Implementation of convolutional neural networks to classify mammograms from a breast cancer cohort.

Cesar Sanchez-Villalobos
Department of Electrical and Computer Engineering
Texas Tech University
Lubbock, TX
cesarasa@ttu.edu

Fernando Koiti Tsurukawa
Department of Electrical and Computer Engineering
Texas Tech University
Lubbock, TX
fetsuruk@ttu.edu

***Abstract*—According to both NIH and the RSNA, Breast Cancer (BC), is the most common type of cancer diagnosed to women in the United States. One of the methods currently used in radiology to diagnose BC, is the mammography screening, where an mammogram is taken from the patient and then a trained physician needs to look into patterns and check if the patient could be diseased. In order to reduce the costs of the mammography screening, it is of high interest use Deep Learning (DL) algorithms to aid the decision making. The following document serves as a report for the final assignment of the Data Science class, where we analyzed a dataset and implemented Convolutional Neural Networks (CNNs) to classify the mammograms into either Control, or diseased.**

***Index Terms*—Breast Cancer, Deep Learning, Convolutional Neural Networks, Data Analysis, Image Processing**

I. Introduction

The following document is a report for our final project during the Data Science class, delivered by Dr. Mary Baker at Texas Tech University. During the project, we joined a kaggle competition named *RSNA Screening Mammography Breast Cancer Detection*. During this competition, the Radiological Society of North America (RSNA), provided a dataset of both control and BC patients with their mammograms to promote the research of DL techniques in their field.

Therefore, we did a brief data analysis, the implementation of data balancing techniques and several CNNs to perform the classification task. Therefore, the following document is divided as follows: first, we have a discussion about our data analysis and the data provided by RSNA. Second, we discuss the problems of data leakage and data imbalance, which are the most common problems in DL and the reproducibility of results. Third, we discuss the methods used for solving these problems and to train the neural networks. In chapter IV, we present our results and finally, we give a brief discussion in the conclusion.

II. RSNA Breast Cancer Dataset

The RSNA dataset is an anonymized dataset publicly released by RSNA as a designed experiment to aid the identification of cancer cases using mammogram screening. This dataset has both the metadata and the mammograms for each patient, in the metadata, we have access to the following information:

- Source hospital: it is an ID number to know where the images were taken.
- ID of the patient: each patient has an assigned number.
- ID of the image: as each patient has several images, it is necessary to know which image are we looking.
- Laterality: each patient has at least two images of the right breast, and two images of the left breast.
- Mammography view: all the patients have a least two views of each breast. In a usual screening, we take two classic types of images, the mediolateral oblique (MLO) view, and the cranial caudal view.
- Implant: we need to know if there are artifacts inside the images, this might be an issue for any implementation.
- Density: a phenotypic trait of the breast.
- Biopsy: some screenings lead to biopsy. In this column we have which one went for a biopsy.
- Invasive: a phenotypic trait of some of the tumors.
- BIRADS: a rating of how likely is for the patient have cancer. There are several NaNs in this column.
- Age: the age of the patient at the moment of the screening.
- Cancer: our target, this is a binary column where the ones are positive values, and the zeros are healthy controls.

A. Types of view in a mammogram screening:

In both the challenge overview and the literature review, we can see that in most mammogram screenings the physician takes two views of each breast, the mediolateral oblique (MLO) and the cranial caudal (CC) views. According to Mohamed et al [4], the CC view is taken from above the breast and the MLO view is taken from one side of the breast, aligning the center of the chest and imaging outwards. These two views are not the only available views in the whole dataset, but they are the views that we are interested in, as most of the patients only have these two images.

In Figure 1, we can see the most common types of views according to Mohamed et al [4], using images from the RSNA dataset. Note that for this patient, we will have an extremely dark image, as the volume of the breast is not spread over all the taken image. Note also that there are annotations with the view on the corners of the images.

