	10	10	20
Asymetries	8	9	17
Miscellaneous	8	7	15
Normals			204
	Totals		322

This work investigates the abnormalities classes of Circumscribed and Spiculated Masses. The database Readme file provides the location of the masses as a pair of values (x, y) that correspond to the location of the mass and a radius r that encloses the abnormality. A program was developed to extract the ROI by using the coordinates and the radius. Consequently, a total of 53 benign and 39 malignant ROI images were obtained by excluding the calcification abnormality category and normal images since they have no coordinates to extract information.

3.4.4 Experimental Procedure

The experimental methodology of this work is partially based on the one proposed by Amiri, Akanbi, and Fazeldehkordi (2014) in their work titled: "*A machine-Learning approach to phishing detection and defense*". There, the authors also address the problem of single models versus ensemble. The experiments carried on this study were conducted in 5 main phases as shown in Figure 3.1.

Phase1: Dataset Generation This phase generates the datasets to be used in the following stages. This phase has 4 main stages: ROI extraction, Pre-processing, Data Augmentation, and Dataset Split Sets generation. The first stage extracts the region of interest from the mammogram image. Then, in the second stage, images are pre-processed and stored in two folders depending on the mass abnormality (benign, malignant). In the third stage, data augmentation is used because there is little data. The library Augmentor (Bloice et al., 2019) was used to create additional images. The last stage consists in forming the train, validation and testing sets. A total of 3 Datasets were created (see Fig. 3.1).

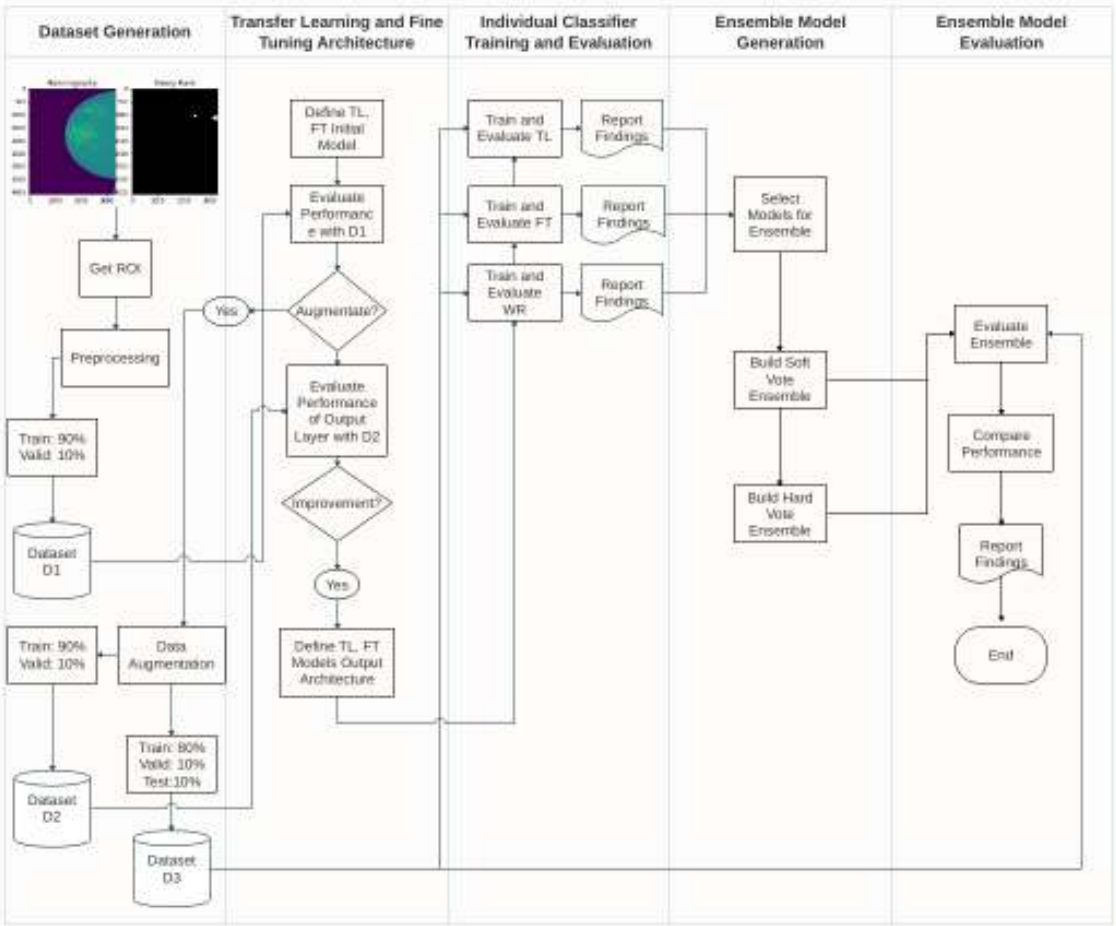

Figure 3.1: Experimental Procedure

ROI extraction Inbreast database has both mammogram and binary masks. By using the provided binary masks, the region of interest images I_{ROI} were generated. For this, a matrix multiplication of the DICOM mammography image with the mask was performed. The boundy box coordinates of the ROI in the resulting image were found by use of Skimage library (**scikit-image**). This permitted to extract mass ROI from the mammogram image. Figure 3.2 shows a mammogram image (**A**), the binary mask (**B**) and the extracted ROIs from the binary mask. Two types of ROI were extracted: without tissue (**C**, **E**) and with tissue or context (**D**, **F**). The procedure followed to obtain the ROI images also produced images with different sizes. In fact, the average width and height of the benign ROI images obtained are: 292 and 278 respectively; and for malignant case: 354 and 355. Getting the average of these values produces a great average of 323×316 (width, height) for the whole set with an average aspect ratio value of 1.15. This distribution is shown in Figure 3.3.

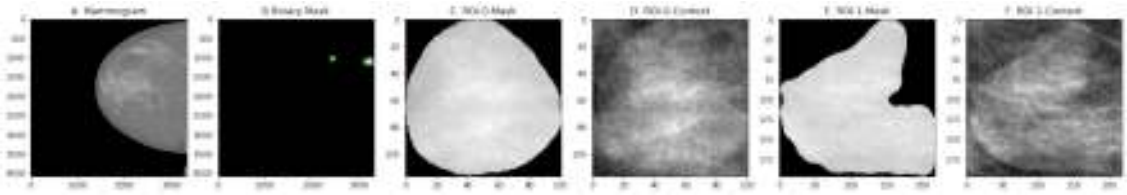

Figure 3.2: Roi extraction procedure

Inbreast ROI patch process. Figure **A**: mamography; **B**: binary mask with two ROIs; **C**: Image ROI-0 without context; **D**: Image ROI-0 conserving tissue; **E** and **F** similar cases for next ROI

In the case of MIAS database, the ROI image was extracted by datamining the ground truth readme file which contains the coordinates of the location of the mass and a radius defined by the specialist. In order to retrieve the locations of the mass and also the pathology class (benign, malignant), a regular expression was used. In the same fashion as with Inbreast database, ROI images for MIAS were generated with different sizes as shown in Figure 3.4. In average, a ROI image had $(410, 407)$ for benign cases and $(496, 485)$; which gave a great average of $(453, 446)$

Pre-processing Inbreast mammogram images are stored in DICOM format in UInt16 bit resolution. This means that pixel values range from 0 to 65 535 in a mammogram. Once the ROI images were obtained, their pixel values were rescaled considering the minimum and maximum values found in I_{ROI} to the range of 0 to 255. It is important to notice that when the image is trained, it is normalized to the range of 0,0 to 1,0. This is done in order to accelerate training process in

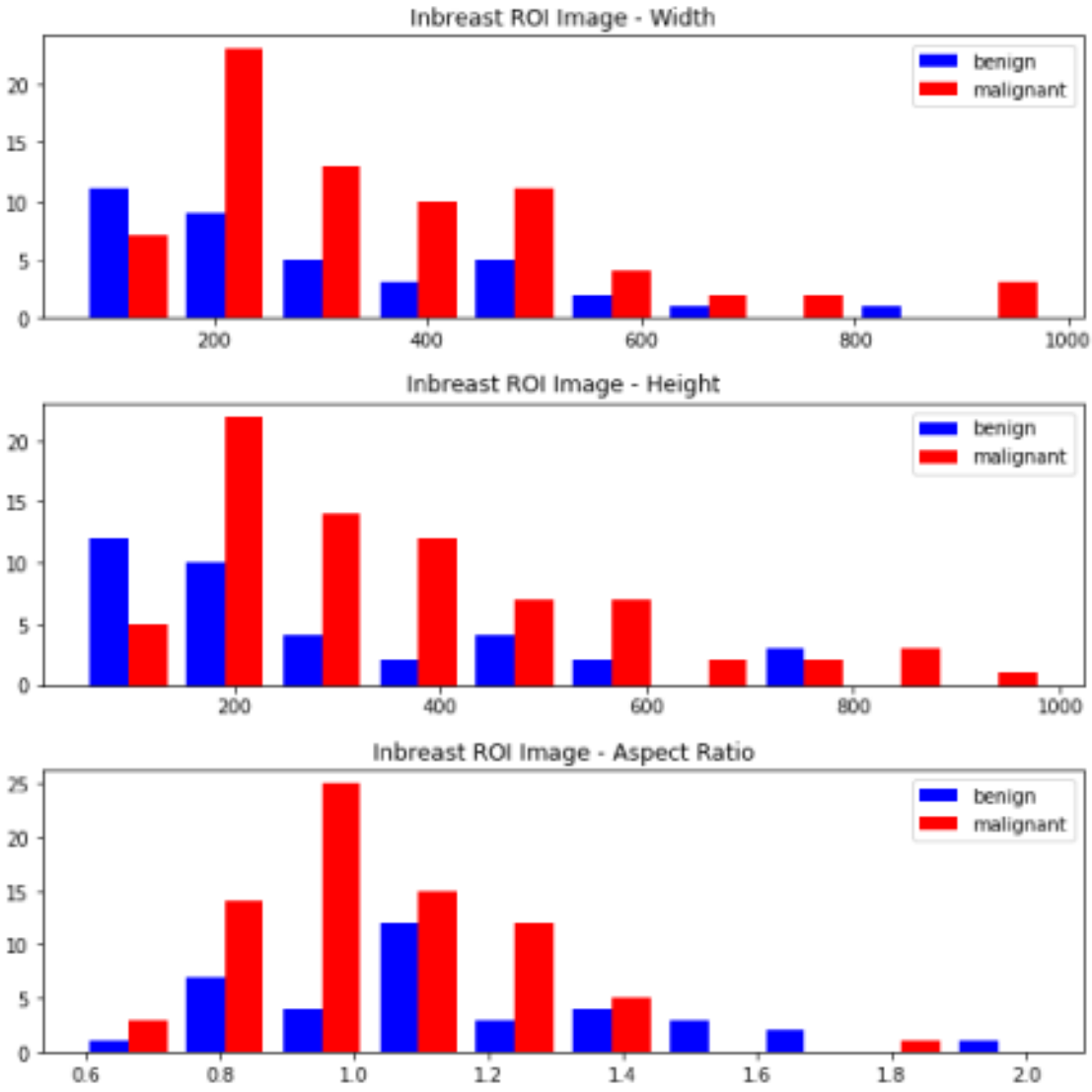

Figure 3.3: Inbreast ROI Images Characteristics

Inbreast ROI images are not of a fixed size at this stage. Averaging the values of width and height through the dataset shows a mean size of 323×316

the ConvNet and it is done when the ConvNet creates the image tensors according to train, validation and testing sets. Our previous work (Falconí et al., 2019) revealed that Otsu binarization and other techniques applied did not contribute essentially in improving the classification results on the CBIS-DDSM dataset. Because of that, the pre-processing of the image in this work was limited to adjusting the image size without affecting the aspect ratio and re-scaling the pixels values. Finally, images were organized in folders according to the pathology, as benign or malignant, considering the BIRADS scale. Values of 1, 2 and 3 were considered benign and 4, 4a, 4b, 4c, 5 y 6 were considered malignant. This created a total of 37 benign cases and 75 malignant cases.

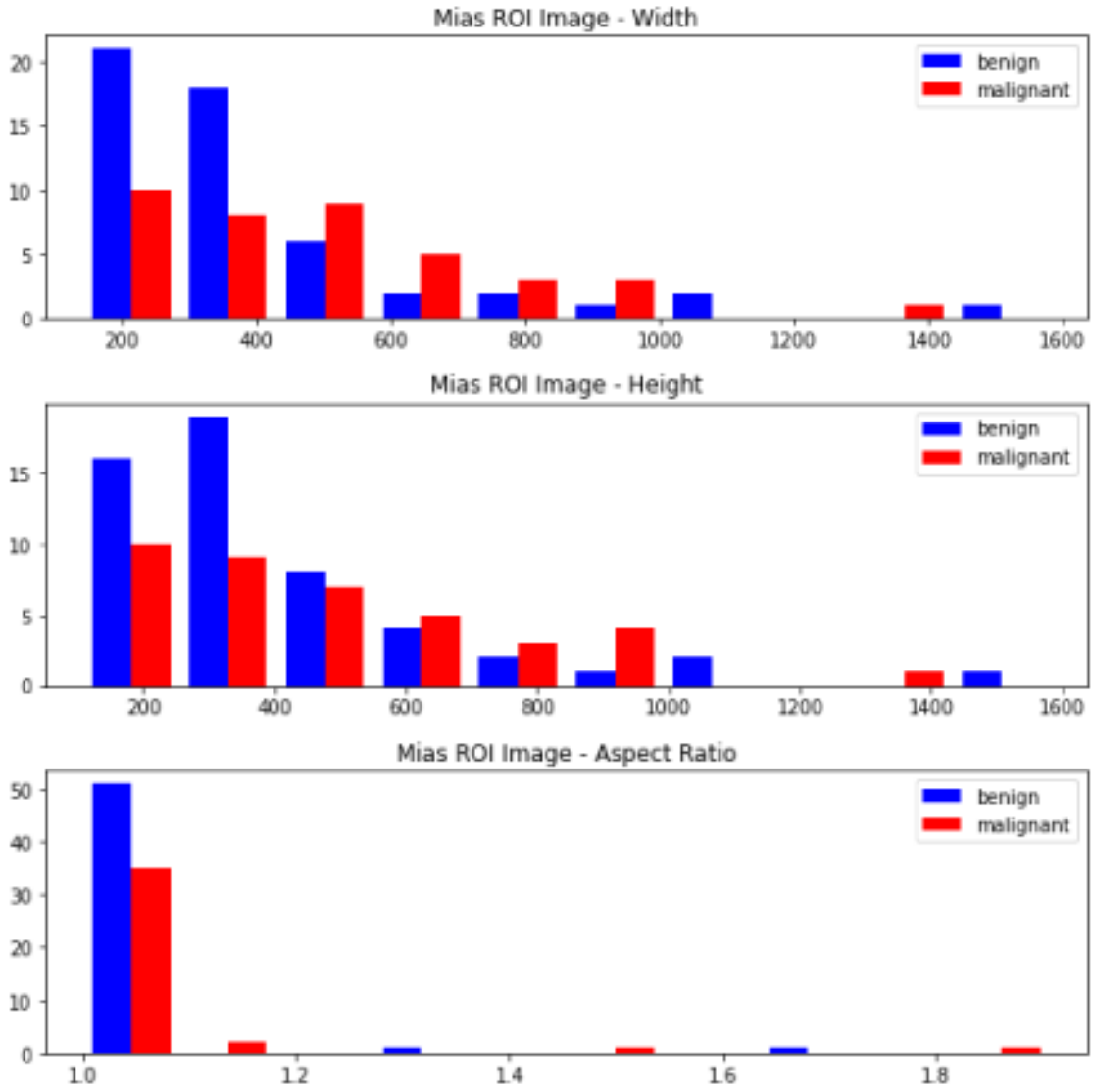

Figure 3.4: MIAS ROI Images Characteristics

MIAS ROI images are not of a fixed size at this stage. Averaging the values of width and height through the dataset shows a mean size of 453×446

In the case of MIAS database, images are stored in PGM file format. The resolution of the images is UInt8, which means that pixels values range from 0 to 255. Similarly to Inbreast, ROI images have different maximum and minimum pixels values. Because of that, the pixels values were also re-scaled according to the minimun and maximum values in the ROI. Moreover, this step enhanced contrast of the ROI image. A total of 53 benign cases and 39 malignant cases were created by this process. The proportion of benign versus malignant is shown in Figure 3.5, for both databases.

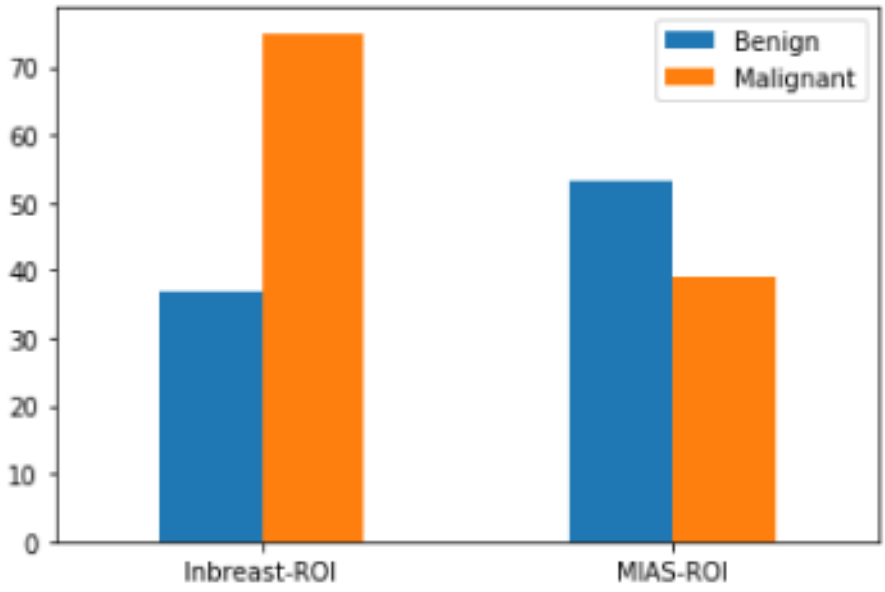

Figure 3.5: Benign and Malignant Cases in ROI Datasets

The number of benign and malignant cases for each database is shown in this Figure. Inbreast has 37 benign cases and 75 malignant. Mias has 53 benign cases and 39 malignant.

Data Augmentation In this step, the total number of samples in the dataset was increased by using data augmentation. Due to the fact that some of the geometrical operations performed on images cause distortions, this work has chosen to use the Augmentor library (Bloice et al., 2019) because it aids in limiting the distortions; specially when rotating the images. This library includes rotations, zoom, histogram equalization and other techniques. To the best of our knowledge, Augmentor has not been used in mammogram classification yet. Through augmentation, a dataset with a total of 3 000 images per category was formed for each database.

Dataset Split This is the final step, where experimental datasets were formed. The first dataset D_1 was formed by splitting the original dataset without data augmentation. Datasets D_2 and D_3 consisted in splitting the augmented dataset generated in the previous stage. However, D_2 was generated splitting the augmented dataset in: train (90%) and validation (10%). This was done in order to validate a first stage in our models training. D_3 is the main experimental dataset that is to be used in this research work and it was created by splitting the dataset source in train (80%), validation (10%), and test (10%) sets respectively. All datasets used images with context (D, F in Fig. 3.2) from both databases. This is because, MIAS dataset does not provide segmentation masks. Table 3.4 shows the generated datasets for experimentation.

Table 3.3: Augmentation Operations

Operation	Probability	Parameters
Rotation	95%	Max Angle 15°
Shear	60%	Max value 25
Histogram Equalization	60%	
Horizontal Flip	70%	
Bright	80%	min value: 0.6 max value: 1.2
Zoom	100%	min value: 1.0 max value: 1.2

Table 3.4: Datasets generated for Experiments

Dataset	Image Type	Data Augmentation	Train-Valid-Test Split	Inbreast	MIAS
D_1	D y F	No	90%, 10%	112	92
D_2	D y F	Yes	90%, 10%	6000	6000
D_3	D y F	Yes	80%, 10%, 10%	6000	6000

Phase2: ConvNets Architecture and training parameters setup In this work, each pre-trained model was evaluated in Transfer Learning, Fine Tuning and Whole Retrain modalities as described in Section 2. Table 3.1 shows the models that were used with the Inbreast dataset. This phase consists in the design of the output layers for Transfer Learning and Fine Tuning by defining the operator $\mathcal{A}$ and two pre-experiments. The first experiment was carried out to validate the need for data augmentation, while the second was used to validate the final output architecture for TL and FT.

Equations (3.5) and (3.6) use the operator $\mathcal{A}$ to define the operations that replace the original FC for each pre-trained ConvNets to adapt it to the target task. In (3.9) and (3.10), the initial configuration of the output layers for Transfer Learning and Fine Tuning are defined.

$$\mathcal{A} = Fc_1 \circ Dro_{0.2} \circ Fc_{1024} \circ GAvg_{2D} \tag{3.9}$$

$$\mathcal{A} = Fc_1 \circ Dro_{0.2} \circ GAvg_{2D} \tag{3.10}$$

Equation (3.9) implies that the last original FC of any pre-trained ConvNet was replaced by a bi-dimensional *Global Average Pooling*(GAvg) layer, followed by a Full Connecting layer of 1024 neurons with Dropout (Dro). The last FC has one neuron and sigmoid activation function for binary classification. This setup is depicted in Figure 3.6.

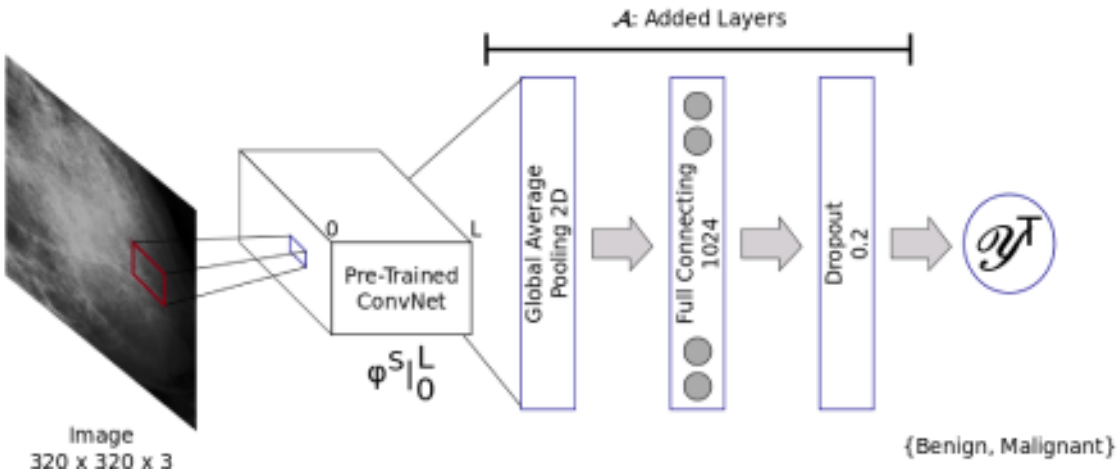

Figure 3.6: Transfer Learning Model for Mammogram Classification

The figure represents the process of Transfer Learning where the last FC was replaced by the operations shown in $\mathcal{A}$

Similarly, (3.10) implies that the last original FC that classifies in 1000 categories was dropped and replaced by bi-dimensional *Global Average Pooling*(GAvg) layer. The output of the GAvg is densly connected to the FC of 1 neuron with a sigmoid function that performs binary classification. Dropout (Dro) was used to prevent overfitting. This is depicted in Figure 3.7.

