# From Easy to Hard: A Curriculum Learning Approach for Breast Lesion Classification

Eduarda Caldeira[1,2]
up201906930@edu.fe.up.pt
Eduardo Meca Castro[1,2]
eduardo.m.castro@inesctec.pt
Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt

[1] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal

[2] INESC TEC
Porto, Portugal

## Abstract

Radiologists are often requested to assess many screenings, which may lead to delays in diagnosis and the treatment of patients. Machine learning (ML) algorithms integrated in computer-aided diagnosis systems may work as a reliable solution to this problem. This study aims to evaluate the performance of a curriculum learning (CL) based algorithm in the breast lesion classification task. We ordered the training samples from the easiest to the hardest, considering the degree of confidence of radiologists in their ground-truth annotation. CL and baseline models achieved similar maximum validation accuracy values (74.42% versus 75.29%) and accuracy values of the best model in the test set (70.71% versus 70.08%). Results suggest that the CL approach was not performing better than the baseline in the lesion classification task.

## 1 Introduction

According to the World Health Organization (WHO), cancer is the leading cause of death worldwide, constituting a global health concern [10]. Sung et al. [7] estimated a total of 19.3 million new cancer cases and almost 10 million cancer deaths worldwide in 2020, estimating a total cancer burden of 28.4 million cases in 2040. Female breast cancer was the most commonly diagnosed cancer (11.7%) and the fifth leading cause of cancer mortality worldwide (6.9%). A breast cancer mammogram consists of a low-dose X-ray of the breast. X-rays of each breast are acquired from two distinct positions: craniocaudal (CC) and mediolateral oblique (MLO). The main goal is to maximise the percentage of the examined tissue. Mammographic screening is performed at regular time intervals to check for breast cancer in women who do not present signs or symptoms of the disease, allowing for early cancer detection [1]. Radiologists are often simultaneously presented with several screenings to classify, which may lead to delays in the classification process and treatment of diseased patients. The usage of machine learning (ML) techniques not only helps solve this problem but also allows for quick and efficient diagnosis and improves early-stage cancer detection [2]. Curriculum learning (CL) is one ML subdiscipline inspired by the human learning process [9]. Human education is organized according to a curriculum, i.e., a way of organizing a group of concepts from the easiest to the hardest. Since the data available in the datasets used for model training is generally heterogeneous in difficulty level, CL has proven to be useful in some ML tasks, enhancing the model's performance and assuring faster convergence. All CL methods follow the same two components framework to design an appropriate curriculum for their tasks [9]:

- **Difficulty Measurer (DM)** — The DM is responsible for organizing the training data by difficulty level. When this data is sorted from the easiest to the hardest samples, it is passed to the Training Scheduler;
- **Training Scheduler (TS)** — The TS defines the weights of the samples in each training epoch. The Baby Step method is a discrete TS that uses ordered information provided by the DM to distribute the training examples by smaller subsets, which are fed to the model after specific criteria are met, starting from the easiest subset to the hardest one.

This study aimed to evaluate the performance of a CL-based model in the breast lesion classification task. To fulfil this goal, we developed two algorithms: a baseline model and a CL model with a Baby Step TS. Additionally, we tested two alternative algorithms: an inverted CL methodology and a task separation strategy.

## 2 State of the Art

According to Sechopoulos et al. [6], the artificial intelligence (AI) revolution in computing led to the usage of ML techniques in breast lesion detection and diagnosis with remarkable results. CL has also proven to be efficient in the medical field. Jiménez-Sánchez et al. [3] conducted a study that compared different CL-based strategies to classify proximal femur fracture types from X-ray images. The CL approaches performed up to 15% better than the baseline, achieving the performance of experienced trauma surgeons. CL strategies have also been applied in the study and classification of breast lesions. Nebbia1 et al. [5] proposed CL methodologies that weighted lesion's features in order to define their classification level of difficulty. This study proved that the CL approaches outperformed the baseline, showing that the incorporation of medical knowledge to build a curriculum improved the classification performance. Although some studies have been conducted with promising results, there is little prior work in CL for medical image analysis [6]. The globally higher performances achieved by this type of model suggest that more studies should be conducted in this field, allowing for its expansion. The main objective of this study is in line with this premise.