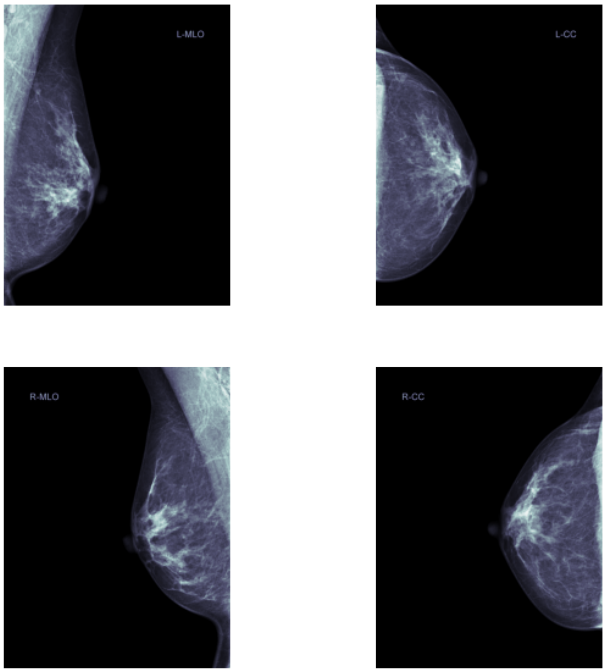

Fig. 1. Example of the views for one patient. In the left column we can see the MLO view, and in the right column we can see the CC view.

B. Exploratory Data Analysis

The RSNA data has a total of 11913 different patients, where 11427 are healthy control subjects and 486 are Breast Cancer subjects. Note that in this case we have a heavily imbalanced problem, as it is usually the case for biomedical imaging problems. The data has records of six different types of views including MLO and CC, and each patient indeed has at least 4 images. A distribution of patients per number of images, is shown in Figure 2.

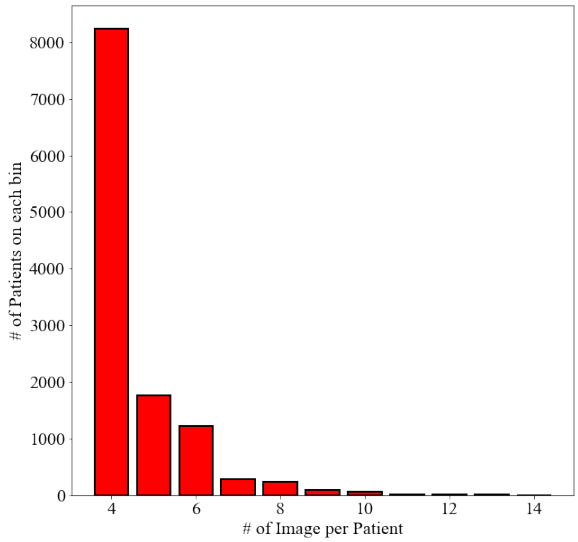

Fig. 2. Distribution of patients per number of Images.

Note that over 8000 of the patients have only 4 images (the CC and MLO views for their two breasts), and only a small number of patients have more than 6 images. All these images sum to a total of 54706 images, where 53548 are labeled as Non-Cancer images and 1158 as showing cancer. It is important to note that from these 53548 images, we also have 1477 that are showing an implant. These implant may be targeted as problematic when we are training, so it is important to highlight that they could be an issue in any screening. Finally, a distribution of the ages of the patients is shown in Figure 3.

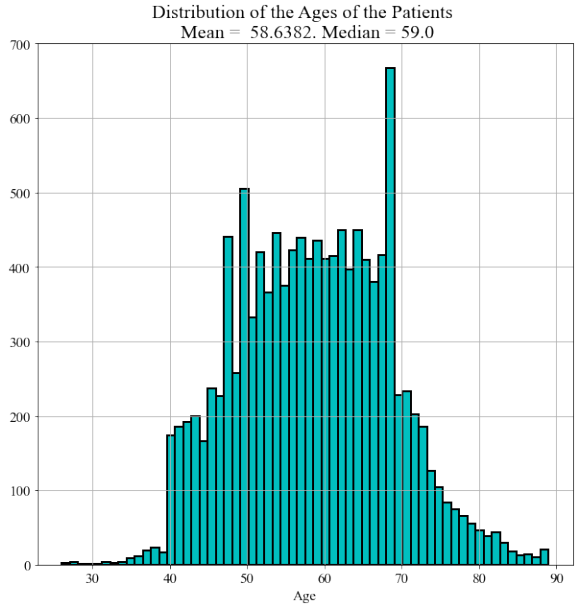

Fig. 3. Age Distribution.

