ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 2, 154-165 (2020)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Transfer Learning and Fine Tuning in Breast Mammogram Abnormalities Classification on CBIS-DDSM Database

Lenin G. Falconi[*,1], Maria Pérez[1], Wilbert G. Aguilar[1], Aura Conci[2]

[1]*Escuela Politécnica Nacional, Facultad de Ingeniería de Sistemas, 170517, Ecuador*

[2]*Federal Fluminense University, VisualLab, Computer Institute, 24210-240, Brazil*

ARTICLE INFO

Article history:
Received: 15 January, 2020
Accepted: 18 February, 2020
Online: 11 March, 2020

Keywords:
Transfer Learning
Fine Tuning
Convolutional Neural Networks
Mammogram classification

ABSTRACT

Breast cancer has an important incidence in women mortality worldwide. Currently, mammography is considered the gold standard for breast abnormalities screening examinations, since it aids in the early detection and diagnosis of the illness. However, both identification of mass lesions and its malignancy classification is a challenging problem for artificial intelligence. In this work, we extend our previous research in mammogram classification, where we studied NasNet and MobileNet in transfer learning to train a breast abnormality malignancy classifier, and include models like: VGG, Resnet, Xception and Resnext. However, training deep learning models tends to overfit. This problem is also carried out in this work. Our results show that Fine Tuning achieves the best classifier performance in VGG16 with AUC value of 0.844 in the CBIS-DDSM dataset.

1 Introduction

This work is an extension of our work originally presented in IWSSIP 2019 [1] about mammogram abnormalities classification using *Transfer Learning* (TL) with Mobilenet [2] and Nasnet [3]. In this paper, we also address the classification problem of mammogram abnormalities using the CBIS-DDSM [4] dataset, but we extend the experimentation in transfer learning to other ImageNet pre-trained convolutional neural network (ConvNet) models like: Resnet, Resnext, Xception; to name a few. Finally, *Fine Tuning* (FT) is used in order to address the overfitting problem and improve previous results on the CBIS-DDSM dataset.

Despite the increase in understanding of breast cancer as a disease, it is still a major public health problem worldwide because of the incidence and mortality rates it presents [5]. According to the International Agency for Research on Cancer (IARC), this illness is the second most frequent form of cancer among women worldwide with 2,088,849 (11.6%) new cases and 626,679 (6.6%) of deaths [6]. The mammogram exam remains the gold standard for screening examination, mainly because it is the only screening test that has proven to reduce mortality [7]. However, mammography has some limitations like the variability of its sensitivity, which is inversely proportional to breast density, the false positive and negative rates, and the patient's exposure to radiation [7]. Other screening tests available are: ultrasound, magnetic resonance (MRI), tomosynthesis, and infrared thermography [8]-[9]; in most cases, the aforementioned screening tests are used as adjunct tests.

The mammogram exam diagnostic relies on the radiologist's experience for detection. However, 10% of all woman screened for cancer are called back for additional testing and just as little as 0.5% of them are diagnosed with breast cancer [10]. This shows that it is important to design CAD systems that aid specialists, and train new ones, in breast lesions detection. A "*classic*" CAD system is comprised of 5 main stages: image pre-processing, image segmentation or region of interest (ROI) definition, feature extraction and selection, classification, and performance evaluation [9, 11]. However, this model can be said to be in change due to advances in the field of machine learning, specifically in *Deep Learning*; which allows to automatically learn representations of data with multiple levels of abstraction through deep convolutional neural networks [12]. For instance, in the field of computer vision, the classification of natural images has shown an incredible increase in performance since 2012, when the AlexNet ConvNet model to classify natural images in 1000 categories presented in [13] achieved a 15.3% top 5 test error rate. As a matter of fact, the stages of feature extraction and classification can be solved directly by a ConvNet [14]. This reduces the need for feature hand engineering, which was tradition-

[*]Corresponding Author: Lenin G. Falconi, lenin.falconi@epn.edu.ec

ally used to create the feature vector, because a ConvNet is able to synthesize its own feature vector [15]. All of this confirms the point indicated in [11] that the development in both image techniques and computer science enhance the interpretation of medical images.

The success of ConvNets and deep learning in computer vision tasks such as image classification heavily relies on the number of examples used in training the model under the supervised learning paradigm [16]. Unfortunately, mammogram public datasets are not "deep enough". In this context, transfer learning and fine tuning are deep learning techniques that can aid the development of accurate enough classifiers by transferring knowledge from another domain where large datasets are available. One of the main problems when dealing with small number of examples in training is overfitting. Transfer Learning and Fine Tuning aid in overcoming this disadvantage of working with ConvNets.

In this work, we aim to classify region of interest images from mass tumors of the CBIS-DDSM [4] dataset. We extend our previous experimentation presented in [1] by using different ConvNets models and also the Fine Tuning technique in order to increase the performance of the classifier of breast mammogram abnormalities in benign and malignant. Our research results indicate that Fine Tuning is able to train an accurate classifier and overcome overfitting. Also, we have included the ROC curve metric to measure the performance of the classifiers here studied.

The remainder of this paper is organized as follows: In Section 2, we perform a review of machine learning concepts related to the current research; specifically convolutional neural networks, transfer learning and fine tuning. Literature review of related works in the field is in Section 3. Our proposed experimental method, dataset, and model are presented in Section 4. Section 5 presents experimental data results. Finally, discussion and future works are presented in Section 6 and 7, respectively.

2 Deep Learning Background

2.1 Convolutional Neural Networks

Models based on the Convolutional Neural Networks (ConvNets) architecture have been able to achieve high accurate results in image classification and detection tasks in the ImageNet[17] dataset, under the supervised learning paradigm and back-propagation. That is the case of residual networks, proposed in [18], which achieved a 3.57% error rate in the ImageNet test set in 2015.

Traditional pattern recognition classifiers rely on a hand designed feature extractor that derives relevant information from the raw input data [16]. Thus, the feature extraction step aims to reduce the dimension of the data while characterizing the raw input data (image, sound, etc.) meaningfully so that a trainable classifier is able to categorize its feature vectors [16, 19]. However, the design of the feature extractor requires specialized knowledge about the data (hand-engineering) that, in some cases, could be unknown [19]. On the contrary, ConvNets eliminate the feature extraction process by absorbing it in their architecture [16]. As pointed out in [14], the structure of a ConvNet combines both: the feature extraction and classification steps in one single model that is trained on back-propagation; the feature extraction task is, therefore, learned from data in the first layers of the model, while the last full connecting layers constitute the classifier task. The LeNet-5 ConvNet, proposed in [15] to solve the handwritten classification task, reduces the input image of 32×32 into a 120 vector that is called *the feature vector*. Thus,the feature vector can be used with any type of trainable classifier to solve the classification task. In fact, this approach is used in [19], where LeNet-5 is used as a black box feature extractor for several Support Vector Machines (SVM) that are trained based on it. Something similar is performed by [14], where the authors build an AlexNet [13] like model, trained it on a large dataset, and then use the trained model as a feature extractor to train new classifiers; however, their approach is more similar to a transfer learning set-up, as it will be discussed in Section 2.2.

