Data Augmentation In Breast Cancer Classification, Using Transfer Learning

Heida, H.G.	Heuvelmans, J.	Pontiggia, F.	Xin Lan	Jiayuan Hu
4271483	5753066	6925480	6690165	6876587

h.g.heida@students.uu.nl j.heuvelmans@students.uu.nl f.pontiggia@students.uu.nl x.lan@students.uu.nl j.hu5@students.uu.nl

***Abstract*—Since the introduction of mammography computer-aided diagnosis (CAD) system, early detection of malign masses in breast cancer screening has been improved. New opportunities are developing due to recent advances in deep-learning and improved data collection. Nevertheless, the amount of available data in the medical domain is still lagging behind due to several practical reasons. In this paper we propose to explore the possibility of improving performances in a situation of little data by employing transfer learning and data augmentation. The choices we make are based on the state of the art literature in the mass lesion classification domain. We analyze the effect of data augmentation in accuracy of the fine-tuned Inception V3 model, pre-trained on ImageNet data. The model is trained and tested on a data set (CBIS-DDSM) of 3566 mammography region of interest (ROI) images, open available from the Image Cancer Archive. Our results (68.3% acc.) are not comparable to the state of the arts due to, amongst other, unavailability of a large dataset and lack of computational power. However, results show the data-augmentation (68.3% acc.) outperforms the model trained without data-augmentation (60.4% acc.).**

***Index Terms*—Transfer learning, Data augmentation, Convolutional Neural Network**

I. Introduction

Breast cancer screening has been found effective to prevent morbidity and mortality related to breast cancer, since it provides a way of early detection of malign masses. Mortality of breast cancer has then decreased since the 1990's in Europe and North America [1]. Breast cancer screening in western countries is performed through mammographies. A mammograph visualizes the internal breast structures by using low-dose x-ray, and it's now regarded as one of the most suitable techniques to detect breast cancer [2].

Screening is not error-free, because, among other causes, experienced radiologists are needed, especially where the quality of images is low due to poor X-ray devices. Radiologists still struggle to interpret the mammographies, around 30% of them are interpreted wrongly [3]. The causes are the complex structures of breast tissues, oversights, poor quality films, and eye fatigue [4]. Since not treating a malign tumor is much worse than doing a biopsy to a healthy person, sensitivity is increased by prescribing biopsies in cases with unclear diagnosis. Unfortunately, biopsies are expensive and invasive for the patient. Therefore also unnecessary biopsies should be prevented, by increasing the specificity of the screening progress.

To make some improvements in this scenario, the development of computer-aided diagnosis (CAD) is of great interest since it could assist radiologists and increase the sensitivity(recall) of mammographies without worsening specificity [5]. CAD systems have been developed since 1980 [6]. Although they were mostly based on hand craft rules, they still led to great results. [7].

Since 2012, deep convolutional neural networks (CNN) have been applied to the image recognition field, and permitted the field to make impressive progress [8]. State of the art CNNs, usually trained on the ImageNet dataset, achieve better performances than is possible by humans [9]. CNNs have then been enforced in many different sub-fields with the hope of improving the performances as well. Then, a lot of research has been spent in improving previous CAD systems using CNN models [10]. When trying to do that, researchers found out that the medical field presents some peculiarities which contrast the use of CNNS.

One of the biggest challenges of deep learning in the medical field is the poverty of available data. This peculiarity is caused by a set of factors: to cite some, privacy requirements of patients and the costs of collecting images and scanning them, since rarely they are already digitized by medical institutions. Another remarkable issue is the way images are produced with different techniques: sometimes they include out of scope objects, or they have very different light levels or contrast, or they are taken from a different viewpoint The rarity of some diseases is also correlated to the risk of having unbalanced sets, where images of healthy patients are much more than images of ill patients. Since the tasks to perform are very specific, these aspects are relevant, because they introduce very strong biases in the model, causing the it to generalize insufficient. In recent years, a lot of efforts have been made to solve this issue, so that the availability of data can't be considered an obstacle to research in this field anymore [16]. The introduction of standards in X-ray image annotation could in the future allow researchers to build easily big and useful datasets [16].

The poverty of available data causes problems when training CNNs because if we train a model with little data, we incur in high chances of overfitting on those data and not to be able to generalize well with new data. Therefor, bigger datasets implies better performances [17]. This statement is no more in discussion in the literature [18]. Consequently to all these reasons, research focus on different techniques to obtain good

results with little data. One of those is transfer learning.

Transfer learning consists in using pre-trained models, which are trained on a big data set and then (partly) fine-tuning the model with a very small different data set. It is demonstrated that even a model which is pre-trained on an unrelated domain can still perform well in the new domain. Researchers showed the power of transfer learning reaching an accuracy of 92.4% in breast mass classification task, by far better than the baseline model, using GoogleNet and AlexNet, which are pre-trained on ImageNet, a large dataset of images not specific for the medical field [14]. Other researchers compared three different well-known models as pre-trained models and the Inception v3 model outperforms and achieves an accuracy of 97.35% on data set DDSM [13]. Transfer learning is an appropriate choice, especially when the available computation power is low.