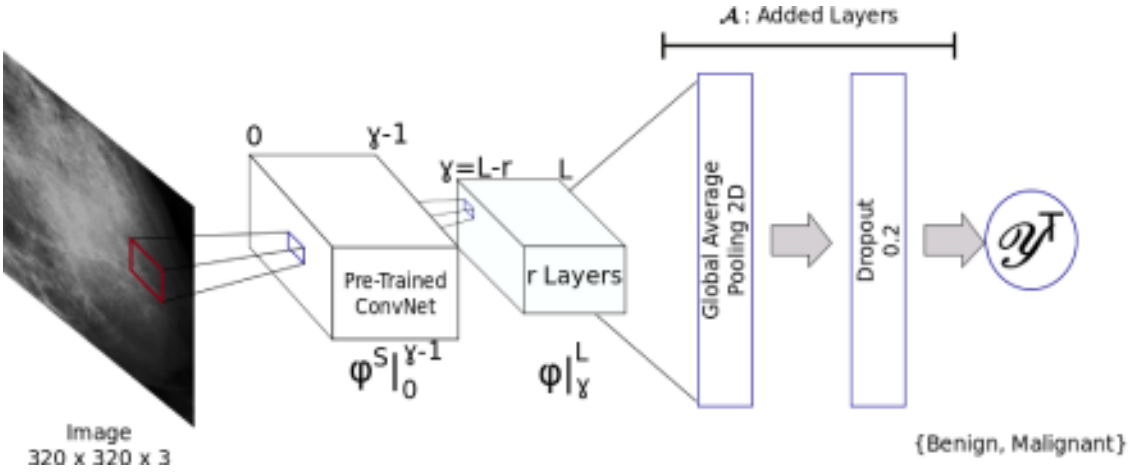

Figure 3.7: Fine Tuning Model for Mammogram Classification

The figure represents the process of Fine Tuning, where γ layers are frozen while the remaining is re-trained with random initialized weights. Additional layers are described by ($\mathcal{A}$)

In the next Chapter, experimental evidence that justifies the need for data augmentation and that validates the final architecture used in this work is presented.

Phase3: Individual Classifier Train and Evaluation: This phase is the main core of the research since it performs TL, FT and WR of 20 pre-trained ConvNets in the ImageNet dataset. Once each classifier was trained, its performance was registered.

Training This procedure used the dataset D_3 to train 20 pre-trained ConvNets in the mammogram classification task by using TL, FT and WR. It is important to notice that the same model used in TL was used for WR with the essential difference that WR does not use pre-trained values. For FT, since different results can be achieved by manipulating the γ layer that splits the ConvNet between the weights of ImageNet and random initialized weights, it was decided to perform at least 3 tests for each ConvNet. This is indicated in Table 3.6 where the different values of γ used are presented for each model trained. Table 3.5 presents the $\mathcal{A}$ operator, the learning rate values, the cost function and optimization algorithm used for training. All models were trained up to 1000 epochs top. Early stopping

with a patience of 50 epochs was used to stop training when the classification performance degrades.

The output layers of $\mathcal{A}$ presented in Table 3.5 are the chosen configuration to experiment. They were validated by a pre-experiment phase whose purpose was to elucidate the number of neurons in the FC layer and the dropout value. This is reported in the Results Chapter. Notice that there are slight differences in the proposed set up of $\mathcal{A}$ for Inbreast and MIAS. This is indicated in Table 3.5. In the case of FT on MIAS dataset, the same output configuration as in TL was used.

Finally, all images were resized to be trained in TL, FT and WR. The resize of the image was performed automatically during training but the size used differs for NasNet models. In fact, all models except *NasNet Large* and *NasNet Mobile* used an input layer size of $320 \times 320 \times 3$. NasNet Large used $331 \times 331 \times 3$ and NasNet Mobile, $224 \times 224 \times 3$.

Table 3.5: Proposed Training parameters

Technique	**Added Layers $\mathcal{A}$**	**Learning Rate**	**Cost Function**	**Optimization Alg.**
		Inbreast		
Transfer Learning	$Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D}$	1×10^{-5}	binary cross-entropy	RMSProp
Fine Tuning	$Fc_1 \circ Dro_{0.2} \circ GAvg_{2D}$	2×10^{-7}	binary cross-entropy	RMSProp
Whole Re-Train	$Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D}$	1×10^{-5}	binary cross-entropy	RMSProp
		MIAS		
Transfer Learning	$Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D}$	1×10^{-5}	binary cross-entropy	RMSProp
Fine Tuning	$Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D}$	2×10^{-7}	binary cross-entropy	RMSProp
Whole Re-Train	$Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D}$	1×10^{-5}	binary cross-entropy	RMSProp

Table 3.6: Fine-Tuning Experiments

Modelo	**Fine Tuning en Profundidad** γ
Vgg16	10, 14, 16
Vgg19	11, 20, 17
DenseNet-121	325, 422, 425
DenseNet-169	490, 590, 593
DenseNet-201	600, 702, 705
Inception-Resnet-v2	675, 775, 778
Inception-V3	209, 306, 309
Mobilenet	77, 82, 85
Nasnet Large	937, 1034, 1037
Nasnet Mobile	669, 764, 767
Resnet-50	160, 170, 173
Resnet-50-v2	180, 185, 187
Resnet-101	330, 340, 343
Resnet-101-v2	365, 372, 375
Resnet-152	500, 510, 513
Resnet-152-v2	550, 559, 562
Resnext-50	230, 234, 237
Resnext-101	460, 470, 472
Xception	127, 130, 137

Model Evaluation Once training was finished, each model was evaluated with the test set of D_3. The measurements and metrics used are presented in Table 3.7. Results were stored in csv and json files for reporting, documenting and analysis. This study used three different levels of metrics by following the advice in Canbek et al. (2017). In the basic level are: accuracy and area under the ROC curve. For the first level, F1 score and balanced accuracy were used. Mathews correlation coefficient was selected as a second level metric. Confusion matrix was used as the base measurement tool to evaluate the performance of each model.

Table 3.7: Used Measurements and Metrics

	Basic	**First Level**	**Second Level**
Measurement	Confusion Matrix (TP, TN, FP, FN)		
Metric	ACC, AUC	F1 score, BACC	MCC

Phase4: Ensemble Model Generation This phase consisted in selecting the best individual classifiers obtained through Transfer Learning, Fine Tuning, and Whole Retrain techniques. Two ensemble algorithms were proposed: hard voting and soft voting. The former consists in a majority vote of each predictor. The most voted category is the one accepted. Hard voting is based on the knowledge of the masses concept. Soft voting is implemented through the probability estimation of the class prediction by the classifier. It consists in averaging the probabilities calculated by each classifier. In order to recommend the weights values for the average sum performed in the soft voting algorithm, this thesis proposes to use a Perceptron to search for the best recommended values. This new model is called *Automatic Soft Voting Ensemble* (ASVE). Hard and soft voting were implemented by using the definitions discussed in Chapter 2, Section 2.7 and the equations: (2.21), (2.22). The *Automatic Soft Voting Ensemble* was implemented using (2.22) and a Perceptron to find the values of w_j.

It is important to notice that ensemble methods require diversity of the individual classifiers that conform them. This diversity must be along all possible problem features: data and architecture. This work did not consider the use of bagging to create different training datasets. Instead, this work used different models

trained with different techniques. Moreover, two independent datasets (MIAS and Inbreast) were used to achieve uncorrelated errors.

Phase5: Ensemble Evaluation In order to evaluate the performance of the ensemble, the test set of D_3 was used. Each ensemble makes a prediction internally using the individual classifiers. The performance of classification was registered in the same fashion as in the case of single classifiers evaluation. The same set of metrics were used to evaluate the Ensemble Model.

3.4.5 Metrics and Measurements

In order to determine the performance of each individual classifier, the use of performance metrics is required. This work considers the observation by Canbek et al. (2017) about the distinction between metric and measurement in binary classification. In their work, the authors propose a total of 22 measurements and 22 metrics that constitute, what the authors call, the periodic table of binary classification. From these 22 measurements and metrics, the authors recommend the use of *F-measures*, the Mathews Correlation Coefficient (MCC) and the Cohen's Kappa (CK). The choice of measurements and metrics to evaluate the classifiers performance in breast cancer is of relevant importance, because the distinction between benign and malignant classes is delicate. For instance, False Positives (FP) and False Negatives (FN) detriment the classifier prediction. In fact, according to Canbek et al. (2017), FN correspond to a type 2 error. This type of error corresponds to omission or sub-estimation, and are generally considered more serious or worse in medicine and engineering: A FN error classifies a true malignant or cancer mass as benign. In contrast, FP errors correspond to type 1; which is called over-estimation (i.e. a true benign mass is classified as malignant).

For this research work, the Confusion Matrix was used to measure the performance of each classifier. From its basic measures, metrics are defined and calculated. These metrics were implemented in the evaluation program. In this section, the metrics and measures used are described.

Confusion Matrix

A confusion matrix is formed by: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Other important measures are: True Classification ($TC = TP + TN$), False Classification ($FC = FP + FN$), Sample Size $Sn = P + N$, Positive Output ($OP = TP + FP$) and Negative Output ($ON = TN + FN$). Therefore, positive cases are $P = TP + FN$ and the negative cases: $N = TN + FP$.

Accuracy (ACC)

This metric is defined in (3.11) and calculates the proportion of correct predictions to the size of the sample. Even though it is the most common metric used, it is not recommended because it is easily distorted; specially when the dataset is imbalanced (*skewed datasets*). This behavior is known as *Accuracy Paradox* because a high value of ACC does not imply a correct performance (Valverde-Albacete & Peláez-Moreno, 2014).

$$ACC = \frac{TC}{Sn} = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + TN} \tag{3.11}$$

F_1 Score

This metric is defined in (3.14) as the harmonic mean of PPV and TPR. It uses the *Precision* (or *Positive Predictive Value*) (3.12) and *recall* (or Sensitivity, or *True Positive Rate*) (3.13).

$$PPV = \frac{TP}{OP} = \frac{TP}{TP + FP} \tag{3.12}$$

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{3.13}$$

$$F_1 = \frac{2}{\frac{1}{PPV} + \frac{1}{TPR}} = \frac{2PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FC} \tag{3.14}$$

Area under the ROC Curve (AUC)

The Receiver Operating Characteristic Curve plots the *true positive rate* (TPR) (3.13) as a function of the *false positive rate* (FPR) (3.15). Considering the *true negative rate* (TNR), it can be concluded that $FPR + TNR = 1$ and, therefore, the ROC curve establishes a relation between sensitivity and $1 - TNR$. The area of the ROC Curve is calculated as indicated in (3.16).

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{3.15}$$

$$AUC = \frac{1}{2}\left(1 + TPR - FPR\right) \tag{3.16}$$

Balanced Accuracy (BACC)

This first level metric is defined in (3.17), which is interpreted as the average of accuracy in either class (Brodersen, Ong, Stephan, & Buhmann, 2010). If the classifier predicts each class correctly, in the same proportion, then (3.17) returns the traditional value of ACC as defined by (3.11). However, if there exists an unbalance, BACC decreases.

$$BACC = \frac{TPR + TNR}{2} \tag{3.17}$$

Where TNR is defined in (3.18).

$$TNR = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{3.18}$$

Matthews Correlation Coefficient (MCC)

It is defined in (3.19) and measures the performance of a classifier. The coefficient has a range of $[-1, 1]$; If $MCC = 1$, the model's prediction is completely correct; If $MCC = -1$, the model is completely wrong. When denominator and numerator are both 0, $MCC = 0$.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{P \cdot OP \cdot N \cdot ON}} \tag{3.19}$$

3.5 Limitations and potential problems

In this Section, threats to validity of this research work are discussed. In any research, it is important to evaluate possible threats that limit the research performed. Inevitably, the first limitation is related to the number of pre-trained ConvNets used. In order to address this problem, all the available models in the Keras API in Tensorflow 1.13.1 were used. Moreover, a comparison among 20 different pre-trained models has not been found in literature. Therefore, this work contributes to expand the impact of using Transfer Learning, Fine Tuning and Whole Retrain in mammogram classification. Also, because TL and FT were compared wrt. WR, this research is not biased by the accepted proposition that training a ConvNet with random initialized weights is worse than training it in TL or FT.

Another threat is the manipulation of the output layer described in this work by the operator $\mathcal{A}$. The output layers used to replace the original FC layer can be arranged in many ways; including using new convolutional layers, FC, etc. A change in the structure of $\mathcal{A}$ can affect the performance of the classification. This was observed and registered in Phase 2 of the Experimental Procedure. There, dataset D_2 was used to test some configurations of the FC layer in order to find the one that achieves the best results for TL. Experiments in Phase 2 were limited by time constrains and because of that only dropout and the number of output neurons in the last FC were reviewed. Other parameters like learning rate can be further explored. Phase 2 shows that care was taken to address possible different $\mathcal{A}$ configurations in an acceptable manner by providing an empirical justification.

In the case of training NasNet Large ConvNet on MIAS dataset by using Whole Retrain modality, the experiments were compelled to use less neurons in the FC layer due to GPU capacity. Even though the program was designed to train models using multi-GPU, NasNet Large requires more than 2 GPUs to be processed without warnings. This was a technological limitation.

The exclusive use of ROI images can be also argued to threat validity. Most of the research tries to process whole mammogram images or provide the whole stages of detection, segmentation and classification. Since, processing a Whole mammogram image is expensive in computation resources, and other researchers also concentrate on the ROI images, this research has focused on this problem.

A key aspect regarding cancer prediction is the fact that only masses were considered in the present study when Micro calcifications are also required to be analyzed in order to confirm malignancy through the mammogram exam. In this case, the decision to use mass images was regarding the scope of work defined, leaving micro calcifications for a future work.

It is recognized that models in this work were not trained several times due to time constrains and computer resources. This limited the possibility to estimate an average of each classification metric. ConvNets training in this work can take as long as 1700 seconds per epoch and convergence was usually achieved in most cases above the 100^{th} epoch. Nevertheless, this work implemented several metrics and compared the results provided by each of them.

A final potential problem is the ensemble model because it does not use Bagging. As discussed in Chapter 2, Section 2.7, ensemble methods aim to improve the generalization capabilities of a set of individual classifiers and Bagging is a common method to re-sample the training set. Instead, two independent different sets (one for film mammography (MIAS) and one for full field digital mammography (Inbreast)) were used. Furthermore, different ConvNets were trained in different modalities and techniques to account for diversity of the classifiers used; and most of all, independence between each base predictor. Also, two different ensemble techniques (hard and soft voting) were used to provide with enough diversity in this work. The aforementioned factors, as far as possible, reduce possible correlated errors and increase diversity.

In the next chapter, the experimental results of this research are presented. The experiments carried out follow the methodology discussed in this chapter.

Chapter 4

Experimental Results of Transfer Learning, Fine Tuning, Whole Retrain and Ensemble Learning of 20 models in Mammogram Classification

4.1 Introduction

Transfer learning in convolutional neural networks has three main advantages: first, the pre-trained weights assure that the early layers of the ConvNet extract general features that are similar to all images (i.e. edge detection) (Koehrsen, 2018); second, they aid to mitigate the overfitting problem, because not all the layers are re-trained; third, because only a few group of layers at the end of the ConvNet are trained, training time is reduced compared to training the whole network. The ImageNet contest (Deng et al., 2009) has provided with state of the art models for image classification as well as object detection. These models trained on natural images are publicly available for research purposes. Literature review reveals that TL is a commonly used technique in mammogram classification. However, just some models have been explored; being AlexNet regularly tested in literature (see Figure 2.7). Consequently, this research proposed to fill the gap by experimenting with 20 pre-trained ConvNets on Imagenet that were fine-tuned on two mammogram datasets (Inbreast and MIAS) in the mass abnormalities binary classification task (benign, malignant). Furthermore, an ensemble

of the best fine tuned models was implemented to increase performance.

As pointed out in Section 1.2 of Chapter 1, the aim of this research is to design a classification model for mass breast lesions using transfer and ensemble learning, to increase literature average performance. Hence, this research experimented on manipulating the intervention (TL, FT, Ensemble) to increase the performance on mammogram classification. In this chapter, the experimental findings of this work are presented. The experimental procedure was carefully detailed in Chapter 3, Section 3.4.4 and represented graphically in Figure 3.1; which aims to work as a road map of the procedure.

It is known from the Discussion Section 2.8.5 in the Literature Review Chapter 2 that research in mammogram pathology classification is active. There are several works that have presented interesting results. That is the case of Chougrad et al. (2018) whose work has achieved $AUC = 0.97$ and $AUC = 0.99$ on Inbreast and MIAS, respectively. This work is similar to that of Chougrad et al. But there are some differences: First, this work studied three training modalities (TL, FT and WR); distinguishing between each of them. Second, the number of pre-trained ConvNets was expanded to 20 models; which corresponds to all the available models in Tensorflow-Keras v1.13.1. Third, in addition to TL, FT and WR, ensemble learning models through hard voting and soft voting were implemented and their results reported. Forth, additional metrics were used to evaluate each experiment by considering the observations proposed in Canbek et al. (2017). Therefore, it is hoped that this work contributes in aiding to fill the gap of potentially unexplored ImagenNet pre-trained ConvNets and providing a systematic comparison of their performance on mammogram classification. Lastly, an adapted formal definition of TL and FT for the case of ConvNets was proposed.

The remaining of the chapter is divided in two parts: a) Pilot Experiments and b) Proper Experiments. The former presents experiments that were carried out to improve some of the parameters of the ConvNets used in experimentation. The latter, presents the results achieved in TL, FT, and WR on D_3 in Inbreast and MIAS by using the best found combination of parameters. The datasets used in each type of experiment are described in Table 3.4. Pilot experiments use datasets D_1 and D_2, while datset D_3 is used for Proper Experiments; D_3 was split in train (80%), valid(10%) and test (10%) for this purpose. Remind that D_1 was

not artificially augmented. The values used for learning rate, optimization algorithm, cost function and operator $\mathcal{A}$ are presented in Table 3.5.

4.2 Pilot Experiments

The experiments in this section were called *Pilot Experiments* because they allowed to establish specific values and procedures that were used later on the *Proper Experiments*. Two experiments were developed in this section: the first one used D_1 on Inbreast to test the performance of Transfer Learning without data augmentation. The second one aimed to compare the effect of changing the number of neurons in the last FC layer and the dropout value on the classification performance.

4.2.1 Transfer Learning in Inbreast without Data Augmentation

Figure 4.1 compares the performance of 20 pre-trained ConvNets on TL with respect to accuracy (ACC), balanced accuracy ($BACC$), area under the ROC curve (AUC), F_1Score and Mathews Correlation Coefficient (MCC). Table 4.1 shows the top 5 results of Transfer Learning on D_1. For a review on the selected Metrics, the reader may refer to Section 3.4.5.

It is observable that Densenet-169 presents $MCC < 0$, which means it is not suitable for classification. Remind that a negative MCC value means that the classifier model does not work. On the other hand, the best model found in this experiment corresponded to ResNet50-v2, which achieved $AUC = 0.685$. The use of the MCC helps to graphically discard models that will not classify mammogram mass accurately despite their accuracy value. This confirms that it is important to use more than one metric when studying binary classification on mammogram. Furthermore, the accuracy paradox is confirmed, because Densenet-169 has a positive above 50% accuracy. Other models present $MCC = 0$ and can also be discarded.

Table 4.1: Transfer Learning Top 5 Results on Dataset D_1

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
resnet-50v2	0.685	0.685	0.816	0.401	0.743
resnet-152v2	0.667	0.667	0.852	0.497	0.771
mobilenet	0.601	0.601	0.784	0.241	0.686
nasnet-l	0.583	0.583	0.821	0.341	0.714
resnet-101v2	0.562	0.562	0.8	0.209	0.686

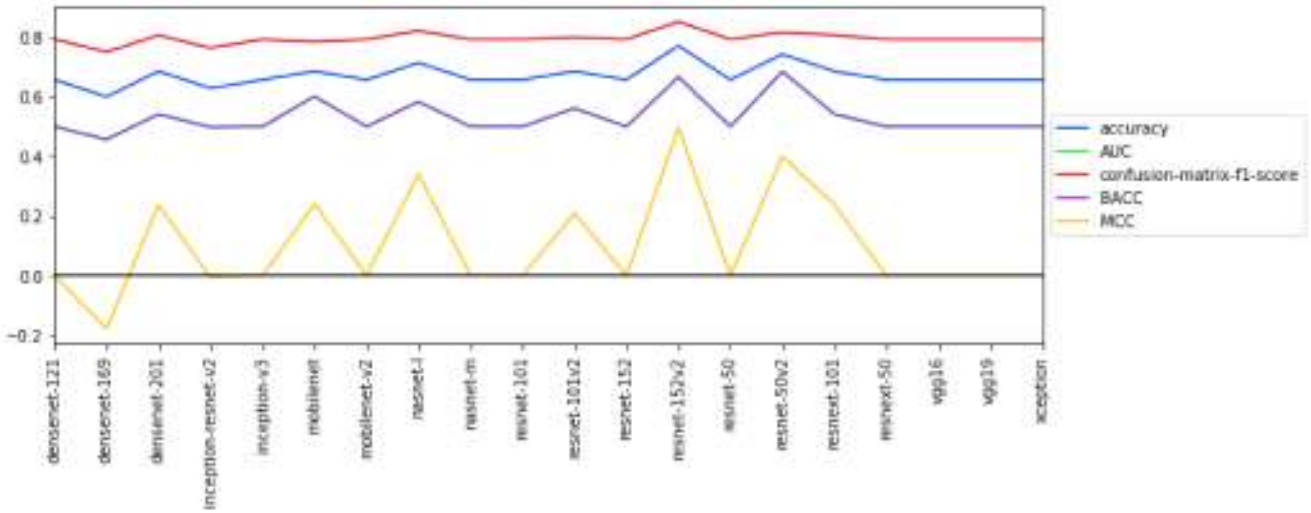

Figure 4.1: Transfer learning for Inbreast on Dataset D_1

4.2.2 Transfer Learning in Inbreast with Data Augmentation

This experiment compared the performance of 20 pre-trained ConvNets on TL wrt. ACC, $BACC$, AUC, F_1Score, and MCC using an artificially augmented dataset D_2. Data augmentation was described in Section 3.4.4 and it mainly consisted in using the Augmentor Library to increase the dataset by the operations described in Table 3.3. Figure 4.2 shows that all metrics improve by using data augmentation. In fact, there is no $MCC < 0$. Table 4.2 presents the top 5 results of Transfer Learning on D_2. Compared to Table 4.1, classification metrics values improved. Also, the best model corresponds to Mobilenet with $AUC = 0.908$. Globally, it can be said that the measurements are more stable than in the case without data augmentation. Therefore, this experiment confirms the need to use data augmentation or at least 3000 images per category.