The deepness in the number of layers of the ConvNets has been increasing in order to obtain better results since year 2012. However, the number of layers cannot be increased indefinitely due to the vanishing and exploding gradient problems. In order to overcome this problem, the structure of the traditional ConvNet, comprised basically of convolutional and pooling layers, has been revisited. An example of those architecture designs are found in Residual Networks[18], MobileNet[2], Inception[20] and NasNet[3] ConvNets. Also the study of regularization functions has been of aid in avoiding the overfitting of the Networks in training [21, 22]. For a review of the state of the art in Convolutional Neural Networks the reader may review the works of: [12, 15, 23, 24].

Thus, ConvNets have some advantages compared with traditional artificial neural networks (ANN): reduction of training parameters by shared weights, local connections and object location invariance [12]. An ANN depends on all the connections between its layers; which increases the number of parameters to train, and makes the training of the model more expensive computationally. On the contrary, a ConvNet reduces the number of parameters trained because the convolution operation is of the local type. Yet, the main disadvantages of the ConvNet model are: the training time, which may be large since it is an hyper-parametrized model, and the susceptibility to overfitting. The most important aspect of ConvNets, as a deep learning method, consists in being a *Representation Learning* method that automatically discovers a representation of the data that is used for classification and detection tasks [12], as previously discussed. This is specially important because it implies that the need for carefully hand engineer feature extractors is not required. Remind that the feature extractor is embedded in the design of the ConvNet and, therefore, learned from data.

The success of a ConvNet relies in its architecture as well as on training the model with enough number of samples. Unfortunately, public mammogram datasets do not have as many examples as the number used per category in the ImageNet dataset. In order to overcome this difficulty, TL and FT are studied as a means to train deep ConvNets in order to classify mammogram abnormalities. In this Section we review the concept of TL and FT as it is used in this article.

2.2 Transfer Learning

A definition of the term is found in [25] and [26]. In their work, the purpose of TL is defined as to improve the performance of a learning algorithm in a target learning task $\mathcal{T}^T$ (i.e. pathology classification) over a target domain $\mathcal{D}^T$ (i.e. mammogram ROI images) by using

the knowledge of the learning algorithm trained in a source learning task $\mathcal{T}^S$ (i.e. 1000 category classification) over a source domain $\mathcal{D}^S$ (i.e. natural images) which is larger than the target domain where:

$$\mathcal{D}^S \neq \mathcal{D}^T \quad (1)$$

$$\mathcal{T}^S \neq \mathcal{T}^T \quad (2)$$

Depending on the relations defined by (1) and (2), different categories of TL are defined in literature. However, it is important to consider that computer vision tasks are particular and different from their corresponding data mining tasks. Thus, despite the fact that a mammogram image is very different from a natural image, the visual properties of objects in an image are general (i.e. edges, textures, shapes, etc.).

In a ConvNet, the knowledge is represented by the value of the weights trained by the back-propagation algorithm on each layer. Therefore, TL implies using the pre-trained ConvNet as a feature extractor or replacing the last original layer with a set of layers that are trained to obtain the target learning task desired. The latter is the approach that we have followed in our experiments. One of the main advantages in this technique is that training time is reduced by not re-training the whole ConvNet, but only the added layers or a trainable classifier in the case of using the pre-trained ConvNet as a feature extractor itself.

2.3 Fine Tuning

In this case, some of the last layers of the pre-trained ConvNet are re-trained with the new images $\mathcal{I} \in \mathcal{D}^T$. Thus, the ConvNet is divided in two parts. Let us define γ as the layer from which the re-training of the ConvNet will occur. If the original ConvNet model has L layers, similarly to TL, we can replace the last original layer and add some layers in order to obtain the desired target learning task. Differently to TL, we also choose a r number of layers before the last one that are also to be trained. This means that the original weights from layer 0 to layer $\gamma - 1$ are preserved or frozen. FT presents more computing resources and time training since the number of parameters to be trained is increased by the r layers that are added to the training queue.

2.4 Over and Underfitting

A machine learning algorithm may suffer of two problems when training: overfitting and underfitting. The former reduces the capacity of the model to predict new unseen data which means that the model has a high variance. The latter, means that the model is not complex enough to reflect the nature of the data and find a pattern [27]. ConvNet models are characterized by being over-parameterized; which means that the parameters of the model exceed the size of the training data [28]. The overfitting problem is reflected when plotting the train vs validation accuracy curve of the model. The difference between the train and validation curve should be minimum. In order to overcome overfitting in ConvNets, Data Augmentation, Regularization, and Early Stopping are usually used. Data augmentation is a basic strategy that consists in increasing the size of the dataset by performing transformations to the original images (i.e. rotation, zoom, reflection, etc). On the other hand, regularization techniques aim to penalize extreme parameter weights values (e.g. L_2 regularization) [27] or controlling the co-adaptation between neurons (e.g. Dropout) [13]. Early Stopping is also considered a regularization technique which aims to interrupt training when the performance of the ConvNet degrades on the validation set. This prevents that the model learn a form of statistical noise [29].

As discussed earlier, TL and FT may also prevent overfitting since the CovNet model is not whole retrained; in other words, TL and FT have less parameters to learn compared to training a model from randomly initialized weights. In the present work, Dropout [30] and Early Stopping have been used as regularization techniques altogether with data augmentation; these are changes introduced in this work that differ from our previous experiment.

3 Related Works

3.1 Search Process

In our previous work [1], we used the methodology by [31], in order to find relevant works for study. Table 1 shows the search string designed to retrieve information from: Springer Link, Science Direct (Elsevier), IEEE Xplore, Scopus, Web of Science, ACM digital library, and PubMed. A total of 174 studies (including our previous work) were gathered from each repository as shown in Table 2. From these studies, a total of 32 primary documents were retrieved according to a selection study process where documents should have experimental methodology with results regarding the use of transfer learning in mammogram breast cancer classification.

Table 1: Search String

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

Table 2: Search Results

Database	**#Publications: 2014-2018**
Scopus	38
IEEE	12
Science Direct	51
PubMed	10
ACM	25
Web of Science	23
Springer	15
Total	**174**

3.2 Literature review discussion

In our literature review, the most common ConvNet used for transfer learning is the model proposed in [13], named *AlexNet*, with a total

of 12 cases. The second most frequent model found is VGG16. These results are shown in Figure 1.

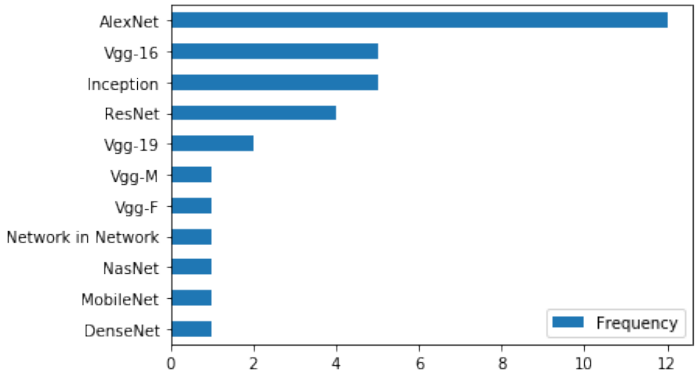

Figure 1: ConvNets used in Literature. This bar plot shows the frequency of the pre-trained ConvNets used in the primary documents retrieved for our literature review. The most used model corresponds to AlexNet (12 cases). At a second place are VGG16 and Inception (5 cases each). Next is ResNet (4 cases) and VGG19 (2 cases).