Another approach to the question is data augmentation, which seems even more important for the aforementioned problem of biased images and small datasets. Data augmentation consists of extracting more information from the existent dataset, so that the overall size of the set increases. There are different kinds of data augmentation: data warping transforms existing images, while oversampling creates synthetic instances. Both the two ways have shown promising results in the image recognition field, such that they have been imported in the medical field as well. In a few years many different techniques have been elaborated to do augmentations. Some examples of data augmentation are flipping, cropping or rotating images (which are all kinds of data warping), while oversampling is usually performed after having extracted some relevant features. This implies that oversampling augmentation is particularly explored in the medical field, where class unbalance is such a big issue and there is need for balancing the classes with artificially created images [19].

We propose to explore the possibility of improving performances in a situation of little data and small computational power employing transfer learning and data augmentation, based on the state of the art of CNN in the breast cancer classification task, trying to determine their capabilities to solve the two problems.

The remainder of this paper is structured as follows: section 2 describes literature, relevant to our research question. Section 3 describes the set up and all our choices in the experiment: in detail, the dataset, the preprocessing of data and the generalization approach we make use of. Section 4 shows the results we obtained, section 5 presents an overview of the current open questions and problems, while Section 6 gives some summarized conclusions.

II. Related work

In recent years a lot of research has been done in the field of breast cancer screening and classification of the mass lesions found. In these two review papers progress have been summarized [10] [11]. The papers, most relevant to our approach, we will briefly discuss.

The following papers, [12] [14] [13], used an extended version of the DDSM-dataset, which we also use(introduced in section 3). In, [12], they proposed a multi-task transfer learning deep convolutional neural network in combination with data-augmentation. They report AUC(Area under the curve) of 0.78. T In [13] transfer learning in the mammography mass lesions classification domain has been explored. They compared the performances of state of the art CNN models trained on ImageNET: VGG16, ResNet50 and Inception v3. They report 97.35% acc. and 0.98A AUC for Inception v3, using data-augmentation. Moreover, they experimentally determine the best fine-tuning strategy to adopt when training a CNN model.So this paper is at amongst other the foundation of this report.

In [14] they report, in contrast with [12] and [13], both the effect of transfer learning and data-augmentation. They compare transfer learning to their own built baseline system, which shows an improvement. They also show the accuracy curves when training on the non-augmented vs. augmented dataset. The curves show that data-augmentation successfully reduces overfitting.

On of the biggest studies performed in the breast cancer screening domain is [15]. They had access to 44.090 images collected in region mid-west Netherlands. For that reason, among others, they were able to build a system, based on CNN deep learning, which is able to perform at the level of a radiologist. Despite their, relatively, large data set, they have also used data augmentation to improve generalization to unseen data.

III. Approach

A. Dataset

To train and test the models described in this paper, the CBIS-DDSM (Curated Breast Imaging Subset of DDSM) has been used. It has been made available by the cancer imaging archive and is an updated version of the DDSM (Digital Database for Screening Mammography), built in 1997.

The CBIS-DDSM collection includes a subset of the DDSM data selected and curated by a trained mammographer. The database contains scanned film mammography studies of normal, benign and cancer cases. Benign and cancer cases could contain one or more abnormalities: in details, the detected abnormalities are calcifications and masses. The providers improved ROI annotations for the abnormalities, to allow researchers to use it both for classification and segmentation tasks, i.e. respectively to classify the abnormalities or to spot the area where the abnormalities are.

We use the cropped ROI images and masks, since we consider only the classification task, of both calcification and mass cases. The data were labelled at the pathology stage, which means that after a surgery or at least a biopsy. This implies that we are certain that the labels are correct. The three pathology assessments of the original data were benign, benign without callback and malignant. We merged the first two ones and obtained two classes: benign (class 0) and malign (class1).

The dataset is split into training and test set to standardize the cases used for training and testing in research. The test set is around 20% of the whole CBIS-DDSM. The split has been carried out based on BI-RADS assessment to allow for correct evaluation of the models, since different BI-RADS levels are correlated with different difficulty to predict the right class.

The dataset consists of 16 bits grayscale decompressed images in DICOM format, formatted similarly to modern computer vision data sets.

Table I provides a description of our data. We have in total 3566 images. 2862 are used for training: 80% of those (2289) compose the training set, 20% of those composing the validation set. The remaining 704 images are used for testing.