Table 4.2: Transfer Learning Top 5 Results on Dataset D_2

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
mobilenet	0.908	0.908	0.909	0.817	0.908
resnet-152	0.903	0.903	0.906	0.808	0.903
resnet-50v2	0.897	0.897	0.895	0.794	0.897
resnet-101	0.895	0.895	0.892	0.791	0.895
resnext-101	0.883	0.883	0.884	0.767	0.883

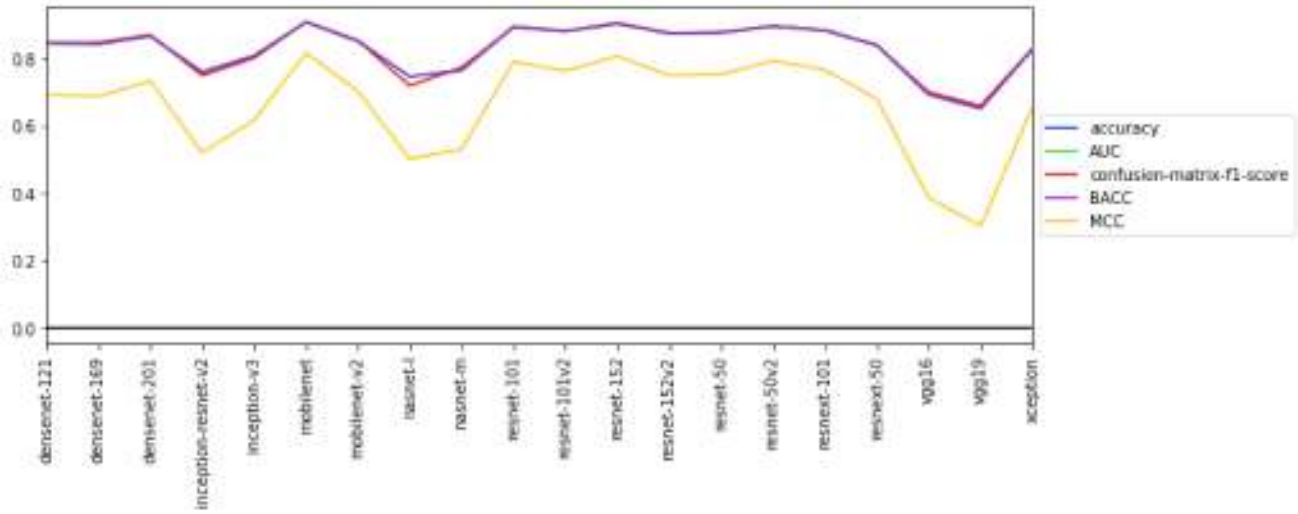

Figure 4.2: Transfer learning for Inbreast on Dataset D_2

4.2.3 Number of Output Neurons in FC and Dropout

The previous experiment confirmed that data augmentation stabilizes the prediction across all the ConvNets trained. Also, it increases classification performance. However, how many output neurons are good enough? Does dropout value affect classification performance? To address this question an experiment was performed using Mobilenet ConvNet where the number of output neurons in the last FC and the value for the dropout layer are changed. Table 4.3 shows that there is an improvement in the classification results if the number of neurons is increased from 1 024 to 8 192. Because the AUC value is very similar for the case with 8 192 and 4 096, it was decided to use 4 096 neurons because using less neurons implies less parameters to train and improves the use of computational resources.

Once selected the 4 096 as the number of neurons to be used in the last FC, an experiment was carried out to check the classification performance wrt. dropout. Table 4.4 shows that classification performance decreases as dropout value is augmented. The best classification was achieved for dropout 0.4. However, since there is little difference with the AUC value achieved using 0.2 value, it was decided to use the latter as the value for the dropout Layer for experimenting.

Table 4.3: Modifying the number of neurons in FC in $\mathcal{A}$ for TL on D_2

$model$	AUC	$FCNeurons$	$Dropout$
mobilenet	0.857	128	0.2
mobilenet	0.877	256	0.2
mobilenet	0.902	512	0.2
mobilenet	0.902	1024	0.2
mobilenet	0.920	2048	0.2
mobilenet	0.930	4096	0.2
mobilenet	0.932	8192	0.2
mobilenet	0.908	1024	0.2
mobilenet	0.903	1024	0.4
mobilenet	0.890	1024	0.5
mobilenet	0.865	1024	0.7
mobilenet	0.783	1024	0.9

Table 4.4: Modifying the dropout rate in $\mathcal{A}$ for TL on D_2

$model$	AUC	$FCNeurons$	$Dropout$
mobilenet	0.925	4096	0.2
mobilenet	0.927	4096	0.4
mobilenet	0.910	4096	0.5
mobilenet	0.900	4096	0.7
mobilenet	0.840	4096	0.9

4.3 Transfer Learning Experiments

The output configuration for the transfer learning experiments is shown in Equation(4.1), where the FC layer has 4096 neurons and the dropout rate has been set to 0.2. The pilot experiments carried out in the last section provided these values. From this section on, the D_3 dataset for Inbreast and MIAS was used for experimentation. This section describes the performance of 20 pre-trained ConvNets in TL for mammogram binary classification on each Dataset.

$$\mathcal{A} = Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D} \tag{4.1}$$

4.3.1 Results of Transfer Learning Training on Inbreast

The Inbreast dataset is formed by Full Field Digital Mammogram images. The dataset here used was augmented with Augmentor (Bloice et al., 2019) and split in three sets: train (80%), valid (10%) and test (10%). Figure 4.3 plots the performance classification metrics against each trained model. As can be seen from the picture, TL improves classification performance. All metrics, except MCC have a similar behavior but with slightly numeric differences. Furthermore, $MCC > 0$ for all models in this experiment. If achieving $AUC > 0.90$ implies that the trained model is a good predictor, a total of 15 out of 20 models could be considered good predictors based on AUC metric.

Table 4.5 presents the value of ACC, $BACC$, MCC, F_1Score and AUC for all 20 ConvNets trained in TL in descending order. Therefore, Mobilenet achieved the best performance for this set of experiments with $AUC = 0.947$; followed by ResNet-50 in second place. The letters TL at the end of each model's name indicate that the model was trained using TL according to the definition proposed for it in Section 2.4 of Chapter 2.

Figure 4.4 shows the behavior of train and validation accuracy as a function of the training epochs. The early stopping finished training around epoch 250 to prevent overfitting. The ROC Curve and the Confusion matrix are shown in Figures 4.5 and 4.6, respectively. Both indicate that Mobilenet trained in TL with 4 096 neurons a dropout of 0.2 and $\mathcal{A}$ defined in (4.1) is capable of classifying a mass as benign and malignant.

Table 4.5: Inbreast Transfer Learning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
mobilenet-TL	0.947	0.947	0.947	0.893	0.947
resnet-50-TL	0.945	0.945	0.944	0.89	0.945
resnet-50v2-TL	0.94	0.94	0.939	0.881	0.94
resnet-101v2-TL	0.937	0.937	0.937	0.873	0.937
resnext-101-TL	0.933	0.933	0.935	0.867	0.933
resnet-152-TL	0.933	0.933	0.934	0.867	0.933
resnet-101-TL	0.93	0.93	0.932	0.861	0.93
densenet-169-TL	0.928	0.928	0.93	0.857	0.928
mobilenet-v2-TL	0.927	0.927	0.925	0.854	0.927
densenet-201-TL	0.927	0.927	0.924	0.856	0.927
densenet-121-TL	0.922	0.922	0.918	0.846	0.922
nasnet-l-TL	0.917	0.917	0.916	0.834	0.917
resnet-152v2-TL	0.913	0.913	0.915	0.827	0.913
xception-TL	0.908	0.908	0.91	0.817	0.908
resnext-50-TL	0.908	0.908	0.91	0.817	0.908
nasnet-m-TL	0.883	0.883	0.878	0.77	0.883
inception-v3-TL	0.88	0.88	0.876	0.762	0.88
inception-resnet-v2-TL	0.843	0.843	0.848	0.688	0.843
vgg16-TL	0.813	0.813	0.81	0.627	0.813
vgg19-TL	0.81	0.81	0.818	0.622	0.81

4.3.2 Results of Transfer Learning Training on MIAS

The MIAS dataset is formed by Film mammogram images. The dataset was augmented with Augmentor (Bloice et al., 2019) and split in three sets: train (80%), valid (10%) and test (10%). The results presented in Figure 4.7 and Table 4.6 consider the metrics of ACC, $BACC$, MCC, F_1Score and AUC. The best performance was obtained on Resnet-101-v2 with $AUC = 0.952$. Considering that $AUC > 0.90$ as a good predictor, similarly to the case of Inbreast, a total of 15 out of 20 models perform above 90% based on AUC metric.

Figure 4.8 shows the behavior of train and validation accuracy as a function of the training epochs. Early stopping finished training around epoch 325 to prevent

Figure 4.3: Transfer learning for Inbreast on Dataset D_3

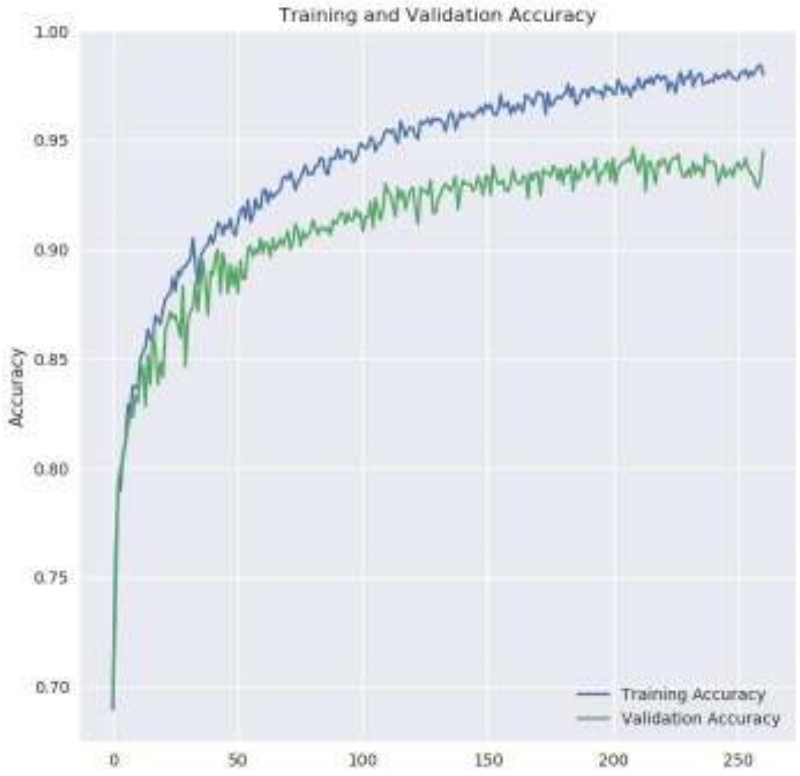

Figure 4.4: Inbreast Mobilenet-TL Training Accuracy Curve

overfitting. The ROC Curve and the Confusion matrix are shown in Figures 4.9 and 4.10, respectively. Resnet-101-v2 is able to predict mass pathology based on the proposed methodology.

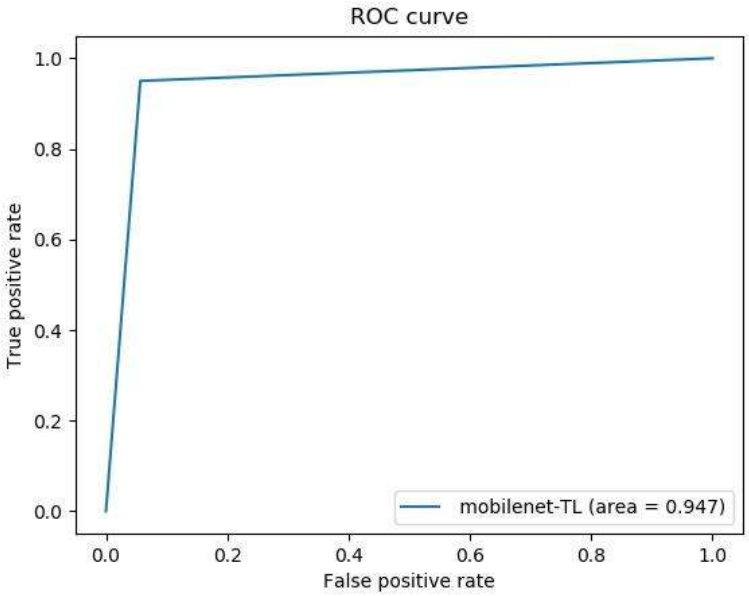

Figure 4.5: Inbreast Mobilenet-TL ROC Curve

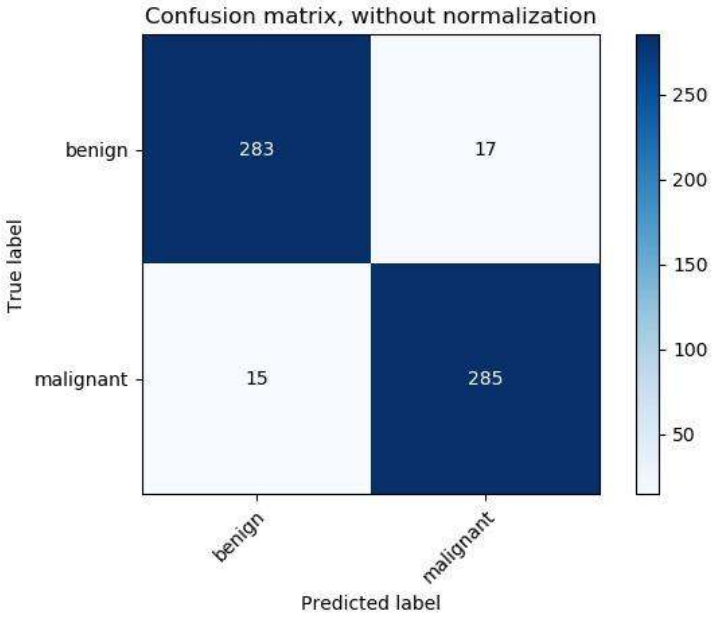

Figure 4.6: Inbreast Mobilenet-TL Confusion Matrix

Table 4.6: MIAS Transfer Learning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
resnet-101v2-TL	0.952	0.952	0.952	0.903	0.952
mobilenet-TL	0.943	0.943	0.944	0.887	0.943
resnext-50-TL	0.94	0.94	0.939	0.88	0.94
densenet-169-TL	0.94	0.94	0.94	0.88	0.94
resnet-152-TL	0.938	0.938	0.938	0.877	0.938
densenet-201-TL	0.937	0.937	0.938	0.875	0.937
resnet-101-TL	0.937	0.937	0.937	0.873	0.937
resnet-50v2-TL	0.935	0.935	0.936	0.87	0.935
resnet-50-TL	0.93	0.93	0.932	0.862	0.93
xception-TL	0.93	0.93	0.932	0.861	0.93
resnext-101-TL	0.93	0.93	0.931	0.86	0.93
densenet-121-TL	0.918	0.918	0.918	0.837	0.918
resnet-152v2-TL	0.912	0.912	0.914	0.824	0.912
nasnet-l-TL	0.912	0.912	0.914	0.825	0.912
mobilenet-v2-TL	0.9	0.9	0.899	0.8	0.9
inception-v3-TL	0.892	0.892	0.893	0.783	0.892
nasnet-m-TL	0.842	0.842	0.845	0.684	0.842
inception-resnet-v2-TL	0.813	0.813	0.816	0.627	0.813
vgg16-TL	0.782	0.782	0.797	0.57	0.782
vgg19-TL	0.723	0.723	0.721	0.447	0.723

Figure 4.7: Transfer learning for MIAS on Dataset D_3

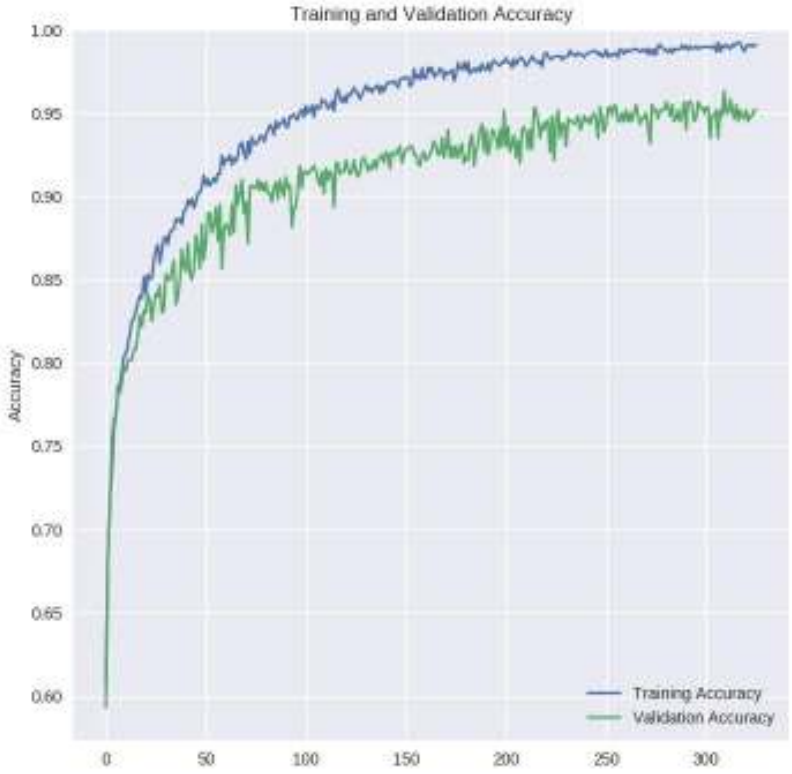

Figure 4.8: MIAS Resnet-101-v2-TL Training Accuracy Curve

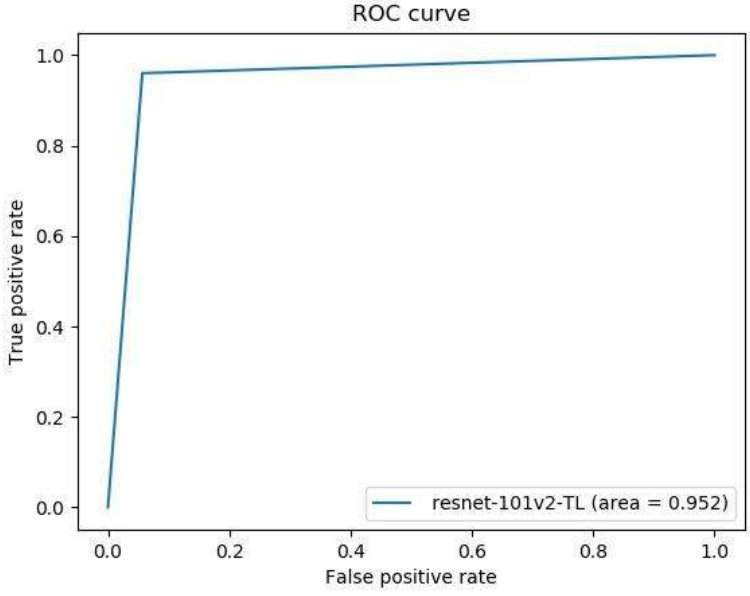

Figure 4.9: MIAS Resnet-101-v2-TL ROC Curve

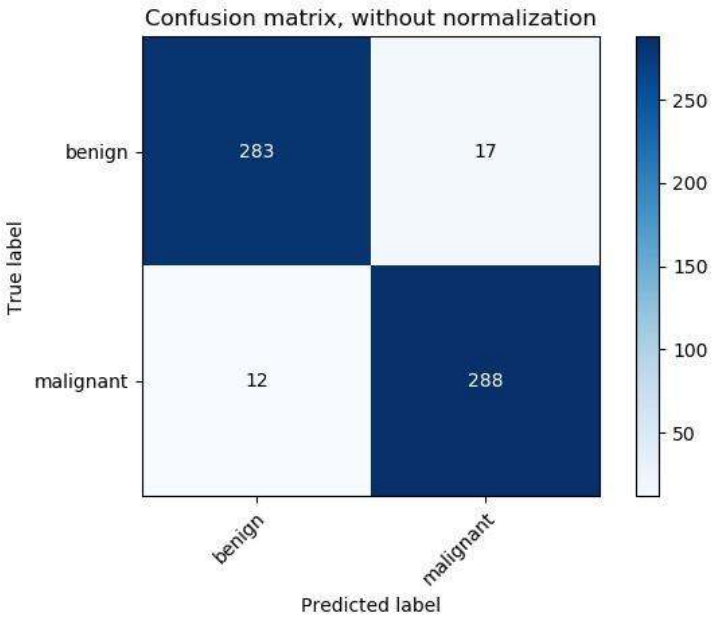

Figure 4.10: MIAS Resnet-101-v2-TL Confusion Matrix

4.4 Fine-Tuning

The output configuration for the Fine Tuning experiments is shown in (4.2) and (4.3) for Inbreast and MIAS, respectively. Notice that the $\mathcal{A}$ operator for Inbreast did not use a FC. On the other hand, MIAS does include a FC of 4096 neurons. Dropout was used in both models with a value of 0.2. As mentioned earlier, the experiments in this section used D_3 dataset on Inbreast and MIAS.