Note that we have low representations of ages before 40 years. This is mainly due to the idea that a risk factor of breast cancer in women is being over 40 years old. Then, the distribution shows a heavy tail on the left and a normal tail on the right. This suggests that the data is slightly skewed, the computed skewnes was 0.103, the Kurtosis was -0.354, the youngest patient was 26 years old, whereas the oldest was 89. The mean and the median are also visible in Figure 3.

III. Problems to Solve

While we studied the RSNA dataset with our exploratory data analysis (EDA), we detected three critical problems to solve before the DL implementation. These problems were Data Imbalance, Data Leakage, and the hetereogenity of the images. In this section we will define these problems.

A. Data Imbalance

In the context of machine learning and statistics, we refer to *class imbalance* to the event where there are more observations of one class, than the other available classes in the dataset. This is specially detrimental when we train any kind of DL model, as it leads to a poor performance when we try to predict the underrepresented class [1].

Data imbalance is a problem in machine learning as it could lead to models that are biased towards the majority class. This is because the model is more likely to learn from the majority class data, and will therefore be more likely to predict the majority class.

In our case, a 96% of our data is labeled as healthy. This represents a heavily imbalanced classification problem, as setting a classifier that always points into the healthy label, will give a 96% of accuracy but 0 recall. The goal then, would be to improve the 0 recall into a more suitable number.

There are four ways to solve this issue in our case:

- **Undersampling the majority class:** the designer of the experiment can sample without replacement from observations in the majority class. This can be done until the number of samples per class is the same [2].
- **Oversampling the minority class:** the designer of the experiment can sample with replacement from observations in the minority class. This can be done until the number of samples per class is the same. However, a drawback of this method is that with a heavily imbalanced class, we will end up with all the instances repeated, which might be also detrimental to the problem [2].
- **Data augmentation of the minority class:** the designer of the experiment can sample with replacement from observations in the minority class, however, to each sampled instance, a random transformation will be applied to it. The transformation can include one or several types of: geometric transformation, probabilistic transformations, non linear transformation, including additive or multiplicative random noise, among others. This will ensure to have different instances and a suitable number of data points, which is desirable for Deep Learning problems [3].
- **Generative methods:** with images, in the past two years we have seen the increase of generative approaches for images like DALL-E and Stable Diffusion, and more classical methods like the Generative Adversarial Networks (GANs). All of these methods end up creating synthetic images, and synthesizing patterns to create new images from the minority class, would be a potential solution for the Data Imbalance problem [3].

B. Data Leakage

We refer to data leakage as the event where some information used in the training stage of the model, will appear in the evaluation stage. In this case, we have to be very careful when we split our dataset into train and test sets, as if we do it image-wise, we will end up with images from the same patient in both sets. Therefore, to solve this issue, we will be splitting the dataset patient-wise [5].

C. Image related issues

As the images are coming from different hospitals and, sometimes, from different imaging machines, they present different sizes and contrasts. Also, the dataset is large, it has a size of 116 GB of information and all the images are stored in DICOM format.

The DICOM format, is highly used in clinical environments as it allows the physician to store metadata related to the patient and the machine. As we will only need the pixel array, we will just save the arrays into JPEG format to save space.

IV. Methods and Preprocessing

A. Image Processing

During the last 6 years, RSNA has launched eight different AI challenges using medical imaging as inputs, all these challenges have:

- Images with different sizes.
- Images with different contrast.
- Images from different places, that could have annotations created by the physicians on top of the images.

In this subsection we will deal with these problems.

1) Images with different sizes: As the CNNs only allow one predefined size, this is a crucial problem. The first way to solve this is to downsample all the images into a specific size using a bicubic resizing algorithm. This technique was implemented using the OpenCV library. As an important drawback of this method, is that when we downsample the images we will lose resolution, and we will not hold the geometrical shape of the breast. The result of doing this is comparable to the top left picture of 1, and it is shown in Figure .

Fig. 4. Deformed breast after downsampling to 256x256.

2) Images with different contrast.: DICOM images have information about the Look-up table (LUT), that allows the machine to transform the contrast of the image. We did the LUT transformation using the OpenCV package with Python. One crucial problem with these images from RSNA is that some machines were taken on a white background and black tissue, which is different to the images from Figure 1. An example is shown in figure 5.