TABLE I
DATASET DISTRIBUTION

	Benign (59.19%)	**Malign** (40.80%)	**Total**
Train (80.26%)	1683 (58.81%)	1179 (41.19%)	2862
Test (19.74%)	428 (60.79%)	276 (39.20%)	704
Total	2111	1455	3566

B. Pre-processing

Due to different sizes of tissues, the ROI images, which have different sizes that vary from around 300×300 to around 800×800, have to be re-scaled into standard pixel arrays, namely 299×299, that we can use as input to the Inception V3 model [20].

There are different methods to re-scale an image, in the most popular computer vision library OpenCV, the following methods are available: nearest-neighbor, bi-linear, re-sampling, bi-cubic, and Lanczos interpolation [21]. In our research, we use re-sampling interpolation, called $INTER\_AREA$, which uses pixel area relation to interpolation, and also known as a preferred method for image decimation for its moiré-free results [21]. For most images of this dataset, they have to be shrunk, not enlarged, to fit the standard size.

The re-sampling algorithm checks whether the original size, both height and width, of n image is an integer multiplied by the target size [22]. Then it takes corresponding ratios as increment for each index to loop over the whole image. For instance, in Figure 1 , it uses the mean of pixel values in the yellow block as the first pixel value of new image, then uses green block to calculate next pixel value in new image. Meanwhile, it is also clear when encountering the non-integer case. How much of an original pixel value will be transferred into a new one, will depend on its weight, namely how much space it occupies in each block, as is shown in Figure 2.

The pixel value of the images is normalized in between 0 and -2, this works best for the Inception V3 model. The distribution of the train and test pixel values is shown in Appendix A. The data is skewed towards 0, which results in slower convergence time, but since our convergence is happens quickly, we decided to leave this as is.

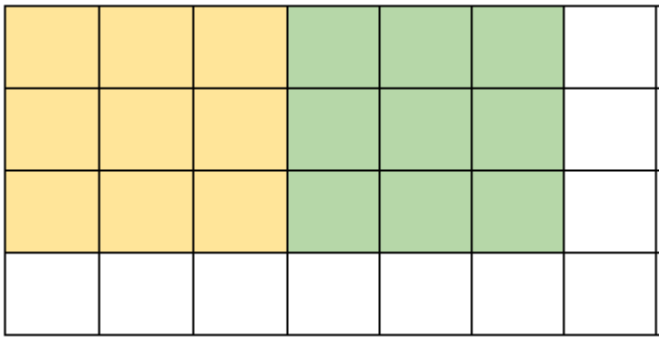

Fig. 1. Ratio of integer

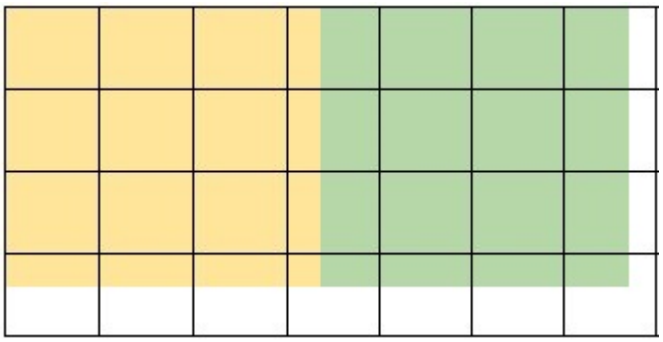

Fig. 2. Ratio of non-integer

C. Architecture

We chose Inception v3 as our pre-trained model and use a softmax 2 output layer on top as a classifier, we then fine-tune the last inception module C on our own data set. When choosing the pre-trained model to use in our network, performance and computational cost are two things that we need to consider. The Inception model uses the so called inception blocks which are sets of multiple different layers. It uses these blocks instead of mapping, and as a result for the same function, it requires less parameters [16]. Therefore, the Inception model requires much less computational cost than VGGNet or other models that perform similar to this model [20]. Besides, researchers compared Inception v3 with two other well-known models namely: VGG16 and ResNet50 and they found that, when using pre-trained weights and fine-tune strategies, Inception v3 achieves the highest accuracy with 97.35% [13].

D. Generalization approach

1) Transfer learning: Literature suggests that the first convolutional layers can extract some general features of the images, which can prevent the model from overfitting at an early stage [23]. Then, fine-tuning the last blocks of the model enables us to learn the specific features of our medical data set.

According to literature [20], fine-tuning the last two blocks (the whole module C) achieves the best performance. Since there is no extensive research about the number of layers to re

train, and many papers don't even state this number, we tried to follow the literature and train both module C's. This however, resulted in a model that was quick to overfit. Moreover, since we have a very small dataset, we thought that the performances could improve with reducing the number of layers to retrain, to limit as much as possible the risk of overfitting, so we focused on trying to fine tune fewer layers than the related work, a good starting point was only training one of the two module C's.

