ProjectionNet and LRNet: Light and Fast ConvNets for Mammography Screening Projections Classification

Radamés Ramírez
UAG

Marco Iñiguez
UAG

Abraham Sánchez
Gobierno de Jalisco

Alejandro Sánchez
Gobierno de Jalisco

Sofía Salas
IMSS

Claudia Haro
IMSS

Miguel Cabrera
IMSS

E. Ulises Moya Sánchez
Gobierno de Jalisco-UAG
eduardo.moya@jalisco.gob.mx

Abstract

Breast Cancer (BC) is the most frequent malignancy in women worldwide. The most effective technique for detecting BC when the treatment can be more successful is the mammography. Previous research using Artificial Intelligence (AI) presented outstanding systems capable of surpassing human experts in breast cancer detection using mammography images. However, the main obstacle to develop the clinical implementation of AI is the availability of sufficiently curated, and representative training data that includes expert labeling. In Mexico, we found a lot of mammography images without any projection label, making it very difficult to prepare these images and test these new AI methods. In this work, we present two new simple, light and fast Convolutional Neural Networks (CNN) models to help classify the mammography projections in terms of left or right (L/R) and Craniocaudal or Mediolateral oblique (CC/MLO). We compare the proposed model, against three state-of-the-art CNN using two datasets. The results confirm that the proposed models achieve the best classification performance with fewer computational resources and significantly less inference time by a factor range between 3.4-13.7. We hope that these models help to reduce the time-intensive data preparation, increasing the data access from more and diverse sources.

1. Introduction

According to the World Health Organization (WHO), Breast Cancer (BC) is the most common malignancy of women. The latest statistics of GLOBOCAN 2020, estimate that BC was the fourth deadliest type of cancer worldwide (6.9 %) [2]. The only method that has proven to reduce the mortality rate by 20-30% is mammography screening. Screening programs should have a coverage that exceeds 70% of the target population, which in turn generates a positive impact on disease-free survival period and overall survival [1, 8].

Recently [5, 16] showed that Deep Learning (DL) models can outperform a group of radiologists in the BC classification of mammograms. However, most of these methods are based on a single (and specific) mammography projection. Typically, two projections, Craniocaudal or Mediolateral oblique (CC/MLO) are performed on each breast (left/right) in the screening process. A major obstacle to developing clinical AI applications is that the data requires a curation and labeling process to obtain better performance [15]. In many hospitals in Mexico, we found a lot of mammography images without any projection labeling, despite labeling being a quality control requirement in all mammographic studies [10], making it very difficult (time-intensive) to prepare a dataset and test these new DL methods. Although most of these hospitals obtain their images in Digital Imaging and Communications in Medicine (DICOM) format, the DICOM-metadata is commonly missing. As a result, we consider it feasible to develop new tools to classify mammography projections to prepare and curate the mammography data.

A popular approach in DL is the use of a pre-trained network and customize it for another task. In general, the benefits of the pre-trained models are saving training time and getting better performance. Nevertheless, for some classification tasks, a simple model could be more accurate and faster. The main contribution of this work, is to present two new simple CNNs capable to classify the L/R and CC/MLO mammography projections with the best tradeoff between accuracy performance and number of weights. ProjectionNet is devoted to classify between CC and MLO projections. LRNet is capable to classify the left and right breast in the MLO projection. We compare this two CNNs against well known multi-propose CNNs: MobileNetV2 [9] DenseNet121 [3], and InceptionV3 [14].

2. Data

We have used the Curated Breast Imaging Subset (CBIS) [4], and mini-MIAS (m-MIAS) database of mammograms [12]. Table 1 presents the number of train, validation and test images. Note that in the testset we use a different dataset to measure how well the models generalize.

Table 1: Train, validation and test data.

Class	Train_{CBIS}	Val_{CBIS}	Test_{CBIS}	Test_{MIAS}
CC/MLO	598	67	167+275	322
L/R	590	66	164	322

3. ProjectionNet and LRNet

Although CNNs architectures are commonly manually designed by experience, we combine our experience with an expert-knowledge, image analysis and Bayesian fine tuning to select the most appropriate architecture.

To obtain the expert-knowledge, we made a review about the mammography screening projections [13, 6, 7]. The review understands the main anatomical characteristics of both projections (pectoralis muscle, perimeter, posterior nipple line, nipple, wide margin at the axilla, among others). According to this information and our experience, this problem is a coarse grain classification, based on the size of the features we expect to be learned by the models. Also, we propose to reduce the input size (image) to 32×32 to help speed up the training and inference stages to accomplish our objective (simplicity).

As image analysis, we have used the first convolutional layer activations (see supplementary material) and the Grad-Cam visualization in order to evaluate the most important features. It is important to remark that most of the mammography screening images are visually labeled with a text landmark and this technique helps us to know if the CNN learns the breast features or the text labels.

Finally, we use a Bayesian optimization search algorithm in order to search and compare similar CNN hyperparameters. Table 2 presents the Bayesian hyperparameter search space. We performed 6 different searches establishing a maximum of 15 trials running for 20 and 50 epochs for CC/MLO and L/R tasks respectively, monitoring the validation accuracy as metric to get the best model on each search. Selected hyperparameters and classification performance over the test dataset are presented in the Tables of *supplementary material*. The proposed models PNet and LRNet have the highest test accuracy and lowest test loss value in their respectively search group. Figure 1 presents the PNet and LRNet architectures.