In mammogram mass abnormality classification and detection there are two main approaches found in literature: a) Processing the whole mammogram image and b) processing the region of interest. The former is found in [32, 33]. Their aim is to find an "*end to end design*". Our approach belongs to the latter, where the ROI image is extracted. In fact processing a whole mammogram images in their original size seems to be a problem itself because mammogram images far exceed the traditional size used in many trained ConvNets in the ImageNet dataset. An interesting approach for an end to end design is proposed in [34]; In their work, the principles of the YOLO [35] architecture are used. However, the mammogram is also resized to 448×448.

As in the case of mammogram classification, Transfer Learning and Fine Tuning are used differently by different authors in literature. In the next subsection this differences are enlightened and discussed.

3.2.1 *Transfer Learning as a Feature Extractor*

In this case, the pre-trained ConvNet is used to extract a feature vector which is later used to train another kind of classifier algorithm like Support Vector Machines (SVM). This case is illustrated in [36]; the author extracts several feature vectors from different layers of the pre-trained AlexNet and trains Support Vector Machines (SVM) for each case. In the end, the author builds an ensemble of SVM. Other similar examples are found in [37, 38].

3.2.2 *Transfer Learning as a new ConvNet Classifier*

In this case, the last full connecting layer of the pre-trained ConvNet may be substituted with a set of additional layers, where the last full connecting layer has only one neuron and the logistic regression for binary classification, or just the number of random initialized neurons required in proportion to the new classification task. For instance, in the case of benign vs malignant classification of the mammogram abnormality this can be achieved with a single neuron or two. Only the added layers are trained while the rest of the ConvNet's weights remain frozen. This approach is tested in both our current and previous work [1]. In a similar fashion, VGG16[39], GoogLeNet[40] and AlexNet[13] are trained in TL in [41].

3.2.3 *Transfer Learning as weight initialization*

In this case the whole ConvNet is re-train but uses the values of the ImageNet pre-trained ConvNet model as initial values for the weights. The last full connecting layer with 1000 categories is substituted by one or two neurons to address the binary classification problem [42]-[43].

3.2.4 *Fine Tuning*

This is the most common technique found in literature. In this case, the model's last full connecting layer is substituted with the number of neurons needed for the new classification or a set of new layers are added before the output layer. Differently to *transfer learning as a new ConvNet*, some of the last layers of the model are re-trained with the new data as indicated in Section 2.3. For example, VGG16[39], InceptionV3[20], and ResNet50 [18] are fine tuned in [44]; the author found that when the number of convolutional blocks exceeds 2, the accuracy of the fine tuned model drops. Also, a comparison of the classification performance between the training of VGG16 in FT and using it as a feature extractor is explored in [45].

3.2.5 *Data Augmentation and Pre-processing*

In literature there is some discussion about the impact of both data augmentation and pre-processing of the medical image. As stated by [46], the achievements in medical images visual tasks with deep learning do not only rely in the ConvNet model but also in the pre-processing of images. For instance, some of the pre-processing methods found in literature are: global contrast normalization (GCN), local contrast normalization, and Otsu's threshold segmentation. However, there is some discussion about improving the image quality. In [37] is reported that global contrast normalization did not aid in improving the experimental results presented in the paper.

Since datasets are not so large, data augmentation is used by almost all researchers. Some of the most common techniques used are: rotations and cropping. However, the rotation operation yields to distortion of the original image. Because of this reason, right angle rotations are preferred to random rotation angles [37, 43, 44].

4 Proposed Approach

4.1 *Transfer Learning and Fine Tuning Model*

The problem to solve is to classify ROI patch mammogram images $\mathcal{I}$ in two classes $\mathcal{Y} = \{benign, malignant\}$. In a supervised learning paradigm this means to find a prediction function $\phi(\cdot)$ that maps an input space $\mathcal{X}$ formed by ROI patch mammogram images ($\mathcal{I} \in \mathcal{X}$) to the output space $\mathcal{Y}$ as indicated in (3)

$$\mathcal{Y} = \phi(\mathcal{X}) \tag{3}$$

However, since TL and FT are to be used to improve the performance of $\phi(\cdot)$, 3 may be written as indicated in 4

$$\mathcal{Y}^T = \phi^{\mathrm{T}}(\mathcal{X}^T) \tag{4}$$

Function $\phi^{\mathrm{T}}(\cdot)$ is to be trained by using pre-trained ConvNets on the ImageNet dataset, which is denominated as $\mathcal{X}^S$. Therefore, our approach satisfies the relations indicated in (1) and (2), since both images and the classification task are different between source and target.

Our approach in both TL and FT consists in replacing the last full connecting layer related to the original ImageNet classification task with a set of layers as indicated in Table 3. The global average pooling (GAvg) layer helps to flatten the original model layer previous to the 1000 full connecting SoftMax classification. The classification layer is comprised of 1 neuron and the Sigmoid function.

Table 3: TL and FT Output

Global Average Pooling 2D
Full Connecting
Dropout
Classification Layer

In TL, only the last layers indicated in Table 3 are to be trained. In FT, we define the γ value that indicates the layer from which the training of the weights is to be performed. It is important to remember that all weights before γ remain with their original value from the ImageNet.

4.2 Dataset

In the present study, as well as our previous work, we use the Curated Breast Imaging Subset of DDSM (CBIS-DDSM)[4] which is an updated and standardized version of the Digital Database for Screening Mammography (DDSM). The dataset includes a subset of the DDSM data selected and curated by a trained mammographer. For our experiments we extract the ROI images from the mammogram images. We have only considered mass problems, leaving micro calcifications for a future work. The dataset is originally organized in train and test sets. The number of images per abnormality class and set type is shown in Table 4.