Fine Tuning experiments contain a total of 60 experiments because 3 different values for γ were used in each ConvNet model. This was done to look for improvement of the overall classification by affecting the deepness of the fine tuned model. Remind that FT allows to split the ConvNet training by dividing the model in two parts. The first part, from the initial layer to $\gamma - 1$, freezes the weight's values. The second part, which includes $\mathcal{A}$, is re-trained. Because there is no way to foreseeing a recommended γ value to improve classification performance, this research tried three different values by increasing the deepness level. For the sake of readability only the top 5 results are shown in the following sections for the experiments carried out on Inbreast and MIAS. However, the complete Table with all the tested models is at the Appendix A.

$$\mathcal{A}_{inb} = Fc_1 \circ Dro_{0.2} \circ GAvg_{2D} \tag{4.2}$$

$$\mathcal{A}_{mias} = Fc_1 \circ Dro_{0.2} \circ Fc_{4096} \circ GAvg_{2D} \tag{4.3}$$

4.4.1 Results of Fine Tuning Training on Inbreast

Figure 4.11 plots the 20 fine tuned ConvNets wrt. ACC, $BACC$, MCC, F_1Score and AUC. The convention used to name the trained models is: *model name-version-*γ-FT. The letters FT stand for Fine Tuning which is the intervention used in these experiments. The version parameter can be a letter or a number (e.g. Mobilenet-v2, Densenet-169). The γ indicates the number of layer that splits between freezing weights values and re-training them. Thus, densenet-121-325-FT means that the 121 layered version of densenet was Fine Tuned at $\gamma = 325$. The densenet-121 version implemented in Tensorflow-Keras has a total of 427 layers. Hence, the number displayed in the version field is according to the given name for the model in literature. As can be seen from Figure 4.11, performance varies among all the tested models. Some models improve achieving high peaks

whereas others decrease in performance. It is true to say that classification performance relies on the γ value that is used to fine tune the network. Furthermore, the MCC curve shows that there is a limit in the γ value after which, classification performance worsen. For instance, densenet-121 was trained with $\gamma = \{325, 422, 425\}$. For the first values (325, 422), $MCC > 0$, but for $\gamma = 425$, $MCC < 0$. Other models perform in the same fashion. This means that FT relies on the γ value to accuratly classify the mammogram image, but if γ is *too deep*, the classifier does not work. Consequently, *Too deep* could be defined as the γ value for which MCC turns negative when a ConvNet is fine tuned. This is an interesting conclusion since now it could be pointed that MCC serves to tell how deep FT can be performed on a ConvNet. Moreover, MCC shows that some models are not good predictors if they have $MCC < 0$. For instance, the model Resnet-101-330-FT has a negative MCC despite the accuracy value it has.

Similarly as before, Table 4.7 presents the performance of the top 5 fine tuned models wrt. the aforementioned metrics. The results are ordered in descending order. The best performance was obtained by VGG16-8-FT (fine tuned from $\gamma = 8$) with $AUC = 0.93$. Considering $AUC > 0.90$ as a good predictor, a total of 4 out of 60 models perform above 90% based on the AUC metric. This is a lower number of models compared to TL. The complete table with all the fine tuning results for Inbreast dataset D_3 can be found in Appendix A.1.

Figure 4.12 shows the behavior of train and validation accuracy as a function of the training epochs for vgg16-8-FT, which was the best model found. Early stopping finished training at epoch 350 to prevent overfitting. As can be seen in Figure 4.12, the distance between validation and training is shorter compared to Figure 4.4. This suggests that FT is better than TL in reducing overfitting. However, the number of parameters trained is bigger than in TL. The ROC Curve and the Confusion matrix are shown in Figures 4.13 and 4.14, respectively. Notice that there were 31 cases of false negative, where there is cancer but the classifier confuses these cases with benign.

Table 4.7: Inbreast Fine Tuning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
vgg16-8-FT	0.93	0.93	0.928	0.862	0.93
vgg19-11-FT	0.918	0.918	0.916	0.838	0.918
vgg16-6-FT	0.917	0.917	0.913	0.836	0.917
vgg16-10-FT	0.912	0.912	0.911	0.823	0.912
vgg19-17-FT	0.898	0.898	0.899	0.797	0.898

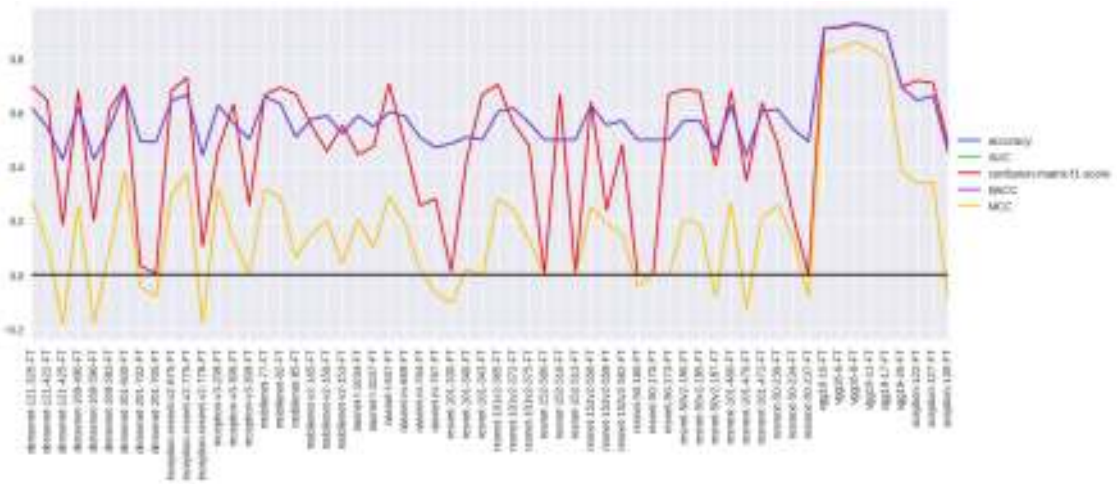

Figure 4.11: Fine Tuning for Inbreast on Dataset D_3

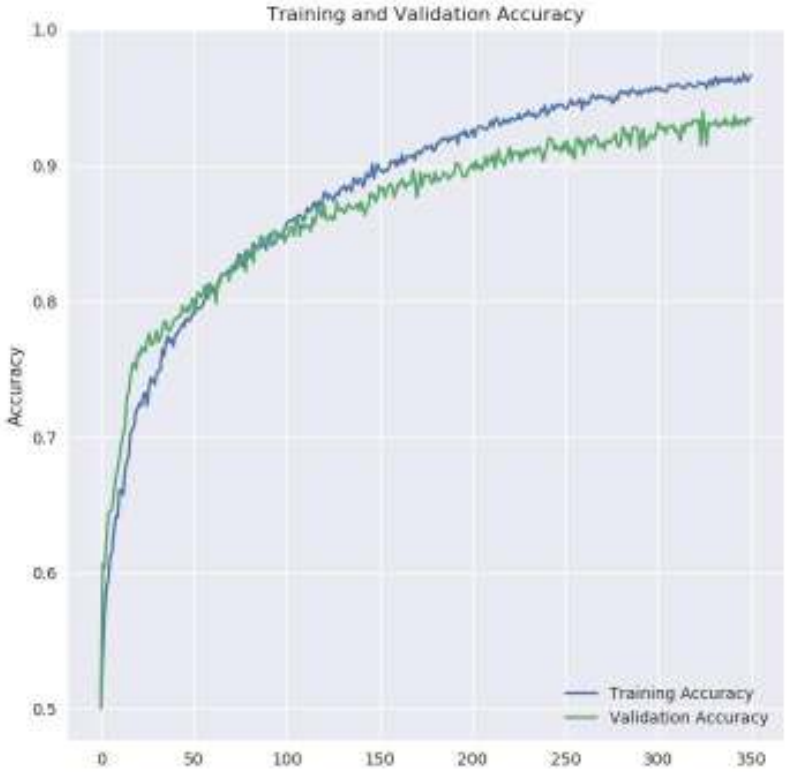

Figure 4.12: Inbreast Vgg16-8 FT Training Accuracy Curve

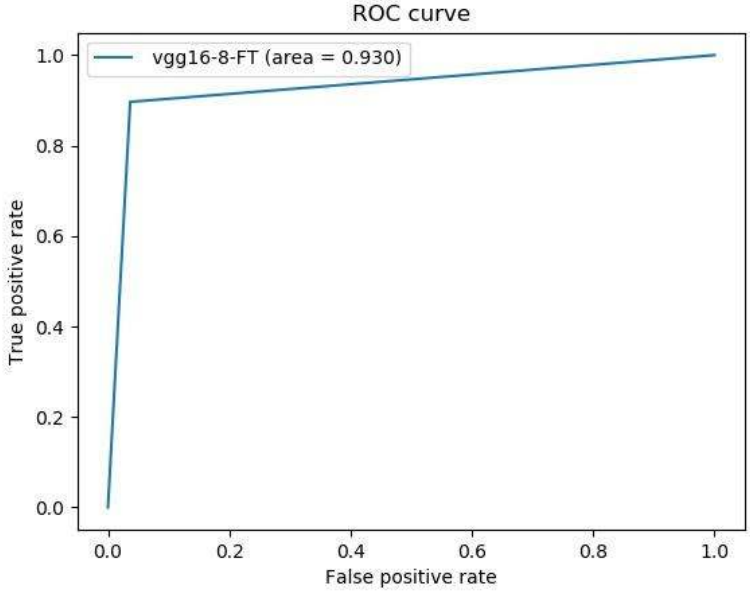

Figure 4.13: Inbreast Vgg16-8 FT ROC Curve

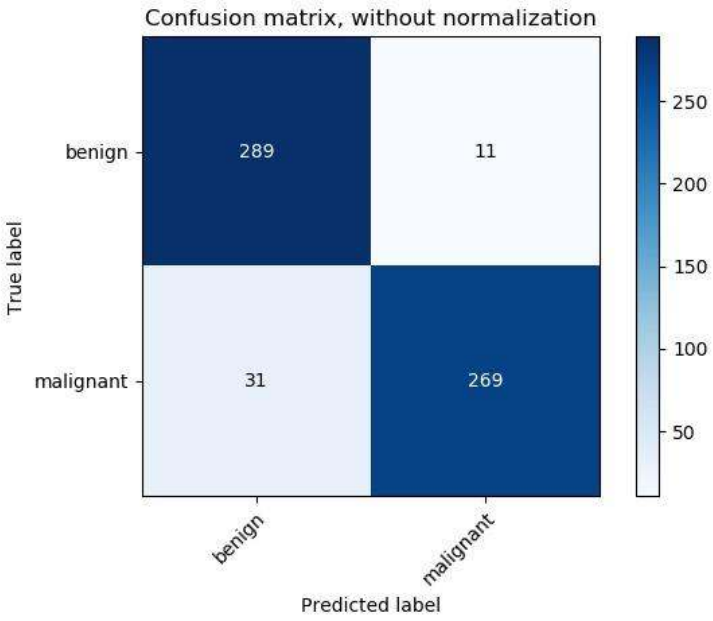

Figure 4.14: Inbreast Vgg16-8 FT Confusion Matrix

4.4.2 Results of Fine Tuning Training on MIAS

Similarly as before, Figure 4.15 plots the 20 fine tuned ConvNets wrt ACC, $BACC$, MCC, F_1Score and AUC for D_3 MIAS. The classification performance variers among the tested models. As can be seen from the picture Vgg16 and Vgg19 models achieved better results than the others. In fact, Vgg models stabilized all performance metrics. Again some models present $MCC < 0$.

Table 4.8 presents the performance of the top 5 fine tuned models in MIAS in descending order. The best performance was obtained by VGG16 fine tuned from $\gamma = 6$ with $AUC = 0.935$. Considering $AUC > 0.90$ as a good predictor, a total of 5 out of 61 models[1] performed above 90% on AUC .

Figure 4.16 shows the behavior of train and validation accuracy as a function of the training epochs for vgg16-6-FT. Early stopping finished training around the 500^{th} epoch to prevent overfitting. Notice the small gap between the train and validation accuracy and compare it with Figure 4.8. Fine Tuning is a good technique to prevent overfitting. Like before, it seems to control overfitting better than TL. The ROC Curve and the Confusion matrix are shown in Figures 4.17 and 4.18, respectively.

Table 4.8: MIAS Fine Tuning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
vgg16-6-FT	0.935	0.935	0.935	0.87	0.935
vgg16-10-FT	0.928	0.928	0.929	0.857	0.928
vgg16-11-FT	0.925	0.925	0.925	0.85	0.925
vgg16-8-FT	0.922	0.922	0.921	0.843	0.922
vgg19-11-FT	0.902	0.902	0.905	0.805	0.902

[1] An additional Vgg16 model was trained

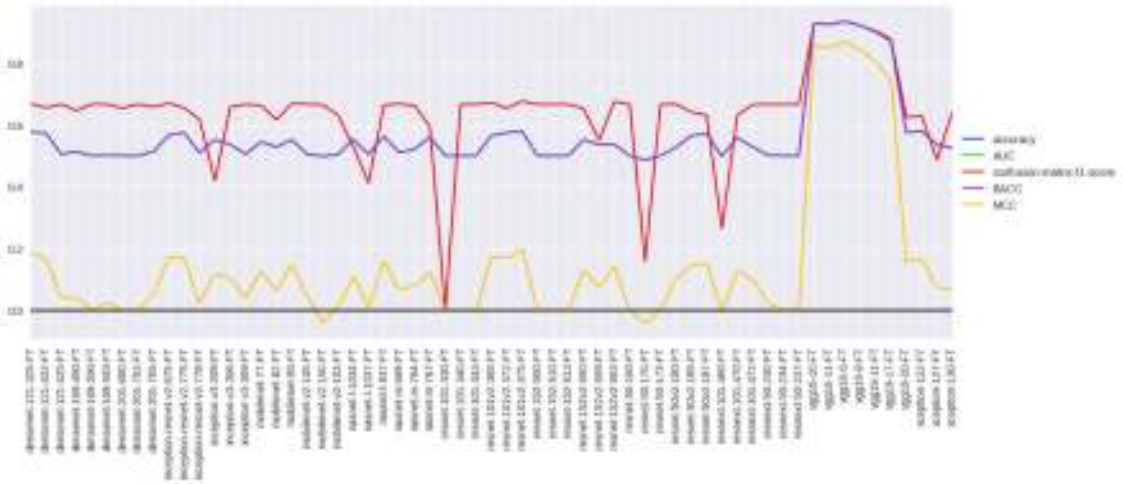

Figure 4.15: Fine tuning for MIAS on Dataset D_3

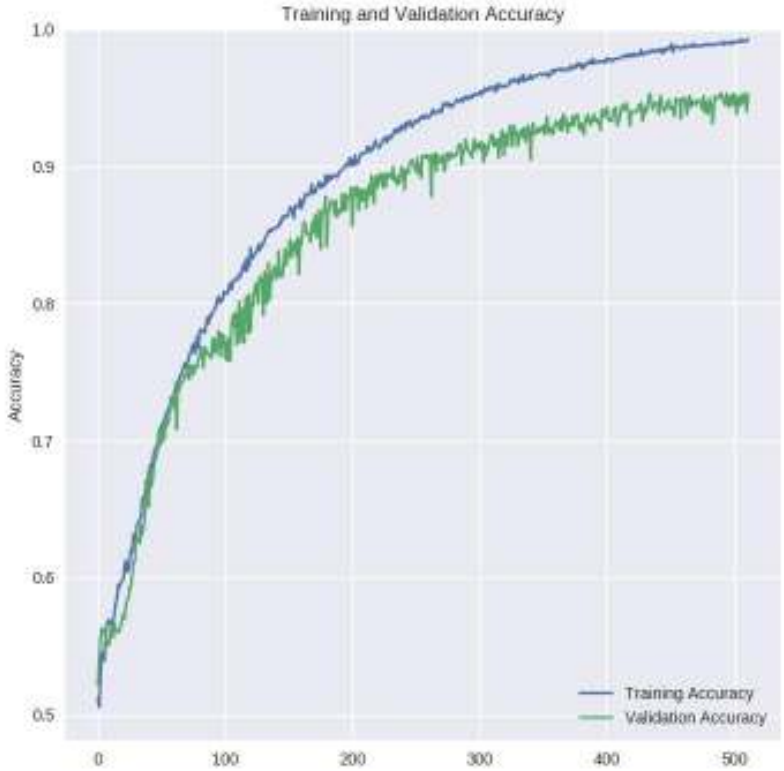

Figure 4.16: Mias Vgg16-6 FT Training Accuracy Curve

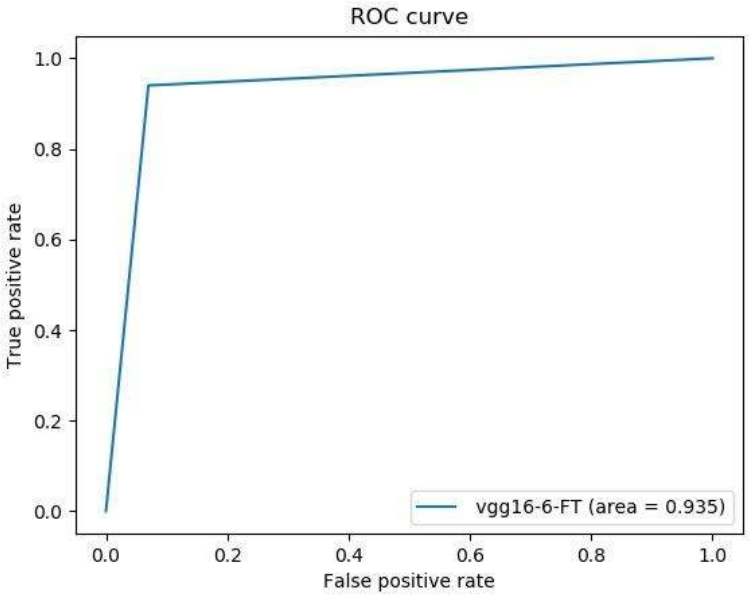

Figure 4.17: Mias Vgg16-6 FT ROC Curve

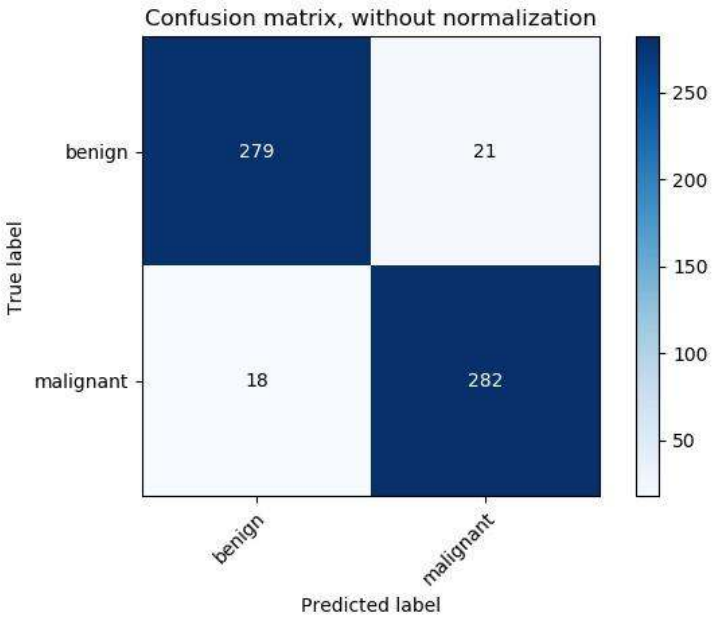

Figure 4.18: Mias Vgg16-6 FT Confusion Matrix

4.5 Whole-Retrain

Experiments carried out in this section retrained all the models with random initialized weights. Remind the definitions of TL, FT and WR provided in Chapter 2. According to the proposed methodology, these set of experiments correspond to the control group type. This means that they permit to evaluate the hypothesis regarding if TL and FT are effective techniques to train ConvNets for mammogram classification. As before, Tensorflow v1.13.1 (Martín Abadi et al., 2015) framework allowed to initialize each ConvNet with a set of random values instead of the ImageNet values. The output architecture used here ($\mathcal{A}$) is the same as in Transfer Learning (see Equation (4.1)); therefore, the number of neurons used in FC layer is 4096 for both Inbreast and MIAS.

4.5.1 Results of Whole Retrain on Inbreast

Figure 4.19 plots each Whole Retrained ConvNet model against the classification performance metrics (ACC, $BACC$, MCC, F_1Score and AUC). No negative value of MCC is found except for Nasnet Mobile which has $MCC = 0$ and $BACC = 0.5$. A value of $BACC \approx 0.5$ indicates low performance of the classifier.

Table 4.9 presents the values of each ConvNet for the aforementioned metrics in descending order. This shows that ResNext-101-WR achieved the best result with $AUC = 0.952$. Notice that WR stands for Whole Retrain. Considering $AUC > 0.90$ as a good predictor, a total of 8 out of 20 models perform above 90% on AUC. It is interesting to notice that ResNext-101 achieved a higher AUC value than Vgg16-8-FT ($AUC = 0.93$) and Mobilenet-TL on Inbreast ($AUC = 0.947$). However, notice that three different models have achieved best performances in three different techniques on the same dataset.

Figure 4.20 shows the behavior of train and validation accuracy as a function of the training epochs for ResNext-101-WR. Early stopping finished training ResNext-101 at epoch 300. However, compared to FT (Figure 4.12) and TL (Figure 4.4), the gap between train and validation accuracy is bigger. Therefore, despite the high value obtained for accuracy and AUC, ResNex-101 is more prone to overfitting compared to Vgg-16-8-FT and Mobilenet-TL. The ROC Curve and the Confusion matrix are shown in Figures 4.21 and 4.22, respectively.