2) Data augmentation: When it comes to medical images augmentation, it is important that we apply the right data transformations [19], this suggests that we have to be careful to preserve the overall features of an image and preserve label recognition post-transformation. By choosing too high values for data augmentations it is possible that the image will be unrecognizable for experts, this will be avoided. According to literature [24], aside from the shear and rotate augmentation, a higher mutual information between original images and augmented images correlates to a higher classification accuracy, in the medical field. Mutual information is a metric to show the similarity between two images [25], so expresses in a way the chances of preserving the label. Operations like flips and shifts result in higher mutual information than operations like adding noise and powers. Although shear and rotation augmentations result in relatively low mutual information, they still perform well in classification tasks. CNNs are rotational invariant, so it's a good choice to perform this kind of augmentation to the dataset. This is because rotations and shears still retain many of the image information despite the low mutual information, and help the model to form abstractions. On the other hand, from the augmented images, we can also see that rotation and shear provides with mass images from different views shapes respectively and according to [36], a vertical flip of a mass would still result in a realistic mass.

Based on the knowledge above and combine the most common methods used in other papers [13] [26] [27], we choose to perform the four operations shown in the table below. We decide to get rid of the zooming operation, although it is used in one paper we based on, after experimenting with it. The reason is that we are already using the ROI (region of interest) which can be seen as an already zoomed image with respect to the original image. Therefore, by zooming in the image again, there is a risk of missing the mass part or losing part of the mass. As for the parameters, there are few papers which clearly state what parameter values they set for each operation. To explore different possibilities, we perform three experiments with different parameters, as shown in the Table II. First, we set the parameter for rotation and shift as a relatively low number since if relevant information appears at the corner of image, we may lose a lot of important information after the augmentation. Then we change the parameters using a multiplying factor of 2 and 0.5 for all the numerical ones.

TABLE II
PARAMETERS FOR DATA AUGMENTATION

Settings	**Standard**	$\times 2$	$\times 0.5$
Rotation	20	40	10
Shift (height & width)	0.1	0.2	0.05
Shear	0.35	0.7	0.175
Fill mode	Wrap	Wrap	Wrap
Filp (horizontal & vertical)	True	True	True

E. Other relevant choices

In fine-tuning we use the adaptive ADAM optimizer [28]. At first, we tried to replicate the learning rates [13], which means setting it primarily to 1^{-4} and dividing it by 10 each time the validation error stops improving. This is a training strategy widely used in research recently [29] [30]. However, during the experiment, this strategy did not make a difference in the accuracies we achieved and thus we decided to explore different values. We opted for a constant learning rate of 1^{-4}, since this learning rate gave us the best resulting accuracy.

In order to reduce overfitting of the model, we added a dropout layer. We explored different possible values of the dropout rate, although the literature agrees on a dropout of 0.5.

Furthermore, we added a dense layer on top of the model. This choice is shared by all the literature. In the dense layer, we chose softmax function in order to get the output probabilities and set the dimensionality of the output space to 2 since we only predict two classes. The softmax function takes a vector from the previous layer and normalizes it into class probability as the output which adds up to 1.

The choices made about learning rate and dropout value are discussed in the results part, since they are based on our own evidence.

IV. RESULTS

The procedure to conduct the experiments is the following: we run a models for 30 epochs, and choose the weights of the epochs with the best validation accuracy. Then we test it on the test set.

Firstly, we looked at the best fine-tuning strategy. Figure 12 and 11 report respectively the results in terms of accuracy and loss function of fine tuning the last two blocks (as in the related work) at different epochs. Fine tuning the last two block means the whole inception V3 module C. We used the data augmentation strategy chosen in the approach part with a multiplying factor of 1. Table III reports the confusion matrix of the predictions of this model on the test set. We found that

TABLE III
CONFUSION MATRIX WHEN FINE TUNING THE LAST TWO BLOCKS AND DATA AUGMENTATION

test_acc = 0.652		Predicted	
		Negative	Positive
Actual	Negative	311	117
	Positive	128	148

retraining only half of the Inception v3 module C increases the performances in terms of loss function and accuracy, with respect to the paper based model. As expected by us before

TABLE IV
CONFUSION MATRIX OF FINAL MODEL WITH DATA AUGMENTATION

With Augmentation test_acc = 0.683		Predicted	
		Negative	Positive
Actual	Negative	335	93
	Positive	130	146

TABLE V
FINAL MODEL: CONFUSION MATRIX OF FINAL MODEL WITHOUT DATA AUGMENTATION

Without Augmentation test_acc = 0.604		Predicted	
		Negative	Positive
Actual	Negative	285	143
	Positive	136	140

the experiment. Moreover, this choice is also the one which gave the best performance in our experiments. It consists of deactivating 279 layers instead of 248. Figure 8 and 4 report respectively the accuracy and the loss function of this model. Table IV reports the confusion matrix we obtained on the test set. The overall accuracy increase from 0.652 to 0.683. Hereunder we will refer to this model as the standard model.