Table 4: Mass Images in CBIS-DDSM

Dataset Set Type	Benign	Malignant
Train	681	637
Test	231	147

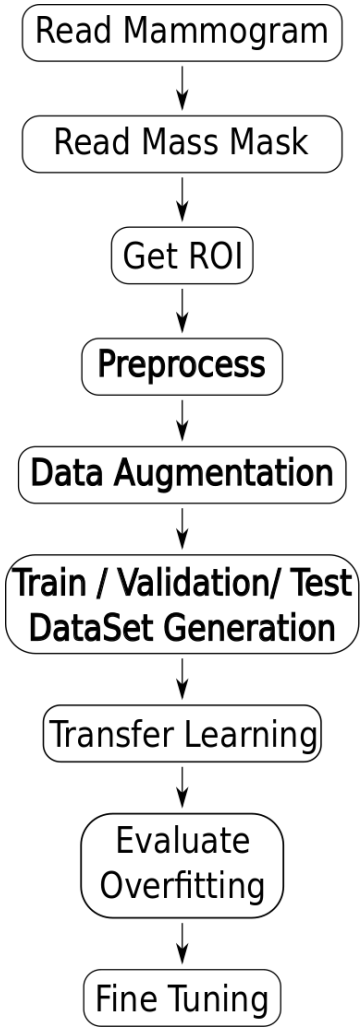

Figure 2: A diagram depicting the methodology followed in our research. First, mammogram and corresponding mass binary mask are read. The mask is used to extract the ROI from the mammogram image. Next, the ROI image is enhanced (pre-process). As a third step, data augmentation is used to increase the number of samples and create train, validation and test sets. Finally, transfer learning is performed and evaluated. The best non overfitting model is selected to be fine tunned

4.3 Methodology

The methodology followed in the current experiments is shown in Figure 2. First, the ROI images are extracted from the mammogram image by using the binary segmentation masks provided in the CBIS-DDSM dataset. After that, images are pre-processed to enhance contrast. Next, a single dataset is formed in order to use data augmentation and create three sets of data: train, validation and test. In this work, we do not use Otsu algorithm to segment the previously obtained ROI by creating an intermediate binary mask. This is because our previous work showed that training with the ROI image segmented with Otsu did not overcome the results obtained with the original background image. Finally, the models are trained in TL and FT, and their performance is evaluated. These steps are described in more detail below.

4.3.1 Image Pre-processing

In this section we present the steps performed in the pre-processing stage of our proposed method. The original mammogram image

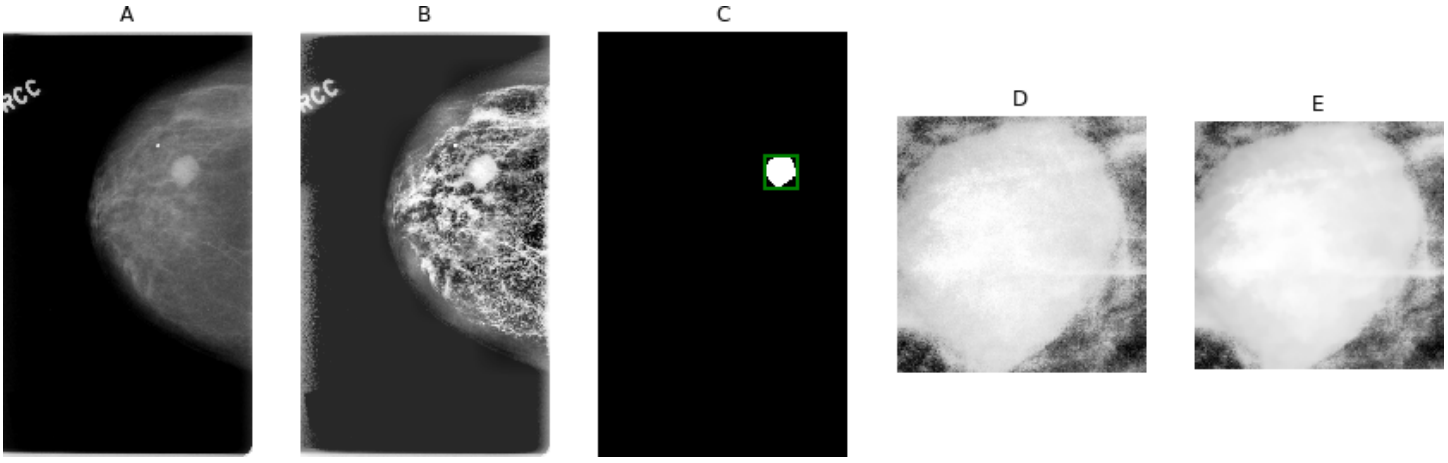

Figure 3: Image Preprocessing. A. The original CBIS-DDSM ROI mammogram image (4704×2744) normalized in UINT 8 resolution ($pixel\_value \in [0, 255]$). B. Mammogram image contrast is improved by using The CLAHE equalization algorithm. C. The binary mass mask provided in the dataset is used to find the ROI. D. The ROI is extracted from B step with a size corresponding to the Bounding Box (358×346) found in the binary mask. E. The ROI image in previous step is resized to 328×328, where we used a 8 pixel padding. Finally, the image is cropped around the center to obtain the final desired training dimensions for the ROI image of 320×320 and Fast Non Local Means Denoising Algorithm is used to further filter ROI image scanning noise.

has a pixel value which ranges up to 65 535 (i.e. 16 bit resolution). In order to generate a dataset of PNG images saved in disk, we change the resolution of the images by normalizing them between 0 and 255, considering the minimum and maximum values of pixel in the original DICOM mammogram image. After that, and different to our previous work, we used the Contrast Limited Adaptive Histogram Equalization (CLAHE)[47] to improve image quality. By using the provided binary masks, we extract the ROI through the coordinates of a bounding box around the suspicious mass. A second normalization of the pixel value is carried out on the ROI considering the minimum and maximum pixel values in it. This originates ROI images with different width and height sizes. As in our previous work, aspect ratio is considered. In (5), the aspect ratio is defined; where r is the aspect ratio, w and h are the width and height of the image respectively.

$$r = w/h \tag{5}$$

Aspect ratio was considered previous to resizing the image in order to preserve the best quality possible from the original image in both, upsampling and downsampling procedures. For upsampling, cubic interpolation was used, whereas for downsampling, area interpolation gives best results. Also, images with an aspect ratio inferior to 0.4 and superior to 1.5 where removed from dataset.

$$\mathcal{I} = \{I | I \in \mathbb{R}^{w \times h}, 0.4 \leq r \leq 1.5\} \tag{6}$$

ROI images were resized to a final size of 320×320. This was achieved only with images whose aspect ratio was inside the limits presented in (6). Resizing the ROI images consisted in two parts: 1) we resized the ROI image to 328×328, 2) cropping around the center of the image. In other words, we have resized the original ROI image to a bigger size (with a padding of 8 pixels) and then cropped in the center of the image to obtain the desired size. Finally, the image is filtered with fast non local means denoising algorithm [48]. This is because CBIS-DDSM images are film mammography and appear to have some noise that has not been removed from the image. Figure 3 presents the steps carried out on image pre-processing for a sample of the training set. Both, the full mammogram image and its corresponding binary mass mask are shown in sub-figures A and C respectively. Image C shows the identification of the bounding box. B presents the application of CLAHE on the mammogram image. Finally, E presents the processed ROI image after performing crop center on D which was extracted from B through the bounding box.

4.3.2 Data Augmentation and Dataset Generation

In our previous work we did not use data augmentation and our models presented overfitting. In order to overcome this difficulty, we implemented the output structure indicated in Table 3 which uses dropout. As pointed out in Section 3.2.5, performing transformations over the images distorts them. Because of that, some researchers use right angles.