Table 4.9: Inbreast Whole Retrain Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
resnext-101-WR	0.952	0.952	0.951	0.903	0.952
resnet-152-WR	0.948	0.948	0.947	0.898	0.948
vgg19-WR	0.94	0.94	0.94	0.88	0.94
nasnet-l-WR	0.938	0.938	0.94	0.878	0.938
vgg16-WR	0.928	0.928	0.925	0.861	0.928
resnet-101-WR	0.918	0.918	0.921	0.838	0.918
inception-v3-WR	0.912	0.912	0.91	0.824	0.912
resnet-152v2-WR	0.903	0.903	0.902	0.807	0.903
densenet-201-WR	0.882	0.882	0.887	0.766	0.882
densenet-169-WR	0.882	0.882	0.883	0.764	0.882
inception-resnet-v2-WR	0.855	0.855	0.844	0.717	0.855
densenet-121-WR	0.843	0.843	0.851	0.69	0.843
xception-WR	0.81	0.81	0.824	0.627	0.81
resnext-50-WR	0.783	0.783	0.759	0.578	0.783
resnet-101v2-WR	0.777	0.777	0.769	0.555	0.777
resnet-50v2-WR	0.762	0.762	0.735	0.534	0.762
resnet-50-WR	0.692	0.692	0.731	0.401	0.692
mobilenet-WR	0.638	0.638	0.591	0.284	0.638
mobilenet-v2-WR	0.625	0.625	0.532	0.272	0.625
nasnet-m-WR	0.5	0.5	0.0	0.0	0.5

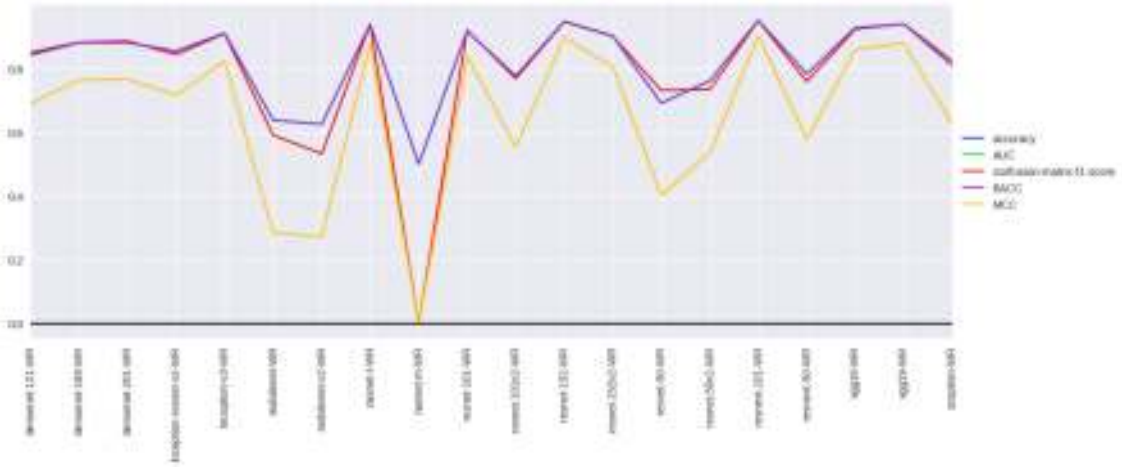

Figure 4.19: Whole Retrain for Inbreast on Dataset D_3

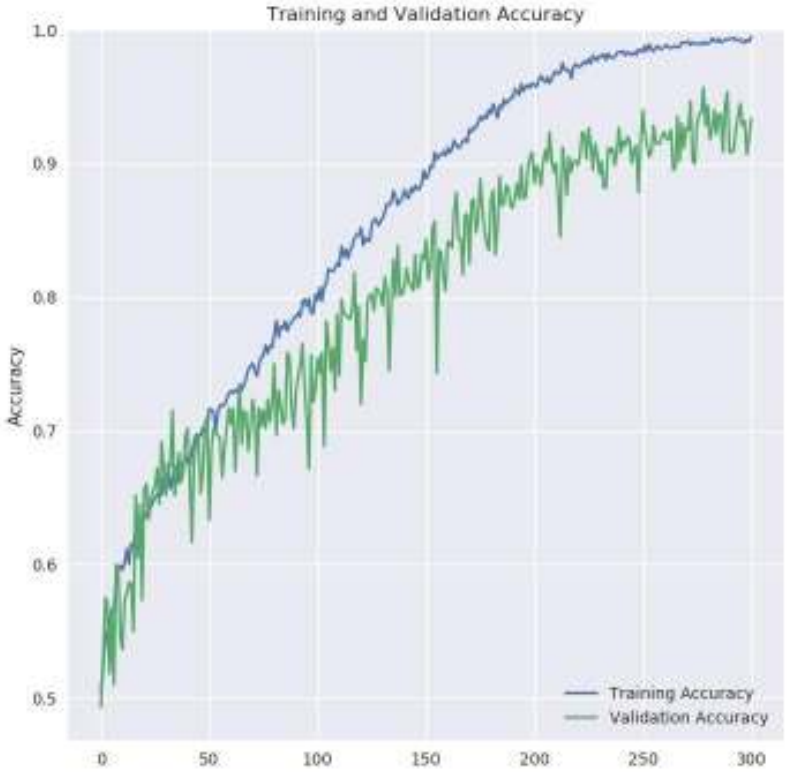

Figure 4.20: Inbreast Resnext 101 WR Training Accuracy Curve

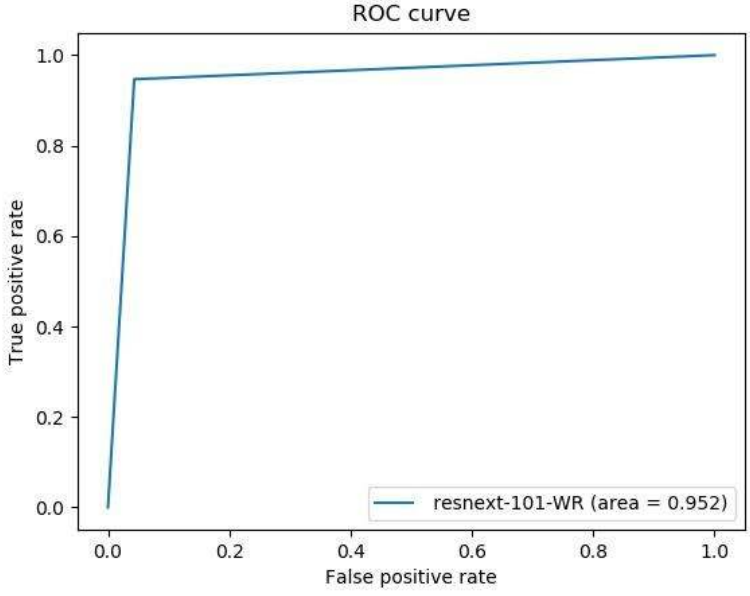

Figure 4.21: Inbreast Resnext-101-WR ROC Curve

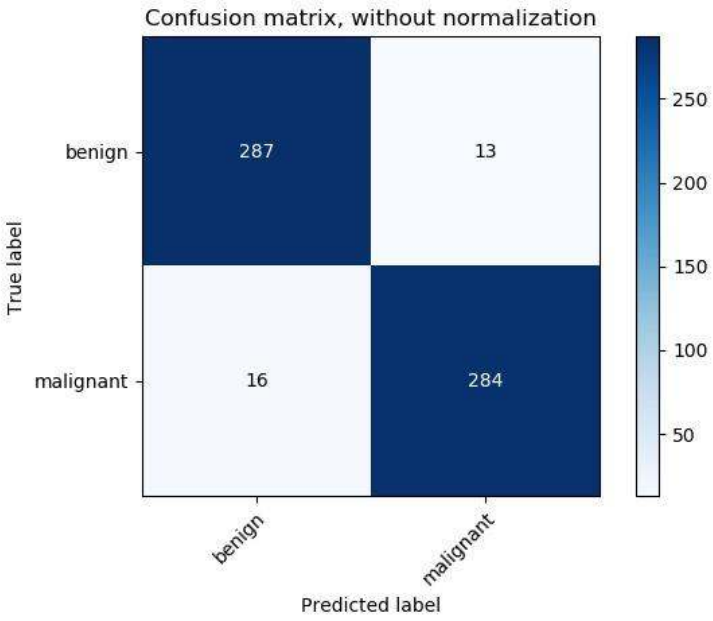

Figure 4.22: Inbreast Resnext-101-WR Confusion Matrix

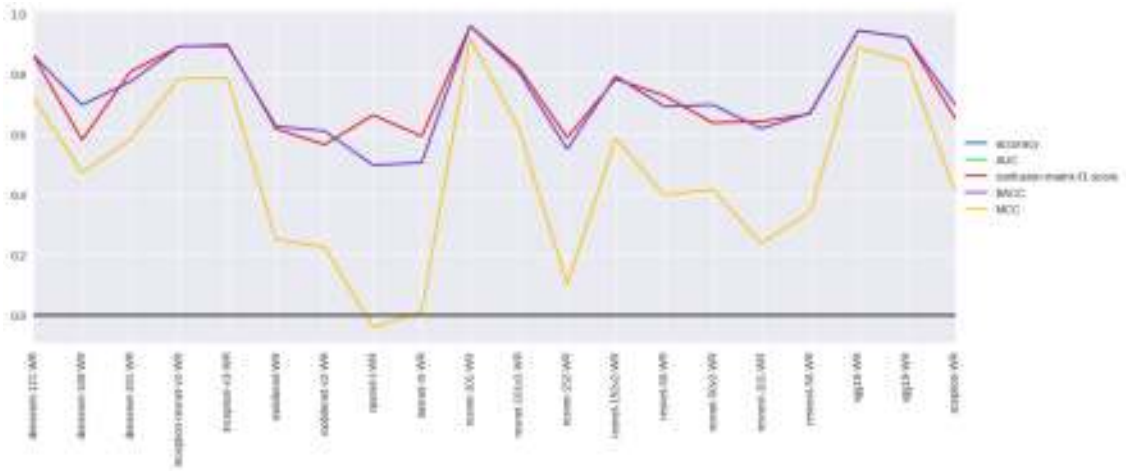

Figure 4.23: Whole Retrain for MIAS on Dataset D_3

4.5.2 Results of Whole Retrain on MIAS

Figure 4.23 plots the 20 Whole Retrained Convnet against the classification metrics (ACC, $BACC$, MCC, F_1Score and AUC). Negative MCC is observed for NasNet-L. In ConvNets Inception, ResNet-101, Vgg16 and Vgg19, the performance is above 80%.

Table 4.10 presents the top 5 ConvNets in WR for MIAS D_3 in descending order. Thus, the best performance was obtained by ResNet101-WR with $AUC = 0.958$, followed by vgg16-WR and vgg19-WR. Notice that the letters WR at the end mean that the technique used in training was Whole Retrain. Considering $AUC > 0.90$ as a good predictor, a total of 3 out of 20 models performed above 90% on AUC. Comparing to TL ($AUC = 0.952$) and FT ($AUC = 0.935$), ResNet-101-WR on MIAS achieved a similar result as TL. But ResNet-101-v2-TL does have a smaller gap between train and validation curves as presented in Figure 4.8.

Figure 4.24 shows the behavior of train and validation accuracy as a function of the training epochs for ResNet-101-WR. Early stopping finished training at epoch 350 to prevent overfitting. Comparing Figure 4.24 with Figure 4.16, the gap between train and validation seems bigger. In FT there is almost a constant distance between the train and validation curves. The ROC Curve and the Confusion matrix are shown in Figures 4.25 and 4.26, respectively.

Table 4.10: MIAS Whole Retrain Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
resnet-101-WR	0.958	0.958	0.958	0.917	0.958
vgg16-WR	0.943	0.943	0.943	0.887	0.943
vgg19-WR	0.92	0.92	0.919	0.84	0.92
inception-v3-WR	0.892	0.892	0.896	0.786	0.892
inception-resnet-v2-WR	0.892	0.892	0.89	0.784	0.892
densenet-121-WR	0.86	0.86	0.859	0.72	0.86
resnet-101v2-WR	0.807	0.807	0.82	0.621	0.807
resnet-152v2-WR	0.792	0.792	0.782	0.586	0.792
densenet-201-WR	0.773	0.773	0.807	0.583	0.773
xception-WR	0.697	0.697	0.65	0.408	0.697
densenet-169-WR	0.697	0.697	0.581	0.472	0.697
resnet-50v2-WR	0.697	0.697	0.637	0.416	0.697
resnet-50-WR	0.692	0.692	0.728	0.397	0.692
resnext-50-WR	0.67	0.67	0.667	0.34	0.67
mobilenet-WR	0.625	0.625	0.617	0.25	0.625
resnext-101-WR	0.618	0.618	0.642	0.239	0.618
mobilenet-v2-WR	0.61	0.61	0.565	0.225	0.61
resnet-152-WR	0.55	0.55	0.588	0.102	0.55
nasnet-m-WR	0.505	0.505	0.593	0.011	0.505
nasnet-l-WR	0.497	0.497	0.663	-0.041	0.497

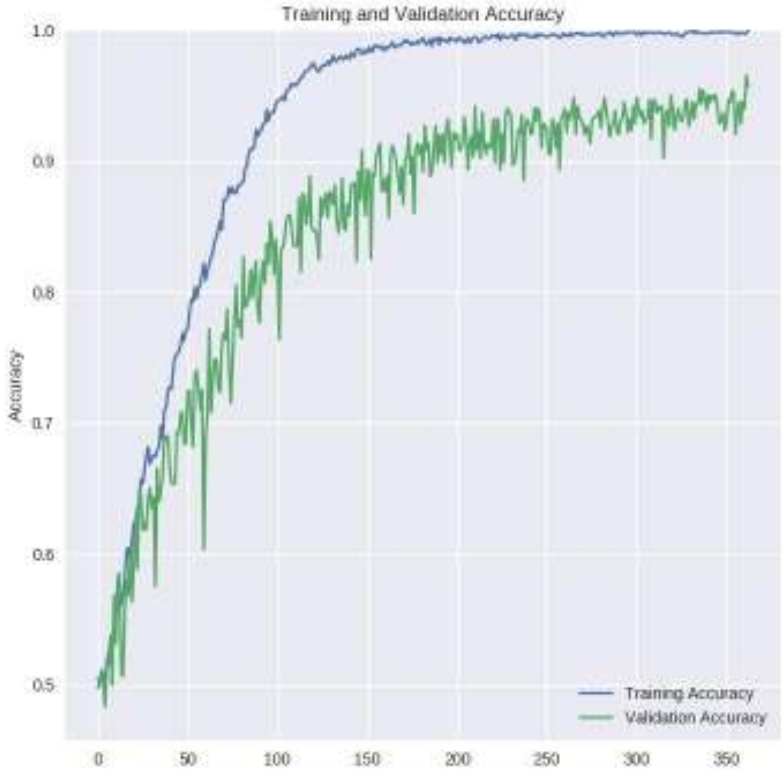

Figure 4.24: Mias Resnet 101 WR Training Accuracy Curve

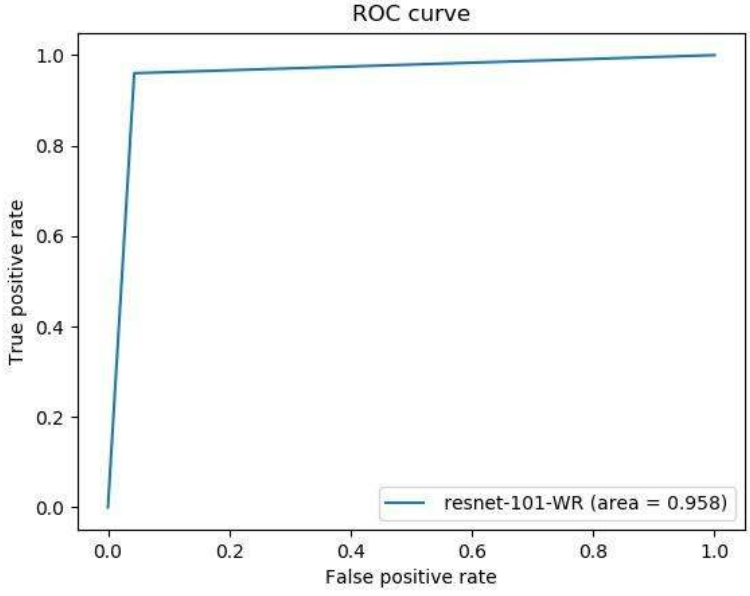

Figure 4.25: Mias Resnet 101 WR ROC Curve

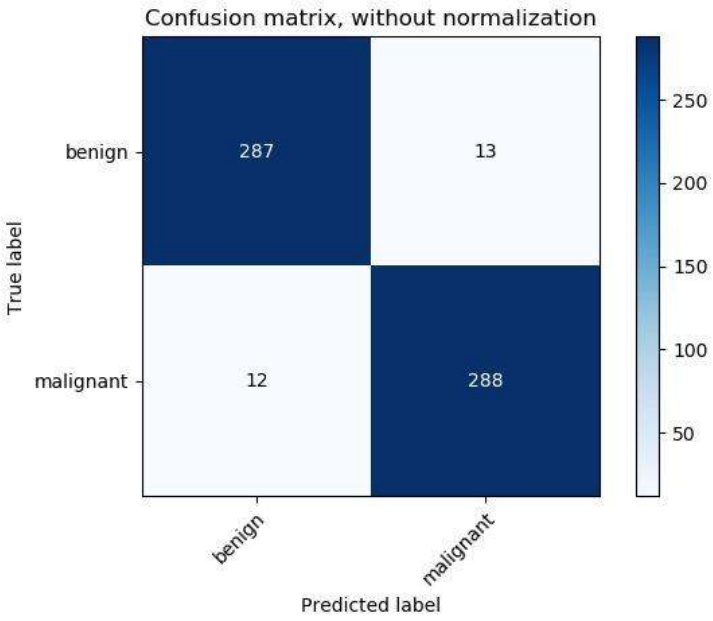

Figure 4.26: Mias Resnet 101 WR Confusion Matrix

4.6 Individual Classifier Comparative Analysis on TL, FT and WR

The previous section detailed the results for TL, FT and WR in Inbreast and MIAS D_3 datasets. According to the value of AUC, in both datasets, the highest value was achieved in the WR technique. This seems at first to contradict what is expected. Most literature assures that TL or FT must be better than whole retraining the ConvNet, but the results indicate the contrary. However, if a probability of success of a technique is defined as the ratio of the number of predictors with $AUC > 0.90$ wrt. the total number of models used, and this value is multiplied with the AUC, an average AUC (AUC_{avg}) is obtained. Table 4.11 presents these calculations for each of the best models in TL, FT and WR on Inbreast and MIAS. As can be seen in Table 4.11, the *probability of success* parameter is of 75% in Transfer Learning. Hence, TL is justified as the technique with more probabilities of success in training ConvNets for mammogram classification. Moreover, TL takes less time to be trained than FT or WR because the number of training parameters is smaller. Even though, the numeric results for WR were the best wrt TL and FT, a closer look to the training curves shows that WR tends to overfit. Because of this second reason, it would be preferred to use TL or FT as a training technique. Also, the difference between TL, FT and WR is small. Probably, experiments with FT could be repeated with new values that increase classification performance. Because of that reason, it cannot be suggested that FT performs worse than the other techniques. In fact, the experiments carried out have been able to present AUC results above 0.90 for each technique (TL, FT, and WR) with different ConvNets. It is interesting to notice that Vgg-16 is the best ConvNet to use in FT for mammogram classification in both datasets (vgg-16-8-FT Inbreast, vgg-16-6-FT, MIAS). Obviously, if experimentation would have been limited to Vgg-16 network only, FT would have achieved the best performance instead of TL and WR. This suggests that a combination of the training technique (TL, FT, WR), the ConvNet architecture, and the size of the dataset assures to find a suitable classification model.

4.7 Ensemble Learning

This section presents, the results achieved by using ensemble learning on dataset D_3 for Inbreast and MIAS. As mentioned earlier in Section 2.7 of Chapter 2, an

Table 4.11: Comparison of TL, FT and WR with probability of Success

Dataset	Technique	Model	Probability	AUC	AUC_{avg}
Inbreast	Transfer Learning	Mobilenet-TL	0.75	0.947	0.71
Inbreast	Fine Tuning	vgg-16-8-FT	0.07	0.93	0.06
Inbreast	Whole Retrain	resnext-101-WR	0.40	0.952	0.38
MIAS	Transfer Learning	resnet-101-v2-TL	0.75	0.952	0.71
MIAS	Fine Tuning	Vgg-16-6-FT	0.08	0.935	0.08
MIAS	Whole Retrain	resnet-101-WR	0.15	0.958	0.14

ensemble classifier combines different *base classifiers* to increase generalization performance. Two ensembles were formed for each mammogram database by combining the best models found in TL (Tables 4.5, 4.6), FT (Tables 4.7, 4.8) and WR (Tables 4.9, 4.10). The global results can be consulted in Table 4.11. The *base models* used to form the ensemble classifier for each database are indicated in Table 4.12. Thus, Mobilenet-TL, Vgg16-8-FT and Vgg19-WR conform the base models used for the Inbreast Ensemble. Likewise, Resnet-101-v2-TL, Vgg16-6-FT and Resnet-101-WR for MIAS Ensemble.

Three different types of ensembles were tested. The first one used hard voting algorithm. The second one used soft voting with all the averaging weights set to 1. The third one was called *Automatic Soft Voting Ensemble* (ASVE) or in short *soft-voting-auto* and was proposed in this thesis. The ASVE algorithm was called as such because it uses a Perceptron to find the best possible weights values for soft voting. As mentioned before, soft voting is defined as the weighted sum of probability scores of each base classifier. If carefully looked, the equation that defines soft voting (2.22) resembles that of a Perceptron. Because of this similarity between the soft voting and Perceptron equations, this work proposed to train a Perceptron to find the w_j values and improve ensemble performance. Figure 4.27 shows an schematic of the types of ensembles designed for testing.

One of the most difficult problems in mammography is to be able to train a model that performs on Film and Digital mammography. To address this issue, an ensemble combining the base models of Inbreast and MIAS was created. This new ensemble model was tested in a new test set created by joining the test set of

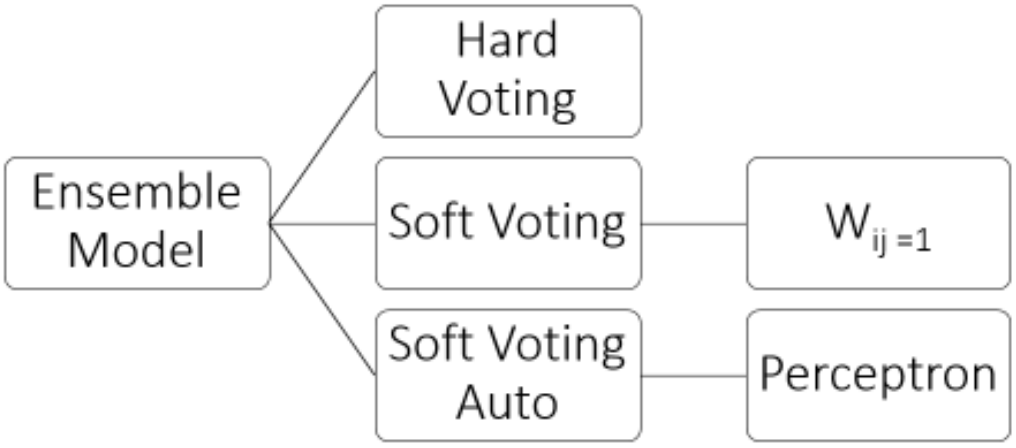

Figure 4.27: Ensemble Models Designed

each D_3 database. The performance of each ensemble is evaluated by using the same set of metrics used for TL, FT, an WR. These metrics were described in Section 3.4.5 of Chapter 3.

Table 4.12: Ensemble Models for Inbreast and MIAS

Technique	**Inbreast Ensemble**	**MIAS Ensemble**
Transfer Learning	mobilenet-TL	resnet-101-v2-TL
Fine Tuning	vgg16-8-FT	vgg16-6-FT
WR	vgg19-WR	resnet-101-WR

4.7.1 Results of Inbreast Ensemble Model

The Inbreast ensemble was formed by combining Mobilenet, Vgg16 fine tuned at $\gamma = 8$, and the whole retrain of Vgg19. Even though Resnext-101-WR achieved higher AUC value, Vgg19 was selected instead due to computer efficiency and because there is little difference in classification performance between them. Similar to the previous tests, this work compared the performance of each ensemble against the performance metrics (ACC, $BACC$, AUC, F_1Score, and MCC). The performance achieved by each ensemble on Inbreast dataset is presented in Table 4.13. Furthermore, the table includes the performance of the base models that form the ensemble. It can be seen from Table 4.13 that *Automatic Soft Voting Ensemble* (in short: soft-voting-auto) achieved the highest performance and improved the results achieved by each base classifier of the ensemble. Figure 4.28 also indicates that better values were obtained in general by using ensemble models.

The proposed Automatic Soft Voting Ensemble (ASVE) algorithm, which uses a Perceptron to find recommended weights, outperformed the single best model (Mobilenet). The ASVE achieved $AUC = 0.982$. One interesting thing about this result is that it outperforms the result reported in the state of the Art (see Table 2.4). The recommended weights by the Perceptron algorithm were: $w_{ij} = [3.313, 2.110, 3.417]$, for $ensemble_{inb} = [vgg19_{WR}, mobilenet_{TL}, vgg-16-8_{FT}]$. Moreover, another interesting fact to notice is that MCC is also increased thanks to the ensemble; achieving values that are closer to 1; which indicates a better classifier.