Fig. 5. MONOCHROME1 labeled image

We identified all these images, and converted them taking their negative by using the equation:

$$I_t(x, y) = 255 - I_s(x, y) \tag{1}$$

Where I_t is the transformed image and I_s is the source image.

3) Images from different places: As the images are from different machines and places, some of them have annotations on top of the image, like the letter "L-MLO" that we can see in Figure 4. To solve this problem, we followed a tutorial on Regions of Interest (ROI) from the wonderful Notebook from Awsaf, hosted in Kaggle. The result is presented in Figure 6.

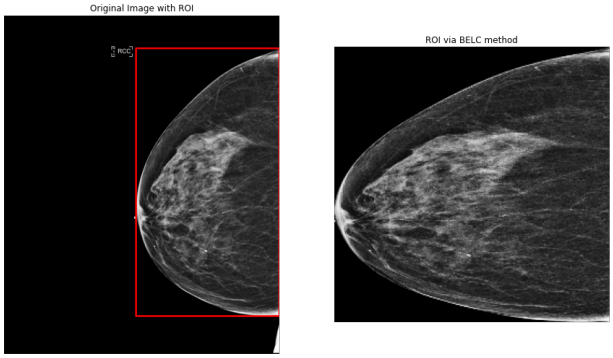

Fig. 6. On the left, the original image with the found ROI. On the right, the cropped and resized to 256x256 image.

The ROI extraction method was done by using the adaptive threshold from OpenCV, then eroding the image with a structural element of a square of ones, then we found the two objects in the image, and took the biggest one (the breast instead of the annotations). From such final image, we downsampled it to 256x256 pixels, as can be seen in Figure 6.