Table 5: 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

In our case, we have used the Augmentor Library [49], which has been designed to permit rotations of the images limiting the degree of distortion. Additionally, the Augmentor library permits to apply other operations for data augmentation like zoom, bright, shear. The function uses a probability value to control the number of artificially created images. To augment the dataset, we first join both Train and Test sets. The dataset was increased to a total of 60 000 images, where 80% of the images are used for training, 10%

for validation and 10% for testing. The augmentation operations used are depicted in Table 5.

4.3.3 Model Training

The generated augmented dataset is tested first in Transfer Learning. A total of 4096 neurons are used for the FC layer, a dropout value of 0.2, and a single neuron in the output layer (or classification layer, see Table 3) for binary classification. Early Stopping, with a patience of 50, was enabled in order to stop training when the performance degrades. A maximum number of 1 000 epochs is proposed. The learning rate for TL is 1×10^{-5}. The loss function is set to binary cross entropy and the optimization algorithm used is RMSProp [50]. Binary cross entropy was chosen because the classification layer of our model uses the Sigmoid activation function to discriminate between benign and malignant mass pathology by using one neuron. RMSProp, an adaptive gradient algorithm, is frequently used in computer vision tasks [51]. For instance, the results achieved in [52] indicate a better training result obtained by using RMSProp instead of Stochastic Gradient Descent (SGD). Also, as presented in [53], RMSProp outperforms other common optimization algorithms [51]. This has inspired the theoretical research in [51], where the authors establish the reasons of the success of RMSProp in deep learning and propose new algorithms for optimization. In our study, we consider that RMSProp is convenient for our image classification task because of the aforementioned reasons.

Transfer Learning was carried out in 20 models provided by the Keras API in Tensorflow v1.13.1 [54]. The models used are shown in Table 6.

Table 6: 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

Each model performance is evaluated in the train-validation curve of accuracy. If the difference between train accuracy and test accuracy is over 10%, we consider that overfitting has occurred and the model is rejected. Once a suitable not overfitting model is found, FT is performed in order to further increase the performance of the selected classifier.

FT training parameters are: learning rate of 2×10^{-7}; Dropout remains at 0.2. Respect to the FC neurons, only Global average pooling is performed. Binary cross entropy is set as the cost function and the optimization algorithm used is also RMSProp.

4.3.4 Computing Resources

In order to train deep learning models it is necessary to use Graphical Processing Units (GPU). In our case we have used: Nvidia Tesla K80, with 12 GB of memory, Nvidia Tesla K40, with 12 GB of memory, and Nvidia GeForce RTX 2080, with 8 GB of memory.

Regarding software resources, our program used Tensorflow v13.1 [54] as the machine learning framework. For image data augmentation, as indicated before, the Augmentor library [49] was used.

5 Experimental Results

In this Section, we present our experimental results. As a difference wrt. to our previous work, we include additional metrics to evaluate the performance of the classification model such as the area under the ROC curve and the F_1 Score. These metrics are described below. According to the methodology proposed in Figure 2, it is important to estimate the overfitting of the trained model $\phi^{\mathrm{T}}(\cdot)$. In order to do so, let us define β as the overfitting ratio by comparing train (*train_acc*) and validation accuracy (*valid_acc*) as in

$$\beta = \frac{train\_acc}{valid\_acc} \tag{7}$$

If $\beta \approx 1$, we could say that *train_acc* $\approx$ *valid_acc* and therefore that there is little overfitting. Values of $\beta > 1$ will reflect that there is a considerable difference between train and validation accuracy, meaning that the model has overfitted.

5.1 Performance Metrics

The confusion matrix compares both the prediction of the trained classifier and the true labels provided in the test set. It consists of four main measures: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The total positive cases are $P = TP + FN$; similarly, the total negative cases: $N = TN + FP$. From these measures, the true positive rate (8), false positive rate (9), and true negative rate (10) are derived.

$$TPR = \frac{TP}{P} \tag{8}$$

$$FPR = \frac{FP}{N} \tag{9}$$

$$TNR = \frac{TN}{N} \tag{10}$$

The elements over the diagonal belong to TP and TN and reflect all the correct classifications made by the model. The FP and FN correspond to wrongly classified cases. For instance, FN corresponds to a true malignant tumor that is classified as benign. From these measurements, metrics like: Accuracy (11), F_1 Score (12), and the area under the ROC curve (13) are defined.

$$ACC = \frac{TP + TN}{TP + TN + FP + TN} \tag{11}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$AUC = \frac{1}{2}\left(1 + TPR - FPR\right) \tag{13}$$

5.2 Transfer Learning Experiment

The ImageNet pre-trained ConvNets presented in Table 6 are trained under TL to predict mammogram abnormalities classes in mammogram roi images.

Table 7: Transfer Learning Results

Model	*AUC*	F_1Score	*ACC*
resnet-50-TL	0.861	0.858	0.861
mobilenet-TL	0.854	0.855	0.854
resnet-152-TL	0.853	0.847	0.853
resnet-101-TL	0.846	0.848	0.846
resnet-152v2-TL	0.844	0.845	0.844
resnet-50v2-TL	0.843	0.854	0.843
resnet-101v2-TL	0.832	0.833	0.832
densenet-201-TL	0.821	0.823	0.821
xception-TL	0.816	0.82	0.816
densenet-169-TL	0.812	0.816	0.812
resnext-101-TL	0.806	0.82	0.806
nasnet-l-TL	0.804	0.802	0.804
mobilenet-v2-TL	0.791	0.801	0.791
densenet-121-TL	0.79	0.79	0.789
inception-v3-TL	0.778	0.781	0.779
resnext-50-TL	0.777	0.764	0.777
inception-resnet-v2-TL	0.744	0.746	0.744
nasnet-m-TL	0.712	0.73	0.712
vgg16-TL	0.644	0.654	0.644
vgg19-TL	0.629	0.659	0.629

Table 8: Transfer Learning Overfitting Ratio β

Model	*train_acc*	*test_acc*	β
vgg16-TL	0.65	0.64	1.01
vgg19-TL	0.64	0.63	1.02
resnet-50-TL	1.00	0.86	1.16
mobilenet-TL	1.00	0.85	1.17
resnet-152-TL	1.00	0.85	1.17
resnet-101-TL	1.00	0.85	1.18
resnet-152v2-TL	1.00	0.84	1.19
resnet-50v2-TL	1.00	0.84	1.19
resnet-101v2-TL	1.00	0.83	1.20
inception-resnet-v2-TL	0.90	0.74	1.21
densenet-201-TL	1.00	0.82	1.22
xception-TL	1.00	0.82	1.22
resnext-101-TL	0.99	0.81	1.22
densenet-169-TL	1.00	0.81	1.23
densenet-121-TL	0.98	0.79	1.24
nasnet-l-TL	1.00	0.80	1.24
mobilenet-v2-TL	1.00	0.79	1.26
resnext-50-TL	0.98	0.78	1.26
inception-v3-TL	0.99	0.78	1.27
nasnet-m-TL	0.91	0.71	1.28

As indicated in Section 4.3.3, 4096 output neurons and a dropout rate of 0.2 are used. The results achieved are shown in Table 7. The overfitting ratio stated in (7) is presented in Table 8. The results indicate that Resnet-50 has the best AUC, however its β shows that there is overfitting. On the contrary, both VGG16 and VGG19 do not present overfitting, but their classification performance is lower compared to Resnet-50. These results are reflected in Figure 4 and 5, where the plot train vs test accuracy is presented for Resnet-50 and VGG16, respectively.