Figures 4.30 and 4.29 present the Confusion Matrix and Roc Curve for the Automatic Soft Voting Ensemble Model. As can be seen, classification performance has been increased due to the reduction in miss classification between the benign and malignant classes as shown in the confusion matrix. This means that False Positives and False Negatives have been reduced, and general classification improved.

Table 4.13: Inbreast Ensemble Performance

$Model$	$BACC$	AUC	F_1Score	MCC	$Accuracy$
soft-voting-auto	0.982	0.982	0.982	0.963	0.982
soft-voting-ones	0.978	0.978	0.978	0.957	0.978
hard-voting	0.97	0.97	0.97	0.94	0.97
mobilenet-TL	0.947	0.947	0.947	0.893	0.947
vgg19-WR	0.94	0.94	0.94	0.88	0.94
vgg16-8-FT	0.93	0.93	0.928	0.862	0.93

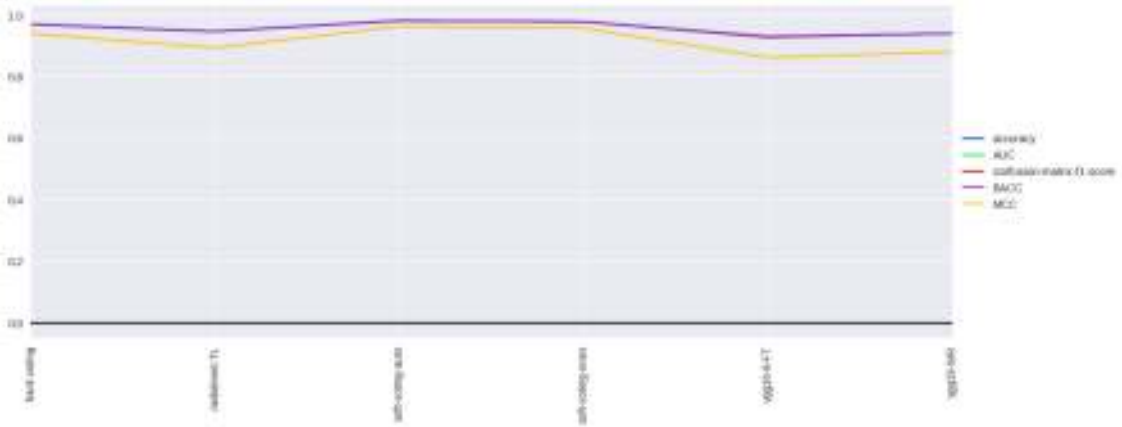

Figure 4.28: Inbreast Ensemble Performance

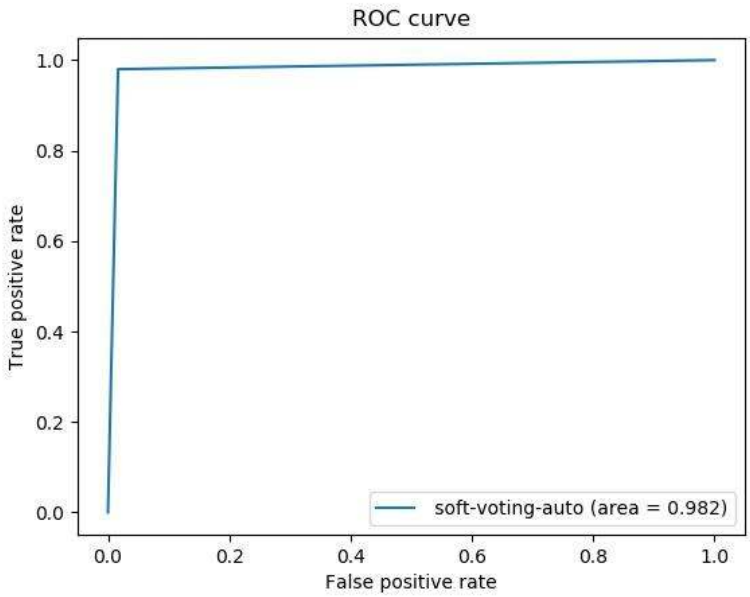

Figure 4.29: Inbreast Automatic Soft Voting Ensemble ROC Curve

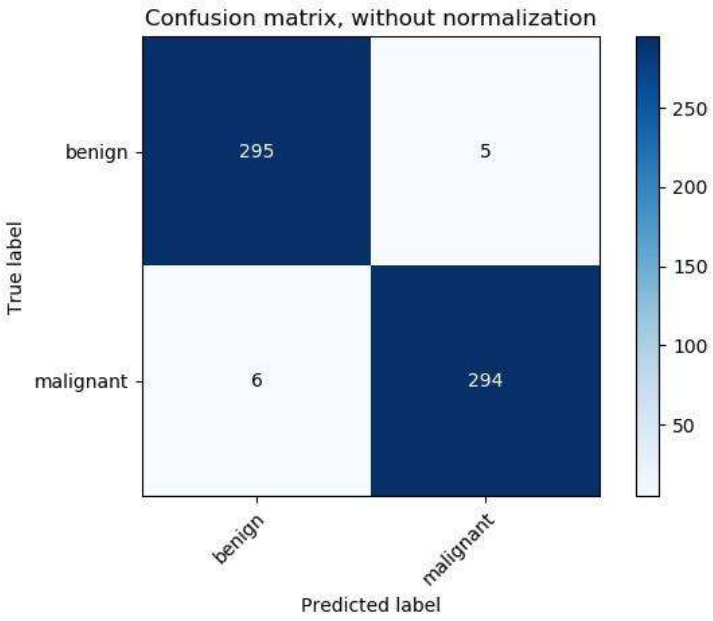

Figure 4.30: Inbreast Automatic Soft Voting Ensemble Confusion Matrix

4.7.2 Results of MIAS Ensemble

The Mias ensemble was formed by Resnet-101-v2-TL, Vgg16-6-FT, and Resnet-101-WR that correspond to the best models for each technique. The performance achieved by each ensemble model for MIAS dataset is compared through the binary classification metrics (ACC, $BACC$, AUC, F_1Score, and MCC). Table 4.14 presents the performance achieved by each ensemble model including the base classifiers. It can be seen from the aforementioned table that this thesis' proposed model called *soft-voting-auto* (ASVE) achieved the best performance with $AUC = 0.978$. In fact, it outperformed the single best model. The Perceptron recommended weights were $w_{ij} = [2.5121.5971.734]$, for $ensemble_{mias} = [resnet-101v2_{TL}, resnet-101_{WR}, vgg-16-g_{FT}]$. Figure 4.31 shows that the performance of ensemble models is better than that of each single model classifier. In fact, MCC has been improved. Compared to the Confusion Matrix of Resnet-101-WR, the Soft-Voting-Auto Confusion Matrix presented in 4.33 indicates a reduction in the miss-classification of benign and malignant classes in MIAS dataset. Figure 4.17 presents the ROC curve for the Automatic Soft Voting Ensemble model.

Table 4.14: MIAS Ensemble Performance

$Model$	$BACC$	AUC	F_1Score	MCC	$Accuracy$
soft-voting-auto	0.978	0.978	0.978	0.957	0.978
soft-voting-ones	0.975	0.975	0.975	0.95	0.975
hard-voting	0.973	0.973	0.973	0.947	0.973
resnet-101-WR	0.958	0.958	0.958	0.917	0.958
resnet-101v2-TL	0.952	0.952	0.952	0.903	0.952
vgg16-6-FT	0.935	0.935	0.935	0.87	0.935

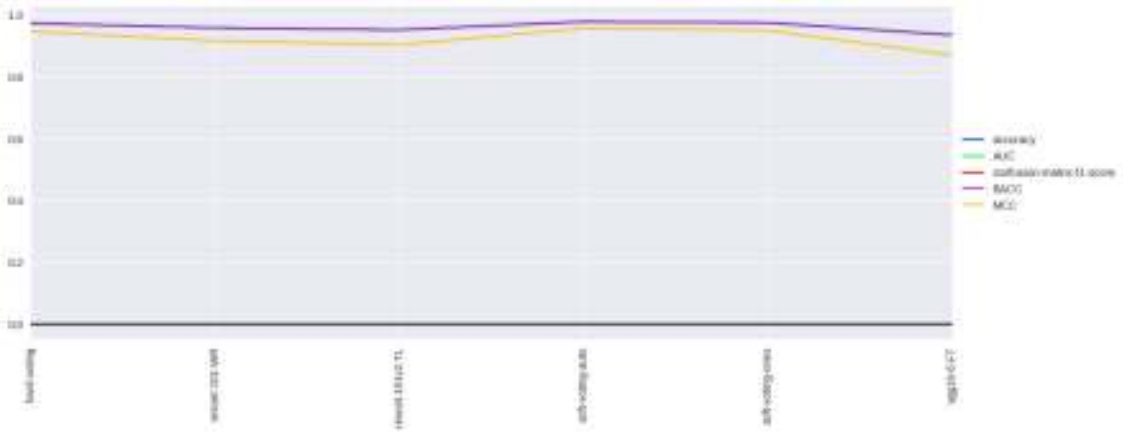

Figure 4.31: MIAS Ensemble Performance

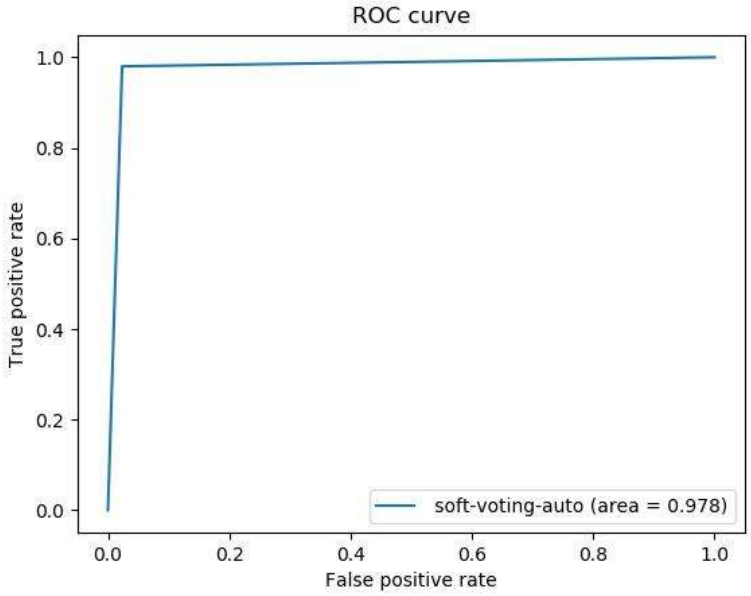

Figure 4.32: MIAS Automatic Soft Voting Ensemble ROC Curve

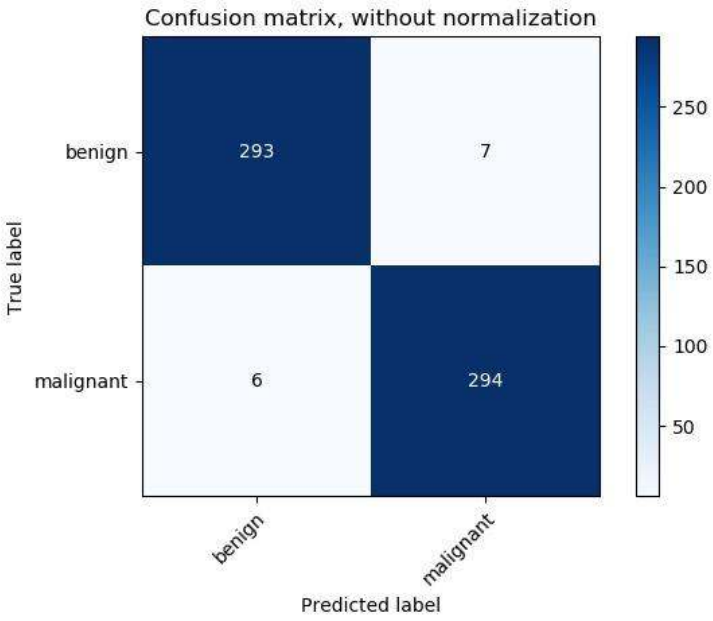

Figure 4.33: MIAS Automatic Soft Voting Ensemble Confusion Matrix

4.7.3 Ensemble Learning model for Film and Digital Mammography

One of the most important aspects of a classifier is its ability to generalize to new samples. If the classifier is able to make accurate predictions on not previously seen data, it could be said that the classifier has learned and not memorized (Goodfellow et al., 2016). Modern ConvNets have reached a great level of performance through increasing the number of layers. NasNet for instance is a 1000 layers ConvNet. Such big architectures are capable of memorizing significant information of the dataset. The key aspect to get a ConvNet to perform accurately is to train it with a great variation of samples that can represent the probability distribution of the input space. However, in the case of medicine, and particularly of breast mammogram images, this is not the case. In fact, this research work used two datasets that are the quiet small.

It is clear that any model trained on the Inbreast dataset has specialized in digital mammogram images, whereas any model trained on the MIAS dataset has specialized in film mammography. Consequently, using a model trained on Inbreast will not perform accurately enough on the MIAS dataset. In order to solve this problem there are several options:

1. Use transfer learning from Inbreast to MIAS and review if the resulting model is capable of predicting samples from both datasets.
2. Create a merged dataset with training samples from both databases and use Transfer Learning to train a model.
3. Use Multi Task Transfer Learning, which is a technique that can be used to train several ConvNets in different datasets. This approach is studied in Samala et al. (2017).
4. Use Bagging
5. Create an Ensemble of models trained on both datasets individually.

This work used the approach of creating an ensemble of trained models on both datasets individually. This means that specialized models on Inbreast and MIAS are combined to form an Ensemble model that is predicted to work on both type of images. This Section details experimentation on such ensemble. As shown by the results provided in the previous section, ensemble classifiers outperform

individual classifiers. In order to test the performance of the proposed ensemble model, a new dataset was created by joining the test sets from MIAS and Inbreast D_3. Therefore, the new model is called *Mixed Ensemble* and was formed by using the base models from Inbreast and MIAS. In other words, the *Mixed Ensemble* is composed of Mobilenet-TL, Vgg16-8-FT, Vgg19-WR, Resnet-101-v2-TL, Vgg16-6-FT, and Resnet-101-WR; all the models presented in Table 4.12.

The proposed *Mixed Ensemble* model was tested in the same three ensemble modalities as before: hard voting, soft voting with all weights set to one, and this thesis' proposed ensemble *Automatic Soft Voting Ensemble* (ASVE or soft-voting-auto), which trained a Perceptron to calculate the weights values of the Soft Voting Algorithm. Table 4.15 presents the performance of the Mixed Ensemble Classifier. Inbreast and MIAS individual models achieve at most $AUC \approx 0.7$. This confirms that Inbreast models classification error is increased when predicting MIAS data and vice-versa. Notice that the ensemble formed by the 6 base models achieves a better performance with $AUC = 0.918$ with the ASVE model. This means that the ensemble performance has aided in the generalization capabilities in order to provide a solution that can predict pathology malignancy class in both digital and film images and outperformed single classifiers. Furthermore, it is worth noticing that the proposed Perceptron Soft Voting Algorithm (ASVE) has achieved the best results compared to the other ensemble techniques tried in this work. Although, there is little difference between soft-voting-auto and hard-voting based on AUC metric, F_1Score and MCC present a bigger difference. This suggests the use of ASVE model. In the same way, Figure 4.34 displays each ensemble and individual classifier performance metrics on the classification metrics. It shows that hard voting and ASVE achieved the highest results, outperforming individual classifiers.

As previously discussed, from the three types of ensemble tested, *Automatic Soft Voting Ensemble* was chosen. Figures 4.35 and 4.36 present the Roc Curve and Confusion matrix for ASVE, respectively. The results show that ASVE is capable of predicting benign and malignant pathology classes in mass mammograms for film and digital images. In fact the confusion matrix shows that False Negatives(22) are smaller than False Positives(76); which is promising.

Table 4.15: Mixed Ensemble Performance on Generalized Test Set

$Model$	$BACC$	AUC	F_1Score	MCC	$Accuracy$
soft-voting-auto	0.918	0.918	0.922	0.84	0.918
hard-voting	0.917	0.917	0.914	0.835	0.917
soft-voting-ones	0.908	0.908	0.913	0.821	0.908
mias-vgg16-6-FT	0.773	0.773	0.79	0.554	0.773
mias-resnet-101v2-TL	0.747	0.747	0.743	0.494	0.747
inb-vgg16-8-FT	0.743	0.743	0.755	0.489	0.743
mias-resnet-101-WR	0.743	0.743	0.738	0.487	0.743
inb-mobilenet-TL	0.727	0.727	0.762	0.475	0.727
inb-vgg19-WR	0.714	0.714	0.737	0.435	0.714

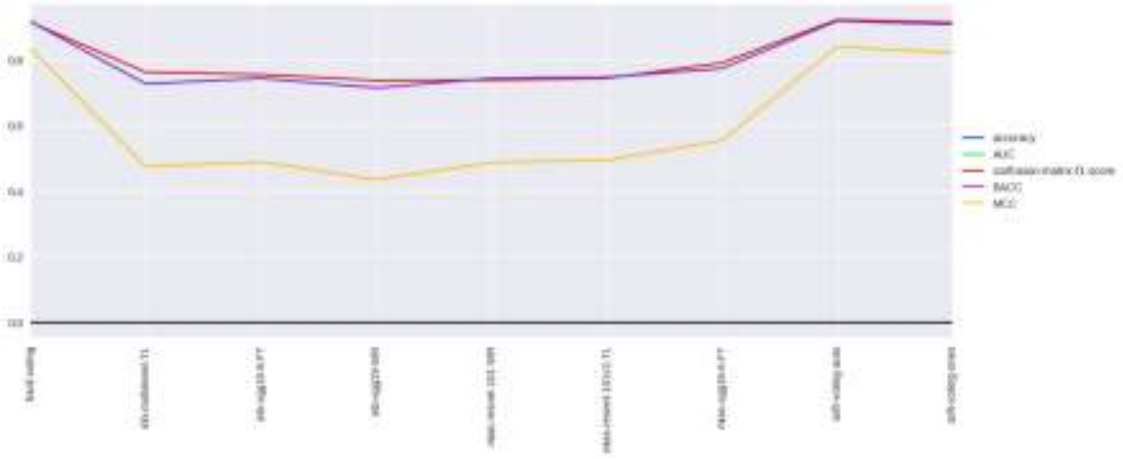

Figure 4.34: Mixed Ensemble Performance on Mixed Test Set

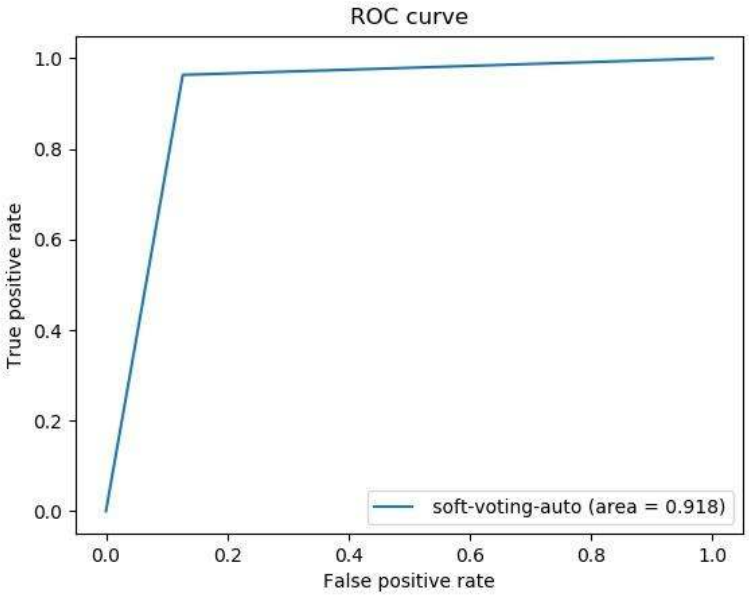

Figure 4.35: Mixed Dataset Soft Voting Ensemble ROC Curve

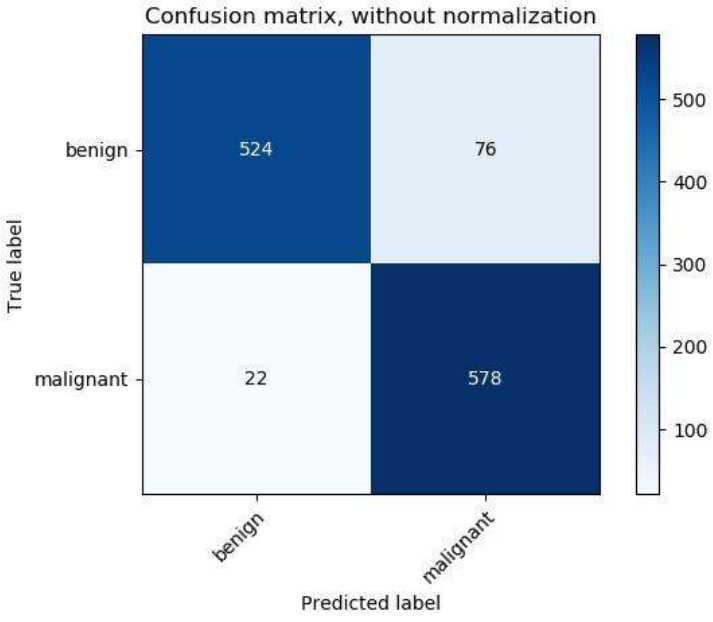

Figure 4.36: Mixed Dataset Soft Voting Ensemble Confusion Matrix

4.8 Experimental Results Discussion

In this Chapter, several experiments were carried out to accomplish the second specific object related to empirical research. The third goal, related to the proposal of an ensemble model that can improve the state of the art classification performance has been achieved for Inbreast Dataset.

Nevertheless, the results showed something unexpected: the individual performance in Whole Retrain achieved numerically better values of AUC compared to Transfer Learning and Fine Tuning. The reason for this is twofold: in both cases (Inbreast and MIAS), the best models in Whole Retrain mode are of the type of ResNet and ResNext. Compared to VGG-16, which is the best model for Fine Tuning, it seems that the increase in the depth of the model along with the residual block aided to extract meaningful features and probably prevented overfitting by using the short cuts. The second reason relies in the overfitting control techniques used for all the models: data augmentation, dropout and early stopping. Also, to the best of our knowledge, ResNext networks have yet not been tested on mammogram images; making these experiments an interesting contribution. Furthermore, if Tables 4.5 and 4.6 are compared, it can be observed that the triad Mobilenet, Resnet, and ResNext perform mammogram classification with an $AUC > 0.90$. Regarding FT, the experiments show that Vgg models outperformed the others. Consequently, it can be suggested that for mammogram classification, Vgg16, Vgg19, ResNet, ResNext and Mobilenet are the most suitable ConvNets independently of the training technique (TL, FT, or WR).