B. Convolutional Neural Network

The objective CNN designed for this problem is presented in Figure 7.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [-1, 8, 256, 256]             224
              ReLU-2          [-1, 8, 256, 256]               0
         MaxPool2d-3          [-1, 8, 128, 128]               0
            Conv2d-4         [-1, 16, 128, 128]           1,168
              ReLU-5         [-1, 16, 128, 128]               0
         MaxPool2d-6           [-1, 16, 64, 64]               0
            Conv2d-7           [-1, 32, 64, 64]           4,640
              ReLU-8           [-1, 32, 64, 64]               0
         MaxPool2d-9           [-1, 32, 32, 32]               0
           Flatten-10                [-1, 32768]               0
           Linear-11                  [-1, 512]      16,777,728
             ReLU-12                  [-1, 512]               0
          Dropout-13                  [-1, 512]               0
           Linear-14                  [-1, 256]         131,328
             ReLU-15                  [-1, 256]               0
           Linear-16                    [-1, 1]             257
          Sigmoid-17                    [-1, 1]               0
================================================================
Total params: 16,915,345
Trainable params: 16,915,345
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.75
Forward/backward pass size (MB): 16.02
Params size (MB): 64.53
Estimated Total Size (MB): 81.29
----------------------------------------------------------------
```

Fig. 7. Convolutional Neural Network

Note that the network has 3 convolutional layers, and 3 fully connected layers, which makes it a simple model for classification tasks, it was trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.0001, a regularization constant equal to 1 and a momentum constant equal to 0.9.

V. Results

A. Baseline Model

In the baseline model, we used the original images and trained the network with those 256x256 images. We used undersampling methods given the computational and time limitations from an academic semester. The result is given in Figure 8 where we can see that we got an test accuracy of 58.62%. Another metric used to measure the performance of the model is given in Figure 9, where we can see that the AUC of the Receiving Operator Curve is 0.63, suggesting an improvement of the 50% of the random flipping. Note that for this baseline, we downsampled the data so the input data was properly balanced.

B. Model with extracted ROIs

Similarly with the images Cropped by their ROI, we obtained the result shown in figure 10, where we can see a slight improvement than what we saw in Figure 9. This improvement

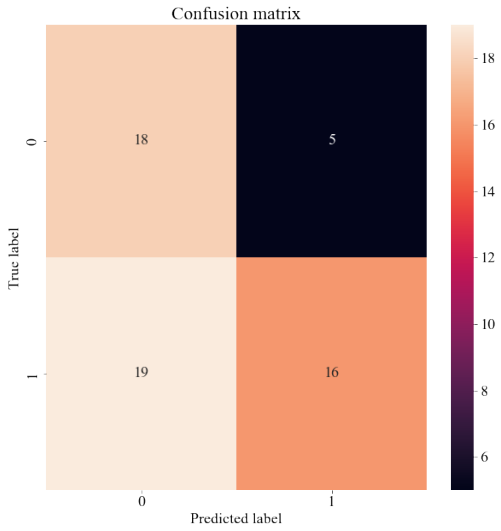

Fig. 8. Confusion matrix of the baseline model inferring the test data.

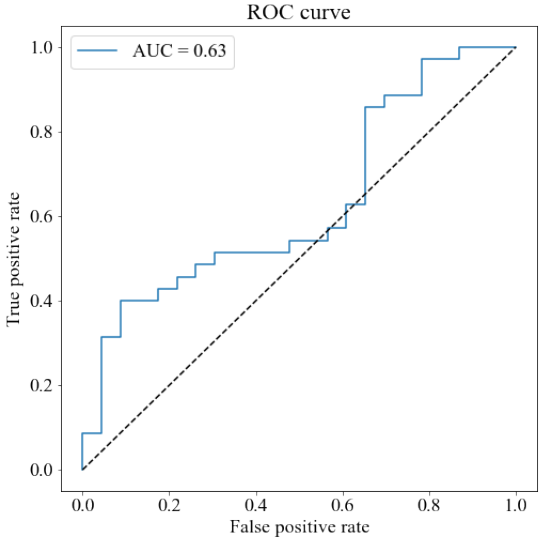

Fig. 9. ROC of the baseline model inferring the test data.

might be seen as insignificant, as the train-test-split was not controlled for the two methods and we might have gotten the result by chance. A better approach might be to seed the split, but this might change between systems and even personal computers, which is very hard to track. The obtained accuracy was of 61.75% on the test set, which is an increment of 3% compared to the previous implementation.

C. Extra models

We tried several combinations and hyperparameter tuning, but we could not find an increased accuracy or AUROC value.

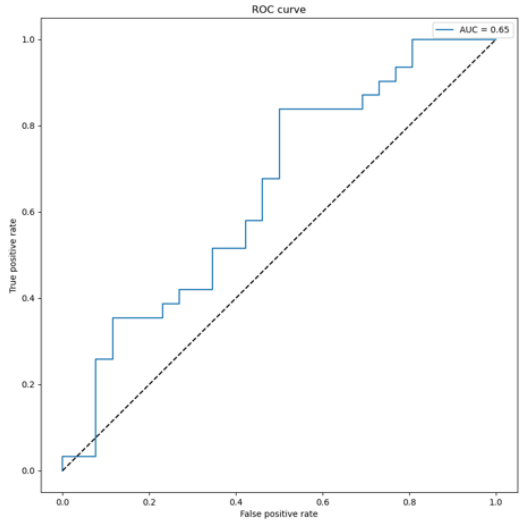

Fig. 10. ROC of the ROI model inferring the test data.

VI. Conclusion and Suggestions

- As one of the most common problems in radiology, the project proved to be challenging for a vanilla neural network. Our best efforts with a reduced dataset, are not performing well in the experiments.
- We were unable to perform the augmentation of the majority class, it was a slow process as the augmentation occurs during the online training of the network.
- The problem of undersampling the data might be an issue [3], as we do not have enough samples to train the network. However, in implementations from Kaggle we saw that even with oversampling, they were unable to perform well.
- A maybe more suitable approach in the future, might be to use a DualCNN to solve this problem, where one branch of the CNN processes the MLO information, and the other processes the CC information. This might give more context to the network and perform better than the vanilla CNN.

References

[1] Alessandro Bria, Claudio Marrocco, and Francesco Tortorella. Addressing class imbalance in deep learning for small lesion detection on medical images. *Computers in Biology and Medicine*, 120:103735, 5 2020.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 6 2002.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[4] Aly A. Mohamed, Yahong Luo, Hong Peng, Rachel C. Jankowitz, and Shandong Wu. Understanding clinical mammographic breast density assessment: a deep learning perspective. *Journal of Digital Imaging*, 31:387–392, 8 2018.

[5] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley

Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 7 2020.