5.3 Fine Tuning Experiment

According to the results indicated in Tables 7 and 8, we proceeded to train the VGG16 in Fine Tuning. Different deepness levels where tried in order to search for classification performance improvement. This is indicated through the γ value. VGG16 was trained from layers $\gamma = 8$, $\gamma = 10$. In order to denominate the trained model, we propose to use the pre-trained ConvNet name followd by the layer from which fine tunning occured and added the keyword FT to distinguish the model from those trained in TL mode. For instance, *VGG16-10-FT* means that VGG16 was fine tuned from layer 10. Similarly, we trained VGG19 at $\gamma = 17$. In all cases, only global average pooling followed by dropout and the classification layer were used; except for the case of *VGG16-8-FT*, where a full connecting layer of 4096 neurons was used.

The results obtained are shown in Table 9. We have complimented the experimental results with the Fine Tuning of models: Xception, Resnet101, Resnet152 and Resnet50. It is observable that the best results are achieved by VGG models. The best result achieved corresponds to the VGG16-8-FT. However, Table 10 suggest that the second best result (VGG16-10-FT) has less overfitting and therefore is preferred. In fact β value for VGG16-10-FT and VGG19-17-FT is similar, but the performance of the latter is poorer.

In Figure 6, the plot of train and validation accuracy vs the number of epochs for VGG16-10-FT is presented. Figure 7 shows the ROC curve obtained, whereas Figure 8 presents the confusion matrix for the test set generated.

Table 9: Fine Tuning Results

Model	*AUC*	F_1Score	*ACC*
vgg16-8-FT	0.844	0.85	0.844
vgg16-10-FT	0.816	0.822	0.816
vgg19-17-FT	0.774	0.774	0.774
xception-127-FT	0.571	0.664	0.571
xception-122-FT	0.526	0.67	0.526
xception-130-FT	0.504	0.663	0.504
resnet-101-343-FT	0.5	0.667	0.5
resnet-101-340-FT	0.5	0.667	0.5
resnet-152-510-FT	0.5	0.667	0.5
resnet-50-173-FT	0.5	0.667	0.5
resnet-101-330-FT	0.5	0	0.5
resnet-152-513-FT	0.5	0	0.5
resnet-50-160-FT	0.5	0	0.5
resnet-50-170-FT	0.5	0	0.5

Table 10: Fine Tuning Overfitting Ratio β

Model	*train_acc*	*test_acc*	β
vgg19-17-FT	0.82	0.77	1.07
vgg16-10-FT	0.87	0.82	1.07
vgg16-8-FT	0.92	0.84	1.09

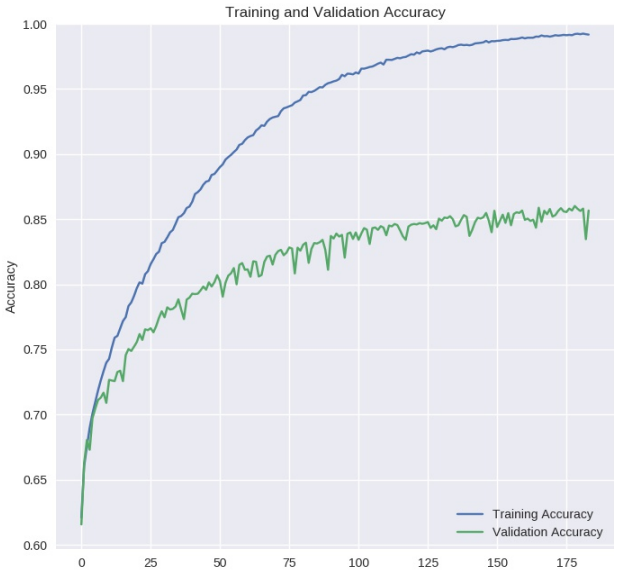

Figure 4: Comparison of the train and validation accuracy for **Resnet50-TL** when trained in transfer learning. As can be seen, despite that Resnet50 achieves an $ACC = 0.86$ accuracy, the model presents overfitting due to the distance between both curves, described by the ratio $\beta = 1.16$

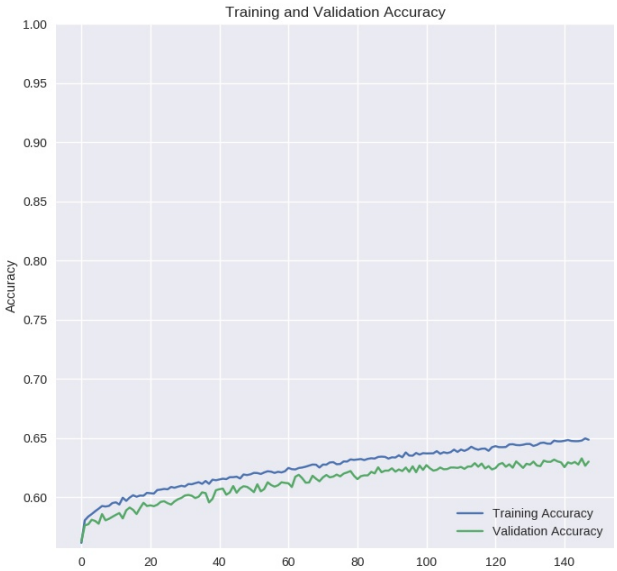

Figure 5: Comparison of the train and validation accuracy for **Vgg16-TL** when trained in transfer learning. VGG16 achieves a lower level of accuracy compared to Resnet50 ($ACC = 0.64, \beta = 1.01$) but there is no overfitting, due to the little distance between train and validation curves.

Figure 6: Fine Tuning VGG16 achieves the best result for our dataset based on CBIS-DDSM ($ACC = 0.87, \beta = 1.07$). Fine Tuning has overcome overfitting and managed to increase classification performance as comparing the train and validation curves for the accuracy metric show in this figure.