## 3 Methodology

The dataset used in this study was extracted from the Digital Database for Screening Mammography (DDSM). This database's images are commonly used in the development of algorithms to aid in the diagnosis of breast lesions [8], such as the one tested in this study. The DDSM contains several reports with information regarding mammography images and patients. To generate an appropriate dataset for the aim of this study, the necessary information was extracted from these reports (see Figure 1). This information included a value between 1 and 5 indicating the subtlety level of the lesion, i.e., how difficult it was to diagnose (higher subtlety values correspond to images easier to classify). The dataset information was divided in three subsets in order to facilitate its usage by a machine learning algorithm: train (71.86% of the images), validation (8.10% of the images) and test (20.04% of the images). To test the CL approach for the classification of mammary lesions, we developed two models: a baseline model and a CL-based model. The only difference between the studied approaches was the strategy used to feed the images to the models during the training phase, since the CL model was initially fed with easier samples and progressively given access to the harder ones, using the subtlety score of the images as the DM. For the CL approach, we defined a Baby Step TS. As a secondary goal of this study, two variations of the previously described methods were tested: an inverted CL model (fed with harder samples first and progressively given access to the easier ones) and a task separation strategy (feature extraction + lesion classification).

## 4 Implementation and Results

### 4.1 Difficulty Measurer Selection

To determine whether the subtlety value was a good DM criterion, we evaluated the confusion matrices of the baseline model on the test set for each subtlety value. The results showed that the percentage of true benign lesions (TBs) increased a total of 29% from subtlety level 1 (52%) to 5 (81%), while the percentage of true malignant lesions (TMs) increased a total of 26% from level 1 (48%) to 5 (74%). Since the percentage of correctly predicted labels was higher for bigger subtlety values, i.e., for lesions that the radiologist considered easier to classify, the subtlety value was considered a good DM criterion.

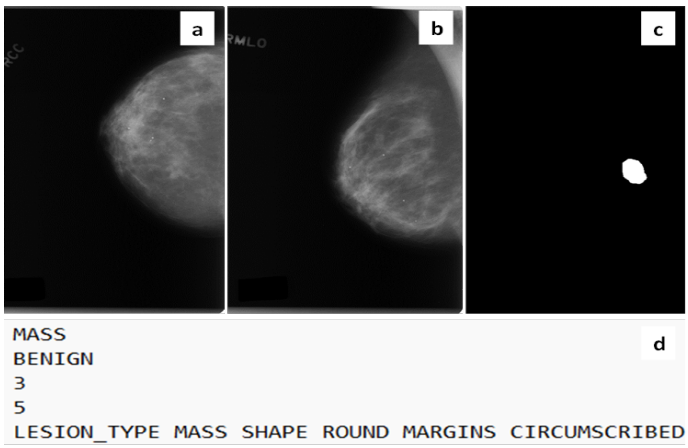


Figure 1: Example of the information extracted from the DDSM dataset: a. CC view; b. MLO view; c. lesion mask of the benign lesion identified in the CC view; d. file with lesion's classification information, based on CC view: type of lesion, lesion's classification, the BI-RADS classification, which standardizes lesion categorization [4], subtlety level and lesion specific attributes (lines 1 through 5).

### 4.2 Results Analysis

The accuracy values considered relevant for this study (maximum accuracy reached in the validation set during the training phase and correspondent accuracy in the test set) were similar for Baby Step and baseline models (74.42% versus 75.29% and 70.71% versus 70.08%, respectively). These results suggest that the CL approach was not influencing the performance of the model, probably due to the used dataset (amount of data available and its distribution by label and subtlety). The percentage of correctly classified benign and malignant lesions for each subtlety value are presented in Table 1. Both models are generally better in the classification of lesions with a high subtlety value. The fact that the TB percentages are higher than the TM percentages for all the subtlety levels may mean that the models are biased towards benign lesions, i.e., that they tend to classify more lesions as benign resulting in more correct guesses when they are, in fact, benign. Regarding the Baby Step model, the easier subsets still show TB and TM percentages much higher than the difficult ones, when compared with the baseline. Since the CL approach should allow for a better understanding of the hard samples by studying the easy ones first, these results led to the conclusion that the Baby Step algorithm did not improve model performance. The analysis of Table 1 also shows that the Baby Step algorithm results in a generally higher percentage of TM classifications. The maximization of this parameter is desired since false negative diagnosis have a highly negative impact in the patient's probability of surviving the tumor. Thus, this model was considered to perform better than the baseline in this matter. However, the baseline model performs better than the Baby Step in a big number of cases, meaning that both models' performances should be considered equivalent, with the Baby Step model performing better in malignant lesions classification. We also verified that the execution time was smaller for the Baby Step model, as expected considering the model's design. None of the alternative approaches managed to fulfill their goals, resulting in a decrease in the model's performance, when compared with the previously implemented strategies.