Together with the analysis of the best number of layers to deactivate, we looked at the dropout value and the learning rate. The quantitative results of the analysis of the learning rate and the dropout values is reported respectively in Figure 3 and Figure 10. The two graphs show the changes in the values of the loss function with the passing of epochs, for different values of learning rate and dropout, respectively. These values are obtained on the standard model.

We found no significant effect of dividing the learning rate when the validation error stops improving on our data set. So we set our learning rate constantly to $10^{-}4$ as shown in Figure 3. The accuracy plots are in Appendix B

Regarding the dropout value, we excluded the really volatile ones, and chose 0.2 as shown in Appendix C.

Then, we evaluated whether the results of the standard model could be improved by changing the multiplying factor

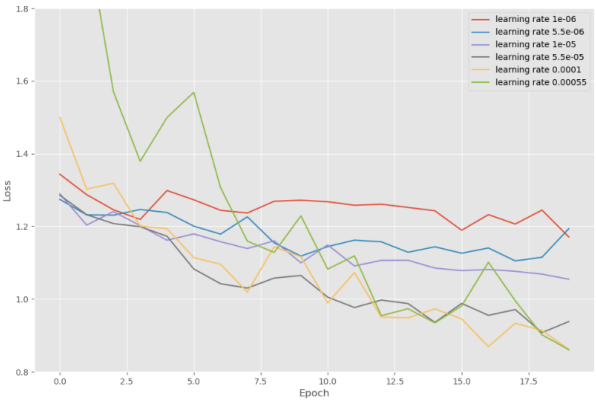

Fig. 3. Effect of the learning rate

of the data augmentation parameters. The accuracy we got on the test set with different multiplying factors is reported in the Table V. From the table, we can see that the standard setting of parameters achieved the highest accuracy. Therefore, the standard model is still the best one.

Finally, we compared the performances between the standard model with and without data augmentation.

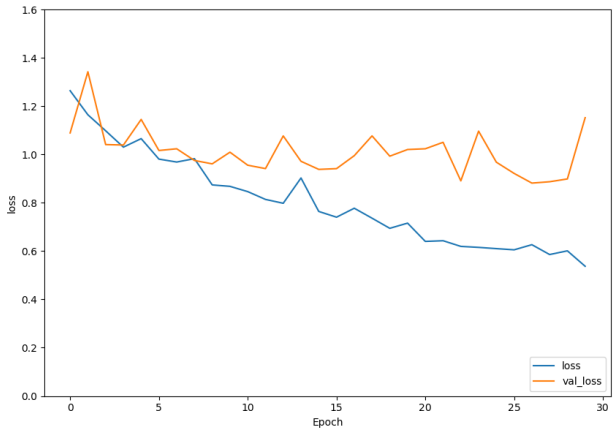

Fig. 4. Loss with data augmentation of the standard model

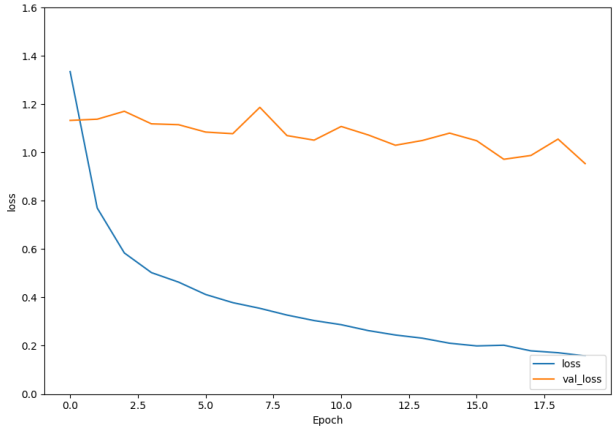

Fig. 5. Loss without data augmentation

The results show a clear difference in the performance of the model with and without data augmentation. The non-data augmented dataset shows that the model is too complex and almost immediately overfits. This is shown by the gap between the accuracy on the training set, which reaches 99%, and the accuracy on the validation set, which remains stable. Employing data augmentation fixes this problem to an extent, since the model achieves an overall higher validation accuracy and the accuracy on the training set starts increasing more

then the validation accuracy only after 10 epochs. Still, our best results do not reach an accuracy of 0.70, and we observe this undesider divergence between the training accuracy and the validation accuracy only a few epochs further. The accuracy with and without data augmentation on training set and validation set is shown in Figure 8 and Figure 9 respectively.

This overfitting trend is also clearly visible when tracking the loss. We find that without data augmentation the model starts to overfit almost instantly, while with data augmentation we can conclude that the validation loss line follows the training loss line more closely and this is a clear sign of less overfitting. Still, after 7-8 epochs the loss also begins to diverge. It seems like our dataset is too small to learn enough features before the model overfits. The loss with and without data augmentation on training set and validation set is shown in Figure 4 and Figure 5 respectively.