Comparing the probability of success, defined as the ratio of the number of classifiers with $AUC > 90$ with respect to the total number of models tested in a technique (TL, FT or WR), TL, as it was defined in Chapter 2, is the best technique to use for the mammogram classification problem in this work. It is also worth to notice that changing the value of γ affects the performance of the Fine Tuning model. The FT experiments tried to vary its value as systematic as possible in order to improve the performance of the single classifier. However, in most cases, this was not achieved, except for VGG and Mobilenet networks. This may be related to the number of samples to train; which is a limitation in the present work. As a matter of fact, the original datasets are scarce in data and creating more than 3000 image per class does not increase the information that is actually being used. Therefore, what is difficult in FT is finding a precise γ value. The

experiment suggests that MCC could be used to explore how deep γ should be, but this fact should be addressed specifically in another work. Also, it must be observed that WR and TL present a tendency to overfit whereas FT seems to control it better because of the little gap between train and validation curves.

Ensemble models formed with the best tuned ConvNets show an improved performance compared to the individual classifiers. Moreover, this thesis' proposed *Automatic Soft Voting Ensembe* (ASVE) model, which uses a Perceptron to find the values of the soft voting weighted sum, outperformed the other ensemble techniques used (hard voting and soft voting with $w_j = 1$). What is more remarkable is that ASVE outperformed other classifiers in the Literature Review as presented in Table 2.4 ($AUC = 0.97$ in Chougrad et al. (2018)) by achieving $AUC = 0.982$ on Inbreast. Regarding MIAS, it could be assumed that $AUC \approx ACC$ when the dataset is balanced and augmented; which is something that the results in this work suggest. Therefore, in the case of MIAS, the average of classifying performance is $AUC \approx 0.95$. In that case, ASVE ($AUC = 0.978$) outperforms literature results as well. This confirms that the third specific objective of this proposed work was achieved.

Tables 4.16 and 4.17 present a comparison of the results reported in literature with respect to this thesis' ASVE model. It can be seen that this work has outperformed state of the art results on Inbreast. In the case of MIAS, ASVE outperformed the average in literature.

Table 4.16: Comparative Results of this Work vs Literature on Inbreast

Article Title	Authors	AUC
"Deep Convolutional Neural Networks for breast cancer screening"	Chougrad, Zouaki, and Alheyane (2018)	0.97
"Automated Analysis of Unregistered Multi-View Mammograms with Deep Learning"	Carneiro, Nascimento, and Bradley (2017a)	0.94
"Deep generative breast cancer screening and diagnosis"	Shams, Platania, Zhang, Kim, and Park (2018)	0.925
"Unregistered multiview mammogram analysis with pre-trained deep learning models"	Carneiro, Nascimento, and Bradley (2015)	0.91
"A deep learning approach for the analysis of masses in mammograms with minimal user intervention"	Dhungel, Carneiro, and Bradley (2017)	0.91
Deep Learning Models for Classifying Mammogram Exams Containing Unregistered Multi-View Images and Segmentation Maps of Lesions	Carneiro, Nascimento, and Bradley (2017b)	0.91
"Deep multi-instance networks with sparse label assignment for whole mammogram classification"	Zhu, Lou, Vang, and Xie (2017)	0.89
"Comparing the performance of various deep networks for binary classification of breast tumours"	Hamidinekoo, Suhail, Denton, and Zwiggelaar (2018)	0.87
	Literature Maximum Value:	**0.97**
	Literature Minimum Value:	0.87
	Literature Average Value:	0.92
	This Thesis' ASVE Model:	**0.982**

Table 4.17: Comparative Results of this Work vs Literature on MIAS

Article Title	Authors	ACC	AUC
"Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks"	Guan and Loew (2017)	0.91	
"Deep Convolutional Neural Networks for breast cancer screening"	Chougrad, Zouaki, and Alheyane (2018)		0.99
	Literature Maximum Value:	**0.91**	**0.99**
	Literature Average Value:	0.95	0.95
	This Thesis' ASVE Model:	**0.978**	**0.978**

Chapter 5

Conclusion

5.1 Introduction

Mass classification in mammograms is critically important because breast cancer is the second most deadly type of cancer due to its high mortality incidence. Advances in computer science and image technology have contributed in improving the interpretation of medical images (Stoitsis et al., 2006). Furthermore, Deep Convolutional Neural Networks (ConvNet) set a turning point in artificial intelligence and machine learning due to their effectiveness in learning automatically representative features from the provided data and, therefore, have become essential in image recognition and classification.

The purpose of this research was to explore Transfer Learning (TL), Fine Tuning (FT) and Ensemble Learning to classify mass mammogram pathologies. Transfer Learning aims to improve the classification performance of a ConvNet model in a new Target Task by re-using some of its pre-trained weights and replacing the final layer to adapt it to the new task. Because literature shows different approaches to the use of TL, this thesis provided with a formal definition of TL, Fine Tuning (FT) and Whole Retrain (WR) that is expected to adapt the original concept for Computer Vision and ConvNets.

Prior work has showed that TL in classification mammogram is better than retraining the whole ConvNet from randomly initialized values. However, this thesis proves that even WR can achieve a performance over 90% in AUC for this problem. The best results for MIAS and Inbreast in literature were achieved in Chougrad et al. (2018). Even though, this work's approach is similar to that of

Chougrad et al. (2018), this thesis did not add 5 dense layers in the $\mathcal{A}$ operator. Instead, Dropout and Global Average Pooling were used with a single Full Connecting Layer. Moreover, the number of experimented pre-trained models was extended to 20 to cover the whole Keras library in Tensorflow. Also, this research extended the number of metrics used to validate the classification results by considering the observations suggested in Canbek et al. (2017).

Ensemble learning is little used in literature for mammogram classification. However, it resembles of importance because it aims to improve the generalization capabilities of classification models. This thesis proposed an Automatic Soft Voting Ensemble model that used a Perceptron to optimize its parameters. This type of ensemble model is new to the best of our knowledge. The results achieved by this model have outperformed the literature classification results for Inbreast dataset. Moreover, the mixed ensemble is able to perform on digital and film mammogram images with an accuracy above 90%.

This chapter presents and discusses the main conclusions, findings, results, and scientific contribution of this research work. First, the research objectives are evaluated by comparing them with respect to the results obtained by using the proposed methodology. Then, the contributions of this work are discussed. The chapter concludes with a discussion of the limitations and future work.

5.2 Research Objectives: Summary of Findings and Conclusions

The goal of this research was to design a classification model for mass lesions in mammograms using Transfer and Ensemble Learning techniques that improved the average performance found in literature. This goal can be interpreted as a set of questions:

1. Is TL on a pre-trained ConvNet able to address mass pathology classification in breast cancer?
2. Among all the pre-trained ConvNets models on the ImageNet dataset, is there one that outperforms others in mass pathology classification in breast cancer?

3. How should TL or FT be implemented to achieve a good classification result?

4. Is an ensemble of such trained models able to outperform other literature results?

The following subsections address questions 1, 2, 3 and 4, by reviewing the main findings from the literature review and the experimental results. One positive result from this research is that the main objective was fulfilled by the proposed approach. The implications of the methodology used are also discussed next.

5.2.1 Transfer and Ensemble Learning for Mammogram Classification

Summary

Transfer learning and Fine Tuning are not new concepts in computer science. The formal definition of transfer learning was proposed in Pan and Yang (2010). In datamining, TL is more close to the study of several techniques that permit the adaptation of a source model to the target data. One of its uses is to deal with unlabeled data. However, Deep Learning changes a bit the base concept because it is able to deal with images (Shao et al., 2014); which are data of different kind. Therefore, learning to detect objects and transferring this knowledge to another model will aid to perform better in the target task, even though the images are of different domain.

Transfer learning in images is used differently by different researchers. One of the most spread forms is to use the pre-trained network as a feature extractor. In this case, the final layer of the ConvNet produces a feature vector which is used in another trainable classifier like a SVM (Perre et al., 2018). Furthermore, other researchers use the pre-trained weights as initial values for their own training (Al-antari et al., 2018). However, the use that this work has adopted and formalized mathematically consists in replacing the last layer of the ConvNet and adding additional layers that will be re-train while the early stages of the network remain with the original weights (Guan & Loew, 2017).

The Equations (2.17), (2.18), (2.19), and (2.20) describe TL and FT wrt ConvNets and Computer Vision. Moreover, (2.20) can be used as the main equation from

which TL and WR are special cases: TL is performed when $\gamma = L$, FT when $\gamma < L$ and WR when $\gamma = 0$. These equations rely on the operator $\mathcal{A}$ that describes the combination of layers used to adapt the pre-trained ConvNet to the new Target Task. This is a useful theoretical contribution.

Respect to the type of solution proposed in the state of the art there are two approaches: whole mammogram classification and ROI patch classification. The former consists in predicting cancer in the complete mammogram image, which is considered more challenging. The latter consists in identifying the region of interest and extract it to classify it. In this thesis, the binary masks provided in the Inbreast dataset and the x, y, $radio$ coordinates provided in MIAS were used to extract the mass area.

The two most used databases in literature were DDSM and Inbreast. This work used Inbreast and MIAS to have a set of images from digital and film mammography, respectively. For Inbreast, the top AUC value is of 0.97 and was achieved in Chougrad et al. (2018). Another finding from literature review is that most researchers have used AlexNet and in a second place ConvNets like VGG and Resnet, besides our previous work that studied Nasnet and Mobilenet (Falconí et al., 2019). Consequently, the extension to study other pre-trained ConvNets was compelling. .

Conclusions

The concept of Transfer Learning in image processing and Deep Learning has been adapted and expressed mathematically in this work. Equations (2.17), (2.18), (2.19), (2.20), and the $\mathcal{A}$ have been proposed to aid in notation by helping to express the modifications that take place in the final layer of the original model. This is an interesting contribution that is hoped to be useful for researchers to express their work. Furthermore, TL and WR can be defined as a special case of FT equation, as follows:

- **Transfer Learning:**

$$\mathbb{T}_{\mathbb{L}} \coloneqq \mathcal{A}\left(\phi^{\mathsf{S}}\left(\phi^{\mathsf{S}}(\mathcal{I}^{T})\Big|_{0}^{\gamma-1}\right)\Big|_{\gamma}^{L}\right) \quad \text{when} \quad \gamma = L \tag{5.1}$$

- **Fine Tuning:**

$$\mathbb{F}_{\mathbb{T}} \coloneqq \mathcal{A}\left(\phi^{\mathsf{S}}\left(\phi^{\mathsf{S}}(\mathcal{I}^{T})\Big|_{0}^{\gamma-1}\right)\Big|_{\gamma}^{L}\right) \quad \text{when} \quad \gamma < L \tag{5.2}$$

- **Whole Retrain:**

$$\mathbb{W}_{\mathbb{R}} := \mathcal{A}\left(\phi^{\mathsf{S}}\left(\phi^{\mathsf{S}}(\mathcal{I}^{T})\Big|_{0}^{\gamma-1}\right)\Big|_{\gamma}^{L}\right) \quad \text{when} \quad \gamma = 0 \tag{5.3}$$

The notation described through the operator $\mathcal{A}$ indicates two important points: First, it clearly establishes that the model's original weights are preserved completely from layer 0 to layer γ. Second, it permits to add additional layers on top of the model.

The literature review section revealed two important points:

- ConvNets like DenseNet, ResNet 101, NasNet-mobile, Xception, ResNext have not been studied in mammogram classification.
- As indicated in Dias Pedro et al. (2018), Yassin et al. (2018), there are no standard metrics in mammogram classification research.
- There is a lack of contrast of the automated results against the radiologist.

Because of the aforementioned reasons, it was concluded that a extended systematic study like the one performed in this Thesis should be performed. The results of the experimental findings are discussed next. Unfortunately, this work has not been able to contrast its results with a radiologist team.

5.2.2 Experimenting with 20 pre-trained ConvNets in Transfer Learning and Ensemble Learning for Mammogram Classification

Summary

The proposed methodology consisted in generating the ROI mass dataset for train test and validation from Inbreast and MIAS mammograms. A total of 6 000 images were generated from each database and then split in 80% training (4 800), 10% validation (600) and 10% testing (600) by using the Augmentor library for data augmentation, as described in Section 3.4.4 from Chapter 3. Then, a total of 20 pre-trained ConvNets were trained in three different techniques (TL, FT, WR), as described by (5.1), (5.2), (5.3), in mass pathology classification (benign, malignant). The idea was to perform a systematic comparison of the performance of each pre-trained ConvNet from the Keras library in Tensorflow in each training technique.

This thesis hypothesis proposed that an ensemble of fine tuned models improves the average classification performance for mass pathology. Literature review usually suggests that TL and FT models will perform better than a whole re-trained model. However, the experiments performed in this thesis suggest that even whole retrained models can achieve an $AUC > 90\%$. This is an unexpected result of this research. From the three techniques (TL, FT, WR), FT shows the least overfitting. Ensemble models were trained using two common techniques: hard voting and soft voting. The ensemble was formed by the best classifiers of each technique (i.e TL, FT, WR). For soft voting, this thesis proposed two options. The first one, used a constant value of ones for the weights used to compute the soft voting ensemble as indicated in (2.22). The second option, used a Perceptron to find the weights values. This was called *Automatic Soft Voting Ensemble* (ASVE) or *soft-voting-auto*. ASVE outperformed other single models in the classification task as well as other ensemble learning techniques for each database. Finally, a merged test dataset was created by using the test sets from each database. In order to be able to classify both types of mammograms, an ensemble was formed by merging the three base classifiers from each database. Again, hard and soft voting were used to form the ensemble of the six base classifiers. The best result was achieved by the ASVE model.

Conclusions

Literature suggests that TL and FT are preferable to WR and perform better than WR. However, the results achieved in Tables 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 show that WR achieved the highest AUC classification performance for both databases in mass pathology classification. This surprising contradictory result needs to be explain. First, a close look to the distance between the train and validation curves (e.g. Figures 4.12 and 4.21) show that WR presents a tendency to overfit. Second, different to the works reviewed in literature, this research has trained 20 models in the modalities of TL, FT, and WR. Consequently, what the results may suggest is that it is not that an specific training technique (e.g. TL, FT or WR) is better than an specific any other, but the combination of the training technique and methodology what achieves an specific result. For example, in MIAS, both Mobilenet-TL and Vgg16-WR achieve $AUC = 0.943$. Moreover, Table 4.11 suggests that the probability of success for TL is much larger than FT or WR. In other words, it is more likely that using a TL on a pre-trained ConvNet would succeed in mammogram classification than using FT or WR. The probability of success was estimated as the number of classifiers that achieved over 90% on AUC wrt. the total number of models trained in each test. This confirms, that TL is preferable to WR as generally accepted. The reasons are twofold: first, TL demands less time to be trained because it trains a shorter number of parameters; second, reducing the number of parameters aids to control overfitting. However, it must be noticed that the smallest gap between train an validation curves is achieved by FT.

This work is similar to Chougrad et al. (2018). However, there are some important differences in the model besides experimenting with ensemble techniques. First, the selected layers in operator $\mathcal{A}$ are different from the ones proposed in Chougrad et al. (2018) and include Dropout to control overfitting. The number of neurons used for the FC layer and the value of Dropout were estimated experimentally to improve the results during Pilot Experiments phase. Furthermore, early stopping was used as a regularization technique that stops training when performance deteriorates. Second, a total of 20 models were trained while AlexNet was the most used in literature. Third, additional classification metrics were used including Mathews correlation coefficient.

Data augmentation was of remarkable importance because it stabilized the performance metrics and improved the results. The comparison between Tables 4.1

and 4.2 show that data augmentation improves classification performance in TL. As for the comparison between Figures 4.1 and 4.2, it shows distances between the different metrics when the dataset is unbalanced. Once, the dataset is balanced through data augmentation, the distance between the metrics is reduced, and they level. Consequently, these results suggest that data augmentation is necessary to achieve a good classification performance or that databases with 3 000 images per category are recommended to solve mass pathology classification.
Now, questions 1, 2 and 3 proposed in Section 5.2 can be answered:

1. Is TL on a pre-trained ConvNet able to address mass pathology classification in breast cancer?

   As indicated in Table 4.11, TL is more likely to tune a pre-trained ConvNet in mammography classification than FT and WR.

2. Among all the pre-trained ConvNets models on the ImageNet dataset, is there one that outperforms others in mass pathology classification in breast cancer?

   It was discussed that it is the relationship between model's architecture, methodology and dataset which has an impact in the accuracy of the result. Nevertheless, the results also show that Vgg16, Vgg19, ResNet, ResNext and Mobilenet were the most suit-able ConvNets independently of the training technique (TL, FT, or WR).

3. How should TL or FT be implemented to achieve a good classification result?

   It was discussed that γ has a relation in the ConvNet's performance, as well as the operator $\mathcal{A}$. However, the presented results rely on data augmentation, because the original datasets have few examples. Because of this, it can be said that a balanced dataset with 6 000 samples is needed to achieve good results. Consequently, TL and FT do not need large datasets but, they do need a balanced dataset with enough samples. In this work, enough samples seem to be around 3 000 samples per category.

5.2.3 Designing and Ensemble Model that improves classification performance in Mammogram Images

Summary

The experimental procedure of this work was presented in Figure 3.1. The ensemble generation consisted in selecting the best single classifiers from TL, FT and WR for each database. Consequently, the ensemble was formed by 3 different trained ConvNets, as described in Table 4.12. Two different types of ensemble were used: hard-voting and soft-voting. The former consist in a majority vote ensemble whereas the latter uses a weight sum of the probabilities to estimate the corresponding class. An improvement to the soft voting ensemble was proposed in this work by using a Perceptron to calculate the values of the weights in (2.22). This new ensemble has been called *Automatic Soft Voting Ensemble* (ASVE) or *soft-voting-auto* and has outperformed single base classifiers on Inbreast and MIAS. Furthermore, on the Inbreast dataset ASVE outperformed the classification performance of related works. Finally a mixed dataset from Inbreast and MIAS was created. Consequently, an ensemble consisting of the six base models presented in Table 4.12 was formed. Hard voting and the proposed ASVE achieved the best performance for the mixed dataset.

Conclusions

An exciting product of this work was the *Automatic Soft Voting Ensemble* (ASVE). This ensemble used a Perceptron approach to tune the weight values in (2.22) and outperformed the base model's classification performance, as can be seen when comparing the confusion matrices presented in Figures 4.28 and 4.6). In fact, this work's proposed ASVE model has been able to improve the classification result for Inbreast with respect to the state of the art by achieving $AUC = 0.98$. The results achieved for ensemble confirmed the hypothesis proposed in this work since an ensemble model formed by fine tuned models was able to improve classification performance in mass abnormalities classification. This also answers affirmatively the fourth question proposed in Section 5.2. One interesting argument for the proposed ensemble models is that they were limited to 3 base models for each database and to 6 base models for the mixed ensemble. It is true that more base models could have been used. However, the limitation in the number and type of base models used was due to computation resources.

5.3 Contributions to Knowledge

As a contribution to knowledge, this work presents the following key points:

- A systematic comparison between 20 models in three different training techniques such as TL, FT and WR. This comparison allowed to discover that WR is able to achieve good results in mammogram pathology classification. Another interesting finding is that TL is more prone to success when used as a retraining technique; this was indicated in Table 4.11.
- The *Automatic Soft Voting Ensemble* (ASVE) or *Soft-Voting-Auto* is an interesting new type of soft voting ensemble for binary classification. The proposed algorithm may be used in other binary classification problems.
- This work provided a mathematical definition for TL, FT and WR for the case of computer vision and ConvNets. These definitions are summarized in (5.1), (5.2), (5.3). Also, the provided operator $\mathcal{A}$ is hoped to aid in notation of the changes implemented by researchers as shown in Table 3.5 and in Equations: (3.9), (3.10).

5.4 Recommendations and Future Works

The lack of mammogram databases with a thousands of examples is a limitation that is beyond the control of the present work. In fact, this thesis has shown that training ConvNets with a database with 3 000 images per category achieves good results in most cases. Consequently, there is a need to develop deep learning datasets for mammography to train classification models based on ConvNets. The use of TL and FT are important to train Deep ConvNets with reduced data. Because of that, it may be not necessary to have a million images mammogram dataset. Research should be encouraged to design new mammogram datasets with enough samples and their corresponding ground truths.

This work focused on ROI images corresponding to mass cases only. This decision was due to time constrains and computational resources. The next stages of this work will explore the whole mammogram image containing both mass and calcifications. Remind that radiologist use the BI-RADS scale to assess the malignancy level of breast cancer. Furthermore, future work should also implement

localization and detection problems together with segmentation.

The experiments presented in this thesis can be further explored. For example, the pilot experiments suggest a relation between the classification performance and the number of neurons in the FC layer. Even though, this work presented experimental evidence for this situation, the pilot experiment only used the Mobilenet ConvNet, because it was the fastest model to train. Also, it is of interest to study the effect of changing the image size in the input layer. The pilot experiments used an image size of 224×224, whereas the proper experiments used 320×320. Some of the results on the Inbreast dataset suggest a better performance using 224×224; which is smaller. Another area that should be also explored are the types of augmentation techniques used. In this work, as it was mentioned, the pre-processing of the image implied resizing the image to its final value and normalizing the pixel values in an 8 bit range. This two operations were used due to the literature review where it was found that some image enhancement techniques present contrary results in increasing the classification performance. However, a possible direction, for a future work, will be studying the effect of different image enhancement techniques on the classification result.

A limitation worth to mention is the fact that the experiments carried out in TL, FT and WR need to be repeated and averaged. This was not done due to time constrains. In this work a total of 20 ConvNets were trained for TL, FT and WR. Considering that in FT, three different values of γ were proposed, a total of 100 different trainings were performed per database. Given that training time for each ConvNet is variable and long (e.g. ResNext-101-WR took 4 days and 6 hours for Inbreast), it was time consuming to repeat each training. Nevertheless, this aspect should be reviewed.

The ensemble model proposed in this work and called *Automatic Soft Voting Ensemble* (ASVE) achieved the best results on Inbreast, MIAS and the Mixed Dataset combining both databases. However, further investigations will explore the use of Bagging to form different training sets and repeat the proposed methodology. It is entirely possible to suppose that an ensemble thus formed performs as good as the ensemble designed in this research work for the mixed dataset.