## 5 Conclusion and Future Work

All the studied models learned the easy samples better, resulting in poor performance for images with low subtlety values, accompanied by undesired bias and overfitting to the easier lesions. Since all the studied models presented these flaws, the tested CL approaches proved to be inefficient in their correction. The studied models were also biased towards benign images, affecting the TM percentages negatively. This information and the fact that the maximum accuracies reached with baseline and Baby Step approaches were similar led to the conclusion that the implemented CL algorithm did not improve the model's performance. However, the CL approach resulted in a bigger percentage of TM classifications accompanied by a reduction in the execution time of the algorithms. Thus, the CL approach based on a Baby Step TS presented some advantages over the baseline. The studied variants of CL didn't improve the performance. On the other hand, it's important to notice that the verified problems could be at least partially attributed to the used dataset, which had predominantly easy and benign images. Datasets with lesions evenly distributed by the subtlety levels and labels should be tested, to reduce both biases. These biases' influence could also be reduced by the implementation of a weighted loss function. This modification is also expected to increase TM percentages, thus improving the predictive performance of the models.

Table 1: Accuracy of baseline and Baby Step models on the test set. Bold cells represent the cases where the usage of the Baby Step method resulted in an improvement of the obtained results.

| Subtlety | Model | TB (%) | TM (%) |
|---|---|---|---|
| 5 | Baseline | 84 | **66** |
| | Baby Step | 84 | **75** |
| 4 | Baseline | 80 | 76 |
| | Baby Step | 69 | 62 |
| 3 | Baseline | 82 | **43** |
| | Baby Step | 74 | **59** |
| 2 | Baseline | 78 | **43** |
| | Baby Step | 70 | **54** |
| 1 | Baseline | 67 | 48 |
| | Baby Step | 63 | 42 |

## Acknowledgements

This work was financed by the ERDF – European Regional Development Fund – through the Operational Programme for Competitiveness and Internationalisation – COMPETE 2020 Programme – and by National Funds through the Portuguese funding agency FCT – Fundação para a Ciência e a Tecnologia – within the PhD grants "*SFRH/BD/136274/2018*" and "*2020.06434.BD*".

## References

[1] DenseBreast-info. Screening technologies - mammography, 3d mammography (tomosynthesis), july 2022. https://densebreast-info.org/screening-technologies/mammography-3d-mammography-tomosynthesis/.

[2] Yassir Edrees Almalki et al. Computerized analysis of mammogram images for early detection of breast cancer. *Healthcare*, 10(5), april 2022.

[3] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A. González Ballester, and Gemma Piella. Medical-based deep curriculum learning for improved fracture classification. *MICCAI*, 2019.

[4] Tariq Mahmood, Jianqiang Li, Yan Pei, Faheem Akhtar, Mujeeb Ur Rehman, and Shahbaz Hassan Wasti. Mammographic classification of breast lesions amongst women in enugu, south east nigeria. *Afri Health*, 17(14):1044–1050, december 2017.

[5] Giacomo Nebbia1, Saba Dadsetan, Dooman Arefan, Margarita L. Zuley, Jules H. Sumkin, Heng Huang, and Shandong Wu. Radiomics-informed deep curriculum learning for breast cancer diagnosis. *MICCAI*, 2021.

[6] Ioannis Sechopoulos, Jonas Teuwen, and Ritse Mann. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. *Seminars in Cancer Biology*, 72:214–225, july 2021.

[7] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer Journal*, 71(3):209–249, may 2021.

[8] USF University of South Florida. Ddsm: Digital database for screening mammography, july 2022. http://www.eng.usf.edu/cvprg/mammography/database.html.

[9] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, march 2021.

[10] WHO World Health Organization. Cancer, july 2022. https://www.who.int/news-room/fact-sheets/detail/cancer.