There is an increase in the total amount of correct predictions. This increase is from an accuracy of about 60.4% without data augmentation to 68.3% with data augmentation.

V. Conclusion

A. Discussion of approach

To implement our models we made some choices taking into consideration different possibilities which are, to date, subjects of debate.

The ImageDataGenerator library used to create augmented images online is automatically reshapes the images to 8 bits. Our data is 16 bit and thus quite a bit of information is lost in this process. In the future it is advisable modify this library to allow 16 bit images to be generated, or alternatively load the pre-processed images in memory and use offline data augmentation.

Transfer learning with fine tuning is becoming the common way of applying transfer learning in the medical field. This is due to the incredible results obtained by [31] and [32]. However, all the research which aims at pointing out whether this is the best approach does not reach an agreement [16]. The alternative is to use the pre-trained network just as feature extractor. In our task, the state of the art results are obtained by [13] and [29]. Among those two, the former applies fine tuning, while the latter applies the feature extraction strategy. We decided to apply fine tuning because we wanted to explore the deep learning models, instead of traditional machine learning methods.

The number of layers to retrain varies in the literature, where all possible options are tried, with no agreement about the best choice to make. Even retraining the entire model has been shown to succeed [31]. We didn't experiment with all possible values, due to computational and time limitations. In Appendix E a loss plot is shown where both C modules are trained, this model overfits quicker and achieves lower accuracy.

More generally, even the employment of transfer learning has been questioned very recently in [33]. However, we based our choices on [34], where transfer learning outcomes training from scratch with a dataset size very similar to ours (1000 images).

B. Discussion of results

The results, in terms of accuracy, are clearly below the ones found of the papers we are referring to. Moreover, they are not satisfactory in the view of deploying the model in a real scenario. We argue here what reasons we found for that.

The dataset size is much smaller than anyone else in the related works. In [13], instead of the CBIS-DDSM dataset, the original DDSM is used and then merged with two other datasets with a resulting dataset of 6116 images. Unfortunately, the original dataset is not available to us since the code libraries to manipulate the images are not supported anymore (they were written in the '90) and we weren't able to download the other two datasets. In [29] the same amount of data is used (even fewer, 900 images overall) from the same subset of DDSM. However, they use the CNN just to extract features and then train two SVMs with the extracted features, which is not our focus in the research since we want to explore with deep learning methods instead of traditional machine learning methods. All the other related works use more than 10000 images.

We didn't perform any kind of cross validation, since it was not feasible with augmenting data online with Keras ImageDataGenerator, which would have augmented validation images as well. Augmenting the validation set and not the test set gives bad estimates, indeed. However, generating images offline requires much more storage space, which is beyond our possibilities, and there is no evidence about how much cross validation could affect the results. [13] makes the same choice as us and uses Keras ImageDataGenerator, anyway.

We have found that a constant learning rate does not affect the performances with respect to reducing it. However, we are aware that in the literature there has been found a reason for reducing the learning rate: using a constant learning rate for all the layers may destroy the features that are learned previously. In this case we based our choice on the empirical evidence of results.

The most relevant related works we referred to in our research did not provide any code or clear settings, making it very hard to replicate their experiments. We suspect that some experiments take advantage of more powerful strategies which haven't been reported in the papers. However, since our goal was to explore data augmentation and transfer learning themselves and not other possibilities, this affects only the results we got and not the scope of our research. Moreoever, performance metrics reported in different papers differs, which make them difficult to compare. They use alternately accuracy [13] [29] [31] , AUC [13] [29] [31] [8] [12] and one even FROC [8], without referring to a common benchmark to compare results. Segmentation tasks uses even different metrics, such as DICE scores and Hausdorff measurements.

C. Conclusion

This report confirms the utility of data augmentation in the context of ROI mass lesion classification in breast cancer screening data. In the extreme situation we were dealing with, data augmentation showed its effect and the plots report the

same trend as in [14]. The training accuracy line follows the validation accuracy line longer: this means that model starts overfitting later. Still, it shows that this method is not a "cure all" method it does not improve the performance enough to allow to deploy the model in a real context and use it, even with the ad hoc solutions we found for the parameter settings.

REFERENCES

[1] Tabár, László, Stephen W. Duffy, Bedrich Vitak, Hsiu-Hsi Chen, and Teresa C. Prevost, "The natural history of breast carcinoma: what have we learned from screening?." Cancer 86, no. 3 (1999): 449-462.

[2] Berlin, Leonard."Radiologic errors, past, present and future." Diagnosis 1, no. 1 (2014): 79-84.

[3] American Cancer Society. Global Cancer Facts & Figures 4th Edition. Atlanta: American Cancer Society; 2018.