Appendix A

A.1 Inbreast Fine Tuning Table

Table A.1: Inbreast Fine Tuning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
vgg16-8-FT	0.93	0.93	0.928	0.862	0.93
vgg19-11-FT	0.918	0.918	0.916	0.838	0.918
vgg16-6-FT	0.917	0.917	0.913	0.836	0.917
vgg16-10-FT	0.912	0.912	0.911	0.823	0.912
vgg19-17-FT	0.898	0.898	0.899	0.797	0.898
vgg19-20-FT	0.692	0.692	0.695	0.383	0.692
densenet-201-600-FT	0.69	0.69	0.7	0.381	0.69
inception-resnet-v2-775-FT	0.667	0.667	0.729	0.375	0.667
xception-127-FT	0.662	0.662	0.71	0.343	0.662
mobilenet-77-FT	0.658	0.658	0.666	0.317	0.658
xception-122-FT	0.645	0.645	0.718	0.339	0.645
inception-resnet-v2-675-FT	0.643	0.643	0.681	0.295	0.643
mobilenet-82-FT	0.633	0.633	0.692	0.288	0.633
inception-v3-209-FT	0.627	0.627	0.464	0.319	0.627
resnext-101-460-FT	0.625	0.625	0.68	0.266	0.625
resnet-152v2-550-FT	0.625	0.625	0.64	0.251	0.625
densenet-121-325-FT	0.617	0.617	0.698	0.277	0.617
densenet-169-490-FT	0.617	0.617	0.681	0.255	0.617
resnet-101v2-372-FT	0.617	0.617	0.561	0.241	0.617
resnext-50-230-FT	0.612	0.612	0.488	0.255	0.612
resnext-101-472-FT	0.605	0.605	0.636	0.213	0.605
nasnet-l-937-FT	0.602	0.602	0.706	0.289	0.602
resnet-101v2-365-FT	0.602	0.602	0.704	0.281	0.602
nasnet-m-669-FT	0.588	0.588	0.489	0.192	0.588
nasnet-l-1034-FT	0.588	0.588	0.445	0.206	0.588
mobilenet-v2-150-FT	0.588	0.588	0.457	0.202	0.588
mobilenet-v2-145-FT	0.575	0.575	0.552	0.151	0.575
resnet-50v2-185-FT	0.57	0.57	0.678	0.188	0.57
resnet-152v2-562-FT	0.57	0.57	0.48	0.149	0.57

Table A.2: Inbreast Fine Tuning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
resnet-50v2-180-FT	0.57	0.57	0.685	0.206	0.57
resnet-101v2-375-FT	0.562	0.562	0.479	0.13	0.562
inception-v3-306-FT	0.56	0.56	0.631	0.13	0.56
resnet-152v2-559-FT	0.553	0.553	0.239	0.19	0.553
nasnet-l-1037-FT	0.548	0.548	0.474	0.101	0.548
densenet-121-422-FT	0.547	0.547	0.647	0.113	0.547
densenet-169-593-FT	0.54	0.54	0.609	0.086	0.54
resnext-50-234-FT	0.54	0.54	0.216	0.142	0.54
mobilenet-v2-153-FT	0.522	0.522	0.552	0.044	0.522
mobilenet-85-FT	0.51	0.51	0.668	0.066	0.51
resnet-101-340-FT	0.51	0.51	0.428	0.021	0.51
nasnet-m-764-FT	0.507	0.507	0.256	0.018	0.507
inception-v3-309-FT	0.502	0.502	0.258	0.004	0.502
resnet-101-343-FT	0.5	0.5	0.667	0	0.5
resnet-152-510-FT	0.5	0.5	0.667	0	0.5
resnet-50-173-FT	0.5	0.5	0.667	0	0.5
resnet-152-500-FT	0.5	0.5	0	0	0.5
resnet-152-513-FT	0.5	0.5	0	0	0.5
resnet-50-170-FT	0.5	0.5	0	0	0.5
resnet-50-160-FT	0.498	0.498	0	-0.041	0.498
resnext-50-237-FT	0.493	0.493	0	-0.082	0.493
densenet-201-702-FT	0.493	0.493	0.032	-0.044	0.493
densenet-201-705-FT	0.492	0.492	0.007	-0.078	0.492
resnet-101-330-FT	0.485	0.485	0.013	-0.103	0.485
nasnet-m-767-FT	0.472	0.472	0.281	-0.067	0.472
resnet-50v2-187-FT	0.462	0.462	0.403	-0.078	0.462
xception-130-FT	0.452	0.452	0.482	-0.097	0.452
inception-resnet-v2-778-FT	0.442	0.442	0.107	-0.176	0.442
resnext-101-470-FT	0.44	0.44	0.349	-0.125	0.44
densenet-121-425-FT	0.427	0.427	0.185	-0.182	0.427
densenet-169-590-FT	0.427	0.427	0.2	-0.178	0.427

A.2 Mias Fine Tuning Table

Table A.3: Mias Fine Tuning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
vgg16-6-FT	0.935	0.935	0.935	0.87	0.935
vgg16-10-FT	0.928	0.928	0.929	0.857	0.928
vgg16-11-FT	0.925	0.925	0.925	0.85	0.925
vgg16-8-FT	0.922	0.922	0.921	0.843	0.922
vgg19-11-FT	0.902	0.902	0.905	0.805	0.902
vgg19-17-FT	0.872	0.872	0.878	0.747	0.872
xception-122-FT	0.58	0.58	0.629	0.166	0.58
resnet-101v2-375-FT	0.578	0.578	0.677	0.198	0.578
densenet-121-325-FT	0.577	0.577	0.668	0.184	0.577
inception-resnet-v2-775-FT	0.575	0.575	0.655	0.169	0.575
resnet-101v2-372-FT	0.575	0.575	0.654	0.169	0.575
vgg19-20-FT	0.575	0.575	0.624	0.155	0.575
densenet-121-422-FT	0.573	0.573	0.656	0.167	0.573
resnet-50v2-187-FT	0.572	0.572	0.631	0.151	0.572
inception-resnet-v2-675-FT	0.567	0.567	0.67	0.171	0.567
resnet-101v2-365-FT	0.565	0.565	0.672	0.171	0.565
resnet-50v2-185-FT	0.565	0.565	0.642	0.144	0.565
nasnet-l-937-FT	0.563	0.563	0.664	0.158	0.563
nasnet-m-767-FT	0.56	0.56	0.594	0.122	0.56
resnext-101-470-FT	0.558	0.558	0.631	0.127	0.558
nasnet-l-1034-FT	0.553	0.553	0.531	0.107	0.553
resnet-152v2-550-FT	0.552	0.552	0.653	0.127	0.552
mobilenet-85-FT	0.552	0.552	0.669	0.147	0.552
inception-v3-209-FT	0.552	0.552	0.419	0.116	0.552
mobilenet-77-FT	0.545	0.545	0.663	0.127	0.545
resnet-152v2-562-FT	0.537	0.537	0.675	0.141	0.537
inception-v3-306-FT	0.537	0.537	0.658	0.104	0.537
resnet-152v2-559-FT	0.537	0.537	0.55	0.073	0.537
xception-127-FT	0.537	0.537	0.487	0.075	0.537
resnext-101-472-FT	0.528	0.528	0.666	0.1	0.528

Table A.4: Mias Fine Tuning Results on Dataset D_3

$Model$	$BACC$	AUC	F_1Score	MCC	ACC
mobilenet-82-FT	0.528	0.528	0.617	0.064	0.528
resnet-50v2-180-FT	0.527	0.527	0.667	0.1	0.527
xception-130-FT	0.525	0.525	0.643	0.067	0.525
nasnet-m-764-FT	0.522	0.522	0.663	0.079	0.522
densenet-169-490-FT	0.513	0.513	0.646	0.041	0.513
densenet-201-705-FT	0.513	0.513	0.658	0.05	0.513
nasnet-m-669-FT	0.51	0.51	0.668	0.066	0.51
inception-resnet-v2-778-FT	0.51	0.51	0.618	0.024	0.51
mobilenet-v2-145-FT	0.505	0.505	0.667	0.047	0.505
densenet-121-425-FT	0.505	0.505	0.667	0.041	0.505
inception-v3-309-FT	0.505	0.505	0.667	0.041	0.505
nasnet-l-1037-FT	0.505	0.505	0.412	0.011	0.505
mobilenet-v2-153-FT	0.503	0.503	0.632	0.009	0.503
densenet-169-593-FT	0.502	0.502	0.667	0.024	0.502
resnext-50-230-FT	0.502	0.502	0.667	0.024	0.502
densenet-169-590-FT	0.5	0.5	0.667	0	0.5
densenet-201-702-FT	0.5	0.5	0.667	0	0.5
resnet-101-340-FT	0.5	0.5	0.667	0	0.5
resnet-101-343-FT	0.5	0.5	0.667	0	0.5
resnet-152-500-FT	0.5	0.5	0.667	0	0.5
resnet-152-510-FT	0.5	0.5	0.667	0	0.5
resnet-152-513-FT	0.5	0.5	0.667	0	0.5
resnet-50-160-FT	0.5	0.5	0.667	0	0.5
resnet-50-173-FT	0.5	0.5	0.667	0	0.5
resnext-50-234-FT	0.5	0.5	0.667	0	0.5
resnext-50-237-FT	0.5	0.5	0.667	0	0.5
densenet-201-600-FT	0.5	0.5	0.654	0	0.5
resnet-101-330-FT	0.5	0.5	0	0	0.5
mobilenet-v2-150-FT	0.498	0.498	0.665	-0.041	0.498
resnext-101-460-FT	0.498	0.498	0.264	-0.004	0.498
resnet-50-170-FT	0.487	0.487	0.158	-0.043	0.487

Bibliography

Amiri, I. S., Akanbi, O. A., & Fazeldehkordi, E. (2014). *A machine-learning approach to phishing detection and defense*. Syngress.

Al-antari, M. A., Al-masni, M. A., Choi, M. T., Han, S. M., & Kim, T. S. (2018). A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *International Journal of Medical Informatics*, *117*(June), 44–54. doi:10.1016/j.ijmedinf.2018.06.003. arXiv: arXiv:1511.00561

Ayash, E. M. M. (2014). Research methodologies in computer science and information systems. *Retrieved November*, *28*, 2014.

Baum, M., & Henderson, C. (2004). *Classic papers in breast disease*. CRC Press.

Bengio, Y. (1997). *Convolutional Networks for Images, Speech, and Time-Series Unsupervised Learning of Speech Representations View project Parsing View project*. Retrieved from https://www.researchgate.net/publication/2453996

Bengio, Y. [Yoshua], Courville, A. C., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 1798–1828.

Berndtsson, M., Hansson, J., Olsson, B., & Lundell, B. (2007). *Thesis projects: A guide for students in computer science and information systems*. Springer Science & Business Media.

Bloice, M. D., Roth, P. M., & Holzinger, A. (2019). Biomedical image augmentation using Augmentor. *Bioinformatics*. doi:10.1093/bioinformatics/btz259. eprint: http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz259/28554765/btz259.pdf

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018a). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. doi:10.3322/caac.21492

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018b). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. doi:10.3322/caac.21492

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121–3124).

Cadène, R., Thome, N., & Cord, M. (2016). Master's thesis: Deep learning for visual recognition. *arXiv preprint arXiv:1610.05567*.

Canbek, G., Sagiroglu, S., Temizel, T. T., & Baykal, N. (2017). Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 international conference on computer science and engineering (ubmk)* (pp. 821–826). doi:10.1109/UBMK.2017.8093539

Carneiro, G., Nascimento, J., & Bradley, A. P. (2015). Unregistered multiview mammogram analysis with pre-trained deep learning models. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9351, pp. 652–660). doi:10.1007/978-3-319-24574-4_78. arXiv: 1505.04597

Carneiro, G., Nascimento, J., & Bradley, A. P. (2017a). Automated Analysis of Unregistered Multi-View Mammograms with Deep Learning. *IEEE Transactions on Medical Imaging*, *36*(11), 2355–2365. doi:10.1109/TMI.2017.2751523

Carneiro, G., Nascimento, J., & Bradley, A. P. (2017b). *Deep Learning Models for Classifying Mammogram Exams Containing Unregistered Multi-View Images and Segmentation Maps of Lesions* (1st ed.). doi:10.1016/B978-0-12-810408-8.00019-5

Charan, S., Khan, M. J., & Khurshid, K. (2018). Breast cancer detection in mammograms using convolutional neural network. In *2018 international conference on computing, mathematics and engineering technologies (icomet)* (pp. 1–5). doi:10.1109/ICOMET.2018.8346384

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Chollet, F. et al. (2015). Keras. https://keras.io.

Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in*

Biomedicine, *157*, 19–30. doi:10.1016/j.cmpb.2018.01.011. arXiv: 1802.00752

Cohen, P. R. (1995). *Empirical methods for artificial intelligence*. MIT press Cambridge, MA.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*, *2*.

Deng, J. [J.], Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09*.

Denning, P. J. (2005). Is computer science science? *Commun. ACM*, *48*(4), 27–31. doi:10.1145/1053291.1053309

Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, *37*, 114–128. doi:10.1016/j.media.2017.01.009

Dias Pedro, R. W., Machado-Lima, A., & Nunes, F. L. (2018). Is mass classification in mammograms a solved problem? - A critical review over the last 20 years. *Expert Systems with Applications*, *119*, 90–103. doi:10.1016/j.eswa.2018.10.032

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15). Springer.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 69–78).

Dromain, C., Boyer, B., Ferre, R., Canale, S., Delaloge, S., & Balleyguier, C. (2013). Computed-aided diagnosis (cad) in the detection of breast cancer. *European journal of radiology*, *82*(3), 417–423.

Drukteinis, J. S., Mooney, B. P., Flowers, C. I., & Gatenby, R. A. (2013). Beyond mammography: New frontiers in breast cancer screening. doi:10.1016/j.amjmed.2012.11.025. arXiv: NIHMS150003

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159.

Falconí, L. G., Pérez, M., & Aguilar, W. G. (2019). Transfer learning in breast mammogram abnormalities classification with mobilenet and nasnet. In *2019*

international conference on systems, signals and image processing (iwssip) (pp. 109–114). doi:10.1109/IWSSIP.2019.8787295

Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".

Goodfellow, I., Bengio, Y. [Yoshua], & Courville, A. (2016). *Deep learning*. http://www.deeplearningbook.org. MIT Press.

Guan, S., & Loew, M. (2017). Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks. In *2017 ieee applied imagery pattern recognition workshop (aipr)* (pp. 1–8). doi:10.1109/AIPR.2017.8457948

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48.

Hamidinekoo, A., Suhail, Z., Denton, E., & Zwiggelaar, R. (2018). Comparing the performance of various deep networks for binary classification of breast tumours. In *14th international workshop on breast imaging (iwbi 2018)* (p. 39). doi:10.1117/12.2318084

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, W. P. (2000). The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography* (pp. 212–218). Medical Physics Publishing.

Hill, P., & Kanagaratnam, U. (2016). Python machine learning sebastian rashka. *ITNOW*, *58*(3), 64–64.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017a). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017b). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, Y., Li, J., & Jiao, Z. (2016). Mammographic mass detection based on saliency with deep features. In *Proceedings of the international conference on internet multimedia computing and service* (pp. 292–297). ACM.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jalalian, A., Mashohor, S., Mahmud, R., Karasfi, B., Ramli, A. R. B., Engineering, C. S., … Branch, Q. (2017). Review article : FOUNDATION AND METHODOLOGIES IN COMPUTER-AIDED, 113–137.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *2009 ieee 12th international conference on computer vision* (pp. 2146–2153). IEEE.

Jiang, F., Liu, H., Yu, S., & Xie, Y. (2017). Breast mass lesion classification in mammograms by transfer learning. In *Acm international conference proceeding series* (pp. 59–62). doi:10.1145/3035012.3035022

Kalaf, J. M. (2014). Mammography: A history of success and scientific enthusiasm. *Radiologia brasileira*, *47*(4), VII.

Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision*, *8*(1), 1–207.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. In *Engineering* (Vol. 2, p. 1051). doi:10.1145/1134285.1134500. arXiv: 1304.1186

Koehrsen, W. (2018). Transfer learning with convolutional neural networks in pytorch. Towards Data Science. Retrieved from https://towardsdatascience.com/transfer-learning-with-convolutional-neural-networks-in-pytorch-dd09190245ce

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems* (pp. 1097–1105). Retrieved from http://code.google.com/p/cuda-convnet/

LeCun, Y., Bengio, Y. [Yoshua], & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-

propagation network. In *Advances in neural information processing systems* (pp. 396–404).

LeCun, Y., Bottou, L., Bengio, Y. [Yoshua], Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, *4*, 170177 EP -. Data Descriptor. Retrieved from https://doi.org/10.1038/sdata.2017.177

Li, M., Soltanolkotabi, M., & Oymak, S. (2019). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. arXiv: 1903.11680 [cs.LG]

Lukong, K. E. (2017). Understanding breast cancer – The long and winding road. *BBA Clinical*, *7*, 64–77. doi:10.1016/J.BBACLI.2017.01.001

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, … Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. Retrieved from http://tensorflow.org/

Al-masni, M., Al-antari, M., Park, J.-M., Gi, G., Kim, T.-S. T.-Y., Rivera, P., … Kim, T.-S. T.-Y. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Computer Methods and Programs in Biomedicine*, *157*, 85–94. doi:10.1016/j.cmpb.2018.01.017

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: Toward a full-field digital mammographic database. *Academic radiology*, *19*(2), 236–248.

Murphy, J. (2016). An overview of convolutional neural network architectures for deep learning.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 807–814).

Niklason, L. T., Christian, B. T., Niklason, L. E., Kopans, D. B., Castleberry, D. E., Opsahl-Ong, B., … Moore, R., et al. (1997). Digital tomosynthesis in breast imaging. *Radiology*, *205*(2), 399–406.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169–198.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Pattanayak, S. (2017). *Pro deep learning with tensorflow. a mathematical approach to advanced artificial intelligence in python*. Apress. Retrieved from http://gen.lib.rus.ec/book/index.php?md5=da448dc4aabf5cd791ee0823a2809b70

Perez, M., Benalcazar, M. E., Tusa, E., Rivas, W., & Conci, A. (2017). Mammogram classification using back-propagation neural networks and texture feature descriptors. In *2017 ieee second ecuador technical chapters meeting (etcm)* (pp. 1–6). doi:10.1109/ETCM.2017.8247515

Perre, A., Alexandre, L. A., & Freire, L. C. (2018). Lesion classification in mammograms using convolutional neural networks and transfer learning. *Lecture Notes in Computational Vision and Biomechanics*, *27*, 360–368. doi:10.1007/978-3-319-68195-5_40

Rasche, C. (2013). Workbook pattern recognition. Citeseer.

Rasche, C. (2019). *Computer Vision I*.

Reboux, G. (2018). Cancer. Retrieved November 23, 2019, from https://www.who.int/news-room/fact-sheets/detail/cancer

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational intelligence magazine*, *11*(1), 41–53.

Röntgen, W. C. (1896). On a new kind of rays. *Science*, *3*(59), 227–231.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton project para*. Cornell Aeronautical Laboratory.

Russakovsky, O., Deng, J. [Jia], Su, H., Krause, J., Satheesh, S., Ma, S., … Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Samala, R. K., Chan, H. P., Hadjiiski, L. M., Helvie, M. A., Cha, K. H., & Richter, C. D. (2017). Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms.

Physics in Medicine and Biology, *62*(23), 8894–8908. doi:10.1088/1361-6560/aa93d4

Sarkar, D., Bali, R., & Ghosh, T. (2018). *Hands-on transfer learning with python: Implement advanced deep learning and neural network models using tensorflow and keras*. Packt Publishing Ltd.

Selvathi, D., & Aarthypoornila, A. (2017). Performance analysis of various classifiers on deep learning network for breast cancer detection. (Vol. 2018-Janua, pp. 359–363). doi:10.1109/CSPC.2017.8305869

Shams, S., Platania, R., Zhang, J., Kim, J., & Park, S. J. (2018). Deep generative breast cancer screening and diagnosis. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 11071 LNCS, pp. 859–867). doi:10.1007/978-3-030-00934-2_95. arXiv: 15334406

Shanmugamani, R. (2018). *Deep learning for computer vision: Expert techniques to train advanced neural networks using tensorflow and keras*. Packt Publishing Ltd.

Shao, L., Zhu, F., & Li, X. (2014). Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, *26*(5), 1019–1034.

Shapiro, L. (1992). *Computer vision and image processing*. Academic Press.

Simonyan, K., & Zisserman, A. (2014a). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K., & Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K., & Zisserman, A. (2015). *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION*. arXiv: 1409.1556v6. Retrieved from http://www.robots.ox.ac.uk/

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stewart, B. W., Wild, C., International Agency for Research on Cancer, & The World Health Organization. (2014). *World cancer report 2014*. Retrieved from http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014

Stewart, B., Wild, C. P. et al. (2019). World cancer report 2014. *Public Health*.

Stoitsis, J., Valavanis, I., Mougiakakou, S. G., Golemati, S., Nikita, A., & Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial intelligence methods. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *569*(2), 591–595.

SUCKLING J, P. (1994). The mammographic image analysis society digital mammogram database. *Digital Mammo*, 375–386.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2016). *Rethinking the Inception Architecture for Computer Vision*. arXiv: 1512.00567v3. Retrieved from https://arxiv.org/pdf/1512.00567.pdf

Theart, R. (2017). Getting started with pytorch for deep learning (part 3: Neural network basics). Retrieved from https://codetolight.wordpress.com/2017/11/29/getting-started-with-pytorch-for-deep-learning-part-3-neural-network-basics/

Thomassin-Naggara, I., Tardivon, A., & Chopier, J. (2014). Standardized diagnosis and reporting of breast cancer. *Diagnostic and Interventional Imaging*, *95*(7), 759–766. Interventional radiology in oncology. doi:https://doi.org/10.1016/j.diii.2014.06.006

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26–31.

Tucker, A. B. (2004). *Computer science handbook* (2nd ed). Chapman & Hall/CRC. Retrieved from http://gen.lib.rus.ec/book/index.php?md5=9D615734BFB5E8D9657D68E

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLOS ONE*, *9*(1), 1–10. doi:10.1371/journal.pone.0084217

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, *5*, 23.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., & He, K. (2016). Aggregated residual transformations for deep neural networks. corr abs/1611.05431 (2016).

Yassin, N. I., Omran, S., El Houby, E. M., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, *156*, 25–45. doi:10.1016/j.cmpb.2017.12.012

Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, X. [X.], Zhang, Y., Han, E. Y. [E. Y.], Jacobs, N., Han, Q., Wang, X., & Liu, J. (2017). Whole mammogram image classification with convolutional neural networks. In *2017 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 700–704). doi:10.1109/BIBM.2017.8217738

Zhang, X. [X.], Zhang, Y., Han, E. Y. [E.Y. Y], Jacobs, N., Han, Q., Wang, X., & Liu, J. (2017). Whole mammogram image classification with convolutional neural networks. (Vol. 2017-Janua, pp. 700–704). doi:10.1109/BIBM.2017.8217738

Zhang, X. [Xiaoyong], Sasaki, T., Suzuki, S., Takane, Y., Kawasumi, Y., Ishibashiz, T., … Yoshizawa, M. (2017). Classification of mammographic masses by deep learning. In *2017 56th annual conference of the society of instrument and control engineers of japan (sice)* (pp. 793–796). doi:10.23919/SICE.2017.8105545

Zhu, W., Lou, Q., Vang, Y. S., & Xie, X. (2017). Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International conference on medical image computing and computer-assisted intervention* (pp. 603–611). Springer.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8697–8710).