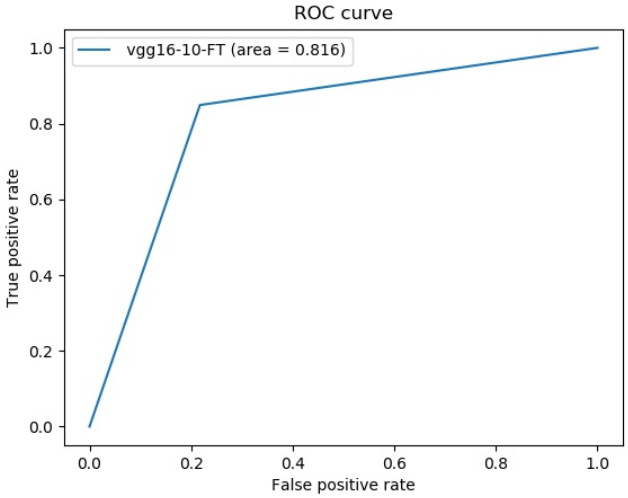

Figure 7: ROC Curve for the binary classification of ROI mammogram images test set. It depicts the performance of the Fine Tuned VGG16-10-FT model, which achieved $AUC = 0.818$

Fine Tuning the VGG16 model from layer 10 ($\gamma = 10$) helps to increase the performance of the classifier while controlling the overfitting ($\beta = 1.07$), which means that train accuracy is 7% over the test accuracy.

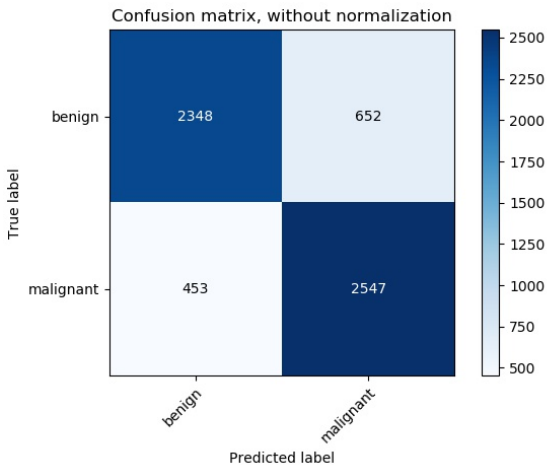

Figure 8: Confusion matrix of the VGG16-10-FT over the test set with 3000 images per category (malignant, benign). In this matrix, True Positive cases are 2547 (84.9%) and True Negatives are 2348(78.2%). False Positive are 652 (21.7%) and False Negatives are 453 (15.1%). The VGG16 fine tuned model has been able to reduce false negative cases.

6 Conclusion

The present study compared the performance of TL and FT of different pre-trained ConvNet models on the ImageNet dataset such as: VGG, DenseNet, Inception, Resnet, Resnext and Xception. In our previous work, we have experimented with Nasnet and Mobilenet without data augmentation. This showed that the models had a trend to overfit. In order to overcome this problem and increase the performance of a Fine Tuned model, in this work we have used data augmentation as indicated in Section 4.3.2. Our experiments showed that increasing the dataset up to 30 000 images per category helped to achieve good results. Special care was taken to increase the dataset by using the Augmentor library [49] which permits to rotate images avoiding excessive distortion. Compared to our previous work, the image pre-processing has also been changed: instead of using histogram equalization, CLAHE was selected to enhance the contrast the images. Also, the image was resized to a bigger value in order to crop around the center. Finally, image filtering was applied to reduce the presence of noise in the image, as indicated in Section 4.3.1.

In order to estimate overfitting, we proposed a simple ratio relation described in (7). This permitted to conclude that models could achieve good results in classification metrics but overfit in the end. Considering this, we decided to increase the complexity of the model by using the fine tuning technique, where weights from layer 1 to $\gamma - 1$ are frozen, and weights from layer γ until the end are trained with back-propagation. This allowed to increase the performance of both VGG16 and VGG19. Increasing the number of neurons in the FC layer of VGG16-8-FT improved the results but there is a slight overfitting.

The CBIS-DDSM dataset, despite the fact of having at most 1 696 examples for mass abnormality classification, their samples present some artifacts and noise, which is the reason why we used some pre-processing and filtering algorithms in order to improve the image. This is probably due to the fact that the original images are of film type. Certainly, our experiments suggest that TL and FT of pre-trained ConvNets is able to classify film mammogram ROI images. However, the data augmentation increase of the original dataset is considerably. Therefore, public mammogram datasets of Full Digital Mammography, which has better quality than film, with enough sample data would aid to train better classifiers.

With respect to our previous work, we have also included new metrics to evaluate the performance of the classifier. This is important, since the accuracy metric is susceptible to be distorted when the dataset is skewed or unbalance. In other words, it tends to benefit the class with a majority of examples.

7 Future Works

The interest in improving CAD systems in mammography is clear since the disease is a public health problem with high rates of both incidence and mortality. In this and our previous work, we have used the CBIS-DDSM dataset mainly. In a future work, we aim to evaluate the performance of TL and FT in other datasets such as INbreast[55] and Mias[56].

Also, we will be addressing the problem of localization and detection of the mass in the mammogram image. This problem is also of interest and could be formulated from the classification problem here addressed. One of the important things to notice is the peculiarity of the size of the mammogram image compared to the size used in ImageNet trained ConvNets. Mammogram images are of considerable size and it could be of interest to design both the classifier and the object detection avoiding to excessively resize the original image.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment 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 those also provided by Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA). The authors also gratefully acknowledge the financial support provided by the Escuela Politécnica Nacional, for the development of the project PREDU 2016-013.

References

[1] L. G. Falconí, M. Pérez, and W. G. Aguilar. Transfer learning in breast mammogram abnormalities classification with mobilenet and nasnet. In *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 109–114, June 2019. doi: 10.1109/IWSSIP.2019.8787295.

[2] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[3] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

[4] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:170177 EP –, Dec 2017. URL https://doi.org/10.1038/sdata.2017.177. Data Descriptor.

[5] Kiven Erique Lukong. Understanding breast cancer–the long and winding road. *BBA clinical*, 7:64–77, 2017.

[6] World health organization. global health observatory. geneva: World health organization; 2018. who.int/gho/database/en/. 2018.

[7] Jennifer S Drukteinis, Blaise P Mooney, Chris I Flowers, and Robert A Gatenby. Beyond mammography: new frontiers in breast cancer screening. *The American journal of medicine*, 126(6):472–479, 2013.

[8] D. Selvathi and A. Aarthypoornila. Performance analysis of various classifiers on deep learning network for breast cancer detection. volume 2018-Janua, pages 359–363. ISBN 9781509067305. doi: 10.1109/CSPC.2017.8305869.

[9] Afsaneh Jalalian, Syamsiah Mashohor, Rozi Mahmud, Babak Karasfi, Abdul Rahman B Ramli, Communication Systems Engineering, Universiti Putra, Health Science, Universiti Putra, and Qazvin Branch. Review article : FOUNDATION AND METHODOLOGIES IN COMPUTER-AIDED. pages 113–137, 2017.

[10] X. Zhang, Y. Zhang, E.Y. Y Han, N. Jacobs, Q. Han, X. Wang, and J. Liu. Whole mammogram image classification with convolutional neural networks. volume 2017-Janua, pages 700–704, 2017. ISBN 9781509030491. doi: 10.1109/BIBM.2017.8217738.

[11] John Stoitsis, Ioannis Valavanis, Stavroula G Mougiakakou, Spyretta Golemati, Alexandra Nikita, and Konstantina S Nikita. 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, 2006.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, jun 2017. ISSN 15577317. doi: 10.1145/3065386.