[4] Huynh, Phan T., Amanda M. Jarolimek, and Susanne Daye. "The false-negative mammogram." Radiographics 18, no. 5 (1998): 1137-1154.

[5] Doi, Kunio, Maryellen L. Giger, Robert M. Nishikawa, and Robert A. Schmidt. "Computer aided diagnosis of breast cancer on mammograms." Breast Cancer 4, no. 4 (1997): 228-233.

[6] Giger, Maryellen L., Nico Karssemeijer, and Samuel G. Armato. "Computer-aided diagnosis in medical imaging." (2001).

[7] Morton, Marilyn J., Dana H. Whaley, Kathleen R. Brandt, and Kimberly K. Amrami. "Screening mammograms: interpretation with computer-aided detection—prospective evaluation." Radiology 239, no. 2 (2006): 375-383.

[8] Ribli, Dezső, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. "Detecting and classifying lesions in mammograms with deep learning." Scientific reports 8, no. 1 (2018): 1-7.

[9] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In Proceedings of the IEEE international conference on computer vision, pp. 1026-1034. 2015.

[10] Gardezi, Syed Jamal Safdar, Ahmed Elazab, Baiying Lei, and Tianfu Wang. "Breast cancer detection and diagnosis using mammographic data: Systematic review." Journal of medical Internet research 21, no. 7 (2019): e14464.

[11] Houssami, N., Kirkpatrick-Jones, G., Noguchi, N., & Lee, C. I. "Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice." Expert review of medical devices, 16(5), (2019): 351-362.

[12] Samala, Ravi K., Heang-Ping Chan, Lubomir M. Hadjiiski, Mark A. Helvie, Kenny H. Cha, and Caleb D. Richter. "Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms." Physics in Medicine & Biology 62, no. 23 (2017): 8894.

[13] Chougrad, Hiba, Hamid Zouaki, and Omar Alheyane. "Deep convolutional neural networks for breast cancer screening." Computer methods and programs in biomedicine 157 (2018): 19-30.

[14] Lévy, Daniel, and Arzav Jain. "Breast mass classification from mammograms using deep convolutional neural networks." arXiv preprint arXiv:1612.00542 (2016).

[15] Kooi, Thijs, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. "Large scale deep learning for computer aided detection of mammographic lesions." Medical image analysis 35 (2017): 303-312.

[16] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. "A survey on deep learning in medical image analysis." Medical image analysis 42 (2017): 60-88.

[17] Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." IEEE Intelligent Systems 24, no. 2 (2009): 8-12.

[18] Sun, Chen, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. "Revisiting unreasonable effectiveness of data in deep learning era." In Proceedings of the IEEE international conference on computer vision, pp. 843-852. 2017.

[19] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of Big Data 6, no. 1 (2019): 60.

[20] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.

[21] Bradski, Gary, and Adrian Kaehler. Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc.", 2008.

[22] Laganière, Robert. OpenCV Computer Vision Application Programming Cookbook Second Edition. Packt Publishing Ltd, 2014.

[23] Jiang, Fan, Hui Liu, Shaode Yu, and Yaoqin Xie. "Breast mass lesion classification in mammograms by transfer learning." In Proceedings of the 5th international conference on bioinformatics and computational biology, pp. 59-62. 2017.

[24] Hussain, Zeshan, Francisco Gimenez, Darvin Yi, and Daniel Rubin. "Differential data augmentation techniques for medical imaging classification tasks." In AMIA Annual Symposium Proceedings, vol. 2017, p. 979. American Medical Informatics Association, 2017.

[25] Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).

[26] Dhungel, Neeraj, Gustavo Carneiro, and Andrew P. Bradley. "A deep learning approach for the analysis of masses in mammograms with minimal user intervention." Medical image analysis 37 (2017): 114-128.

[27] Jung, Hwejin, Bumsoo Kim, Inyeop Lee, Minhwan Yoo, Junhyun Lee, Sooyoun Ham, Okhee Woo, and Jaewoo Kang. "Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network." PloS one 13, no. 9 (2018).

[28] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[29] Jiao, Zhicheng, Xinbo Gao, Ying Wang, and Jie Li. "A deep feature based framework for breast masses classification." Neurocomputing 197 (2016): 221-231.

[30] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105. 2012.

[31] Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." Nature 542, no. 7639 (2017): 115-118.

[32] Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." Jama 316, no. 22 (2016): 2402-2410.

[33] Raghu, Maithra, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. "Transfusion: Understanding transfer learning for medical imaging." In Advances in Neural Information Processing Systems, pp. 3342-3352. 2019.

[34] Menegola, Afonso, Michel Fornaciali, Ramon Pires, Sandra Avila, and Eduardo Valle. "Towards automated melanoma screening: Exploring transfer learning schemes." arXiv preprint arXiv:1609.01228 (2016).

APPENDIX A. DISTRIBUTION OF PIXEL VALUES

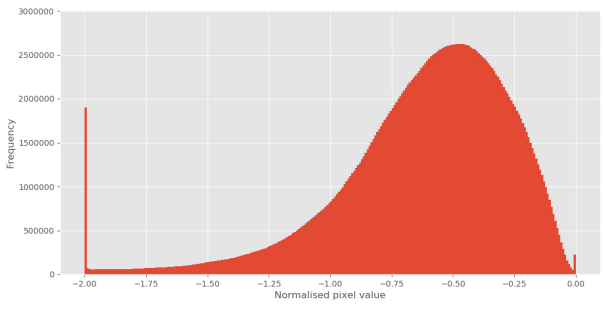

Fig. 6. Pixel value distribution train set

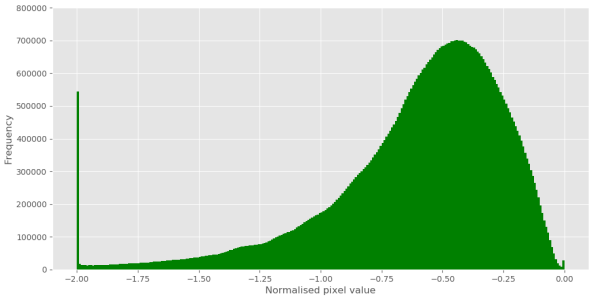

Fig. 7. Pixel value distribution test set

Appendix B. Accuracy Plots with and without data augmentation

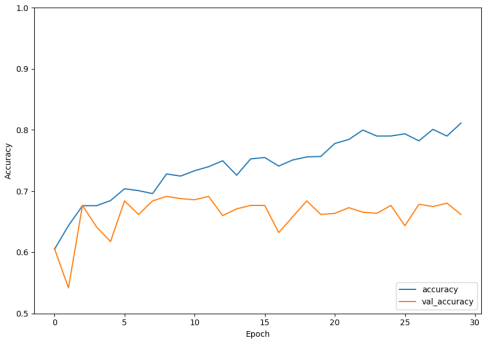

Fig. 8. Accuracy with data augmentation of the standard model

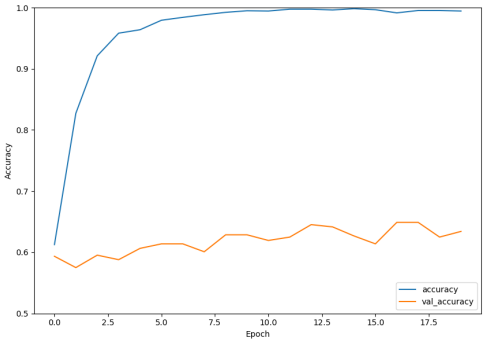

Fig. 9. Accuracy without data augmentation

Appendix C. Loss plot on dropout rate

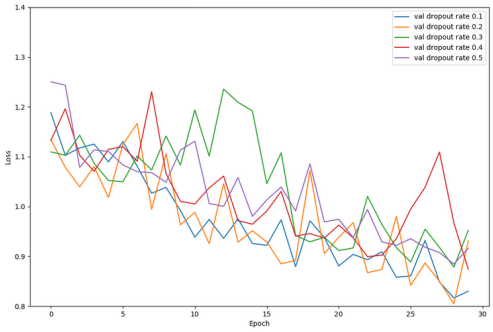

Fig. 10. Effect of the dropout rate

Appendix D. Loss plot both C modules unfrozen

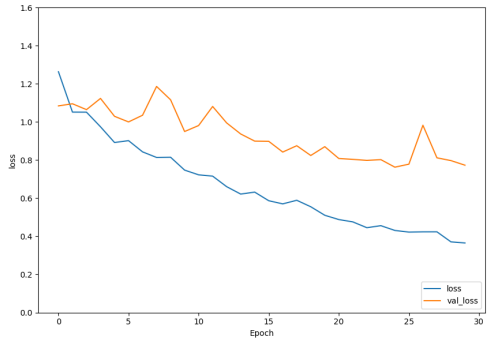

Fig. 11. Loss data augmentation and both C modules unfrozen

Appendix E. Accuracy plot both C modules unfrozen

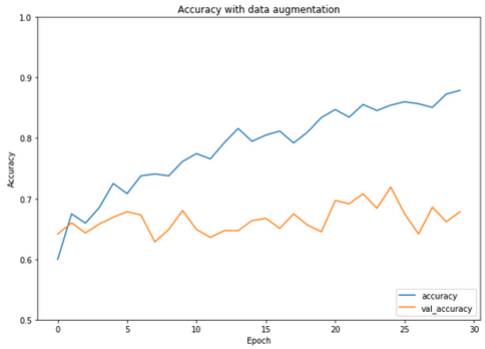

Fig. 12. Accuracy data augmentation and both C modules unfrozen