[14] L. Hertel, E. Barth, T. Käster, and T. Martinetz. Deep convolutional neural networks as generic feature extractors. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4, July 2015. doi: 10.1109/IJCNN.2015.7280683.

[15] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series, the handbook of brain theory and neural networks, 1998.

[16] Y Bengio. Convolutional Networks for Images, Speech, and Time-Series Unsupervised Learning of Speech Representations View project Parsing View project. Technical report, 1997. URL https://www.researchgate.net/publication/2453996.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. doi: 10.1109/cvprw.2009.5206848.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. Technical report. URL http://image-net.org/challenges/LSVRC/2015/.

[19] Fabien Lauer, Ching Y. Suen, and Gérard Bloch. A trainable feature extractor for handwritten digit recognition. *Pattern Recognition*, 40(6):1816 – 1824, 2007. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2006.10.011. URL http://www.sciencedirect.com/science/article/pii/S0031320306004250.

[20] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, and Jonathon Shlens. Rethinking the Inception Architecture for Computer Vision. Technical report. URL https://arxiv.org/pdf/1512.00567.pdf.

[21] Evgeny A. Smirnov, Denis M. Timoshenko, and Serge N. Andrianov. Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks. *AASRI Procedia*, 6:89–94, 2014. ISSN 22126716. doi: 10.1016/j.aasri.2014.05.013.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Technical report, 2014.

[23] Waseem Rawat and Zenghui Wang. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9):2352–2449, sep 2017. ISSN 0899-7667. doi: 10.1162/neco_a_00990. URL http://www.mitpressjournals.org/doi/abs/10.1162/neco{_}a{_}00990.

[24] Neena Aloysius and M. Geetha. A review on deep convolutional neural networks. In *2017 International Conference on Communication and Signal Processing (ICCSP)*, pages 0588–0592. IEEE, apr 2017. ISBN 978-1-5090-3800-8. doi: 10.1109/ICCSP.2017.8286426. URL http://ieeexplore.ieee.org/document/8286426/.

[25] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

[26] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[27] Patrick Hill and Uma Kanagaratnam. *Python Machine Learning Sebastian Rashka*, volume 58. OUP, 2016.

[28] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *CoRR*, abs/1903.11680, 2019. URL http://arxiv.org/abs/1903.11680.

[29] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *CoRR*, abs/1903.11680, 2019. URL http://arxiv.org/abs/1903.11680.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[31] Barbara Kitchenham and S Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. In *Engineering*, volume 2, page 1051. 2007. ISBN 1595933751. doi: 10.1145/1134285.1134500.

[32] Gustavo Carneiro, Jacinto Nascimento, and Andrew P. Bradley. *Deep Learning Models for Classifying Mammogram Exams Containing Unregistered Multi-View Images and Segmentation Maps of Lesions*. Elsevier Inc., 1 edition, 2017. ISBN 9780128104095. doi: 10.1016/B978-0-12-810408-8.00019-5. URL http://dx.doi.org/10.1016/B978-0-12-810408-8.00019-5.

[33] Stephen Morrell, Zbigniew Wojna, Can Son Khoo, Sebastien Ourselin, and Juan Eugenio Iglesias. Large-scale mammography CAD with deformable conv-nets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11040 LNCS, pages 64–72, 2018. ISBN 9783030009458. doi: 10.1007/978-3-030-00946-5_7. URL http://link.springer.com/10.1007/978-3-030-00946-5{_}7.

[34] M.A. Al-masni, M.A. Al-antari, J.-M. Park, G. Gi, T.-S. T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. T.-Y. Kim. 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, 2018. doi: 10.1016/j.cmpb.2018.01.017.

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[36] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J. Med. Imag*, 3(3):34501, 2016. doi: 10.1117/1.JMI.3.3.034501. URL http://benhuynh.github.io/tl{_}jmi.pdf.

[37] Ana Perre, Luís A. Alexandre, and Luís C. Freire. Lesion classification in mammograms using convolutional neural networks and transfer learning. *Lecture Notes in Computational Vision and Biomechanics*, 27:360–368, jul 2018. ISSN 22129413. doi: 10.1007/978-3-319-68195-5_40. URL https://www.tandfonline.com/doi/full/10.1080/21681163.2018.1498392.

[38] Y. Hu, J. Li, and Z. Jiao. Mammographic mass detection based on saliency with deep features. In *ACM International Conference Proceeding Series*, volume 19-21-Augu, pages 292–297, 2016. ISBN 9781450348508. doi: 10.1145/3007669.3007714.

[39] Karen Simonyan and Andrew Zisserman. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. Technical report, 2015. URL http://www.robots.ox.ac.uk/.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[41] Azam Hamidinekoo, Zobia Suhail, Erika Denton, and Reyer Zwiggelaar. Comparing the performance of various deep networks for binary classification of breast tumours. In *14th International Workshop on Breast Imaging (IWBI 2018)*, page 39, 2018. ISBN 9781510620070. doi: 10.1117/12.2318084.

[42] Oliver Diaz, Robert Marti, Xavier Llado, and Richa Agarwal. Mass detection in mammograms using pre-trained deep learning models. In Elizabeth A. Krupinski, editor, *14th International Workshop on Breast Imaging (IWBI 2018)*, volume 10718, page 12. SPIE, jul 2018. ISBN 9781510620070. doi: 10.1117/12.2317681.

[43] F. Jiang, H. Liu, S. Yu, and Y. Xie. Breast mass lesion classification in mammograms by transfer learning. In *ACM International Conference Proceeding Series*, pages 59–62, 2017. ISBN 9781450348270. doi: 10.1145/3035012.3035022.

[44] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Deep Convolutional Neural Networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, 157:19–30, apr 2018. ISSN 18727565. doi: 10.1016/j.cmpb.2018.01.011. URL https://linkinghub.elsevier.com/retrieve/pii/S0169260717301451.

[45] Shuyue Guan and Murray Loew. Breast Cancer Detection Using Transfer Learning in Convolutional Neural Networks. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–8. IEEE, oct 2017. ISBN 978-1-5386-1235-4. doi: 10.1109/AIPR.2017.8457948. URL https://ieeexplore.ieee.org/document/8457948/.

[46] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis, dec 2017. ISSN 13618423. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841517301135.

[47] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press Professional, Inc., 1994.

[48] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-Local Means Denoising. *Image Processing On Line*, 1:208–212, 2011. doi: 10.5201/ipol.2011.bcm_nlm.

[49] Marcus D Bloice, Peter M Roth, and Andreas Holzinger. Biomedical image augmentation using Augmentor. *Bioinformatics*, 04 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz259. URL https://doi.org/10.1093/bioinformatics/btz259.

[50] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

[51] Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2545–2553. JMLR.org, 2017.

[52] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[53] Tom Schaul, Ioannis Antonoglou, and David Silver. Unit tests for stochastic optimization. *arXiv preprint arXiv:1312.6055*, 2013.

[54] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

[55] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[56] P SUCKLING J. The mammographic image analysis society digital mammogram database. *Digital Mammo*, pages 375–386, 1994.