Table 2: Bayesian optimization hyperparameters search space.

Hyperparameter	CC/MLO	L/R
Act func [conv]	[ReLU, tanh]	
Neurons [conv]	[mn: 8, mx: 128, s: 8]	
Kernel	[3, 5, 7]	
Stride	d:1	[1, 2]
Max pooling	[2, 3, 4, 5]	
Dropout	[mn:0.0, mx:0.7]	[mn:0.0, mx:0.5]
Act func [dense]	[ReLU, tanh]	
Neurons [dense]	[mn:32, mx:512, s:32]	[mn:8, mx:128, s:8]
Learning rate	[mn:1e-5, mx:1e-2, sp:"LOG", d:1e-3]	

*[mn=min value, mx=max value, s=step, sp=sampling, d=default value]

*[LOG=logarithm probability distribution]

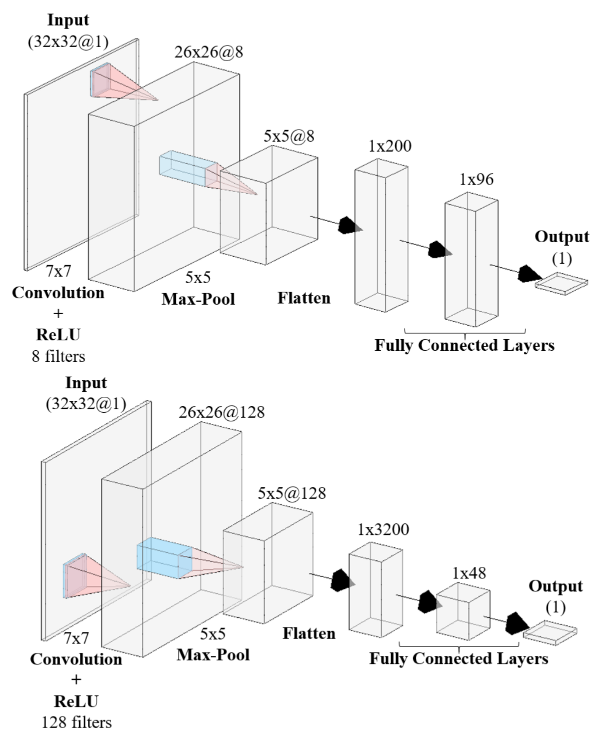


Figure 1: ProjectionNet (top) and LRNet (bottom) architectures.

4. Experimental setup

We compare the proposed CNNs: ProjectionNet (input shape: 32, 32) and LRNet (input shape: 32, 32) against MobileNetV2 (input shape: 224, 224) [9] DenseNet121 (input shape: 224, 224) [3], and InceptionV3 (input shape: 180, 180) [14] using the same hardware[1] and same DL framework: Tensorflow 2.1. We used 10-fold validation in order to obtain the best model. For all models we use the same preprocesing steps and the same training conditions (20 epochs, ADAM optimizator, data normalization, binary cross-entropy loss function, etc). The proposed CNN (PNet and LRNet) were trained from scratch, whereas the others architectures were pretrained with the ImageNet weights. As data augmentation in all the cases we used random rotation of 20 degrees and random zoom of 0.1.

The evaluation of their classification performance was made by measuring the test accuracy, test loss, and AUC. Additionally, we compare the number of trainable parameters and the inference time of each model. We use the Grad Cam technique [11] in order to analyze the most important features of the models.

5. Results and analysis

Tables 3 and 4 show the classification performance comparison with pre-trained models MobileNetV2 (MNetV2), DenseNet121 (DNet121) and InceptionV3 (IncepV3). [2]

Table 3: Performance comparison between PNet and pre-trained CNNs for CC/MLO classification task

Metric	**PNet**	MNetV2	DNet121	IncepV3
Test acc	**0.995**	0.977	0.958	0.940
Test loss	**0.026**	0.090	0.174	0.767
AUC	**0.999**	0.998	0.997	0.979
Parameters*	**20**	2,320	7,090	21,820
Inf. Time**	**7**	37	99	50

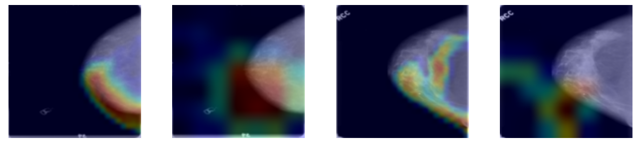
*Thousands(K) **Average time (milliseconds) per test image

Table 4: Performance comparison between LRNet and pre-trained CNNs for L/R classification task

Metric	**LRNet**	MNetV2	DNet121	IncepV3
Test acc	**0.988**	0.798	0.502	0.770
Test loss	**0.030**	4.017	39.820	3.999
AUC	**0.999**	0.80	0.734	0.847
Parameters*	**160**	2,320	7,090	21,840
Inf. Time**	**6**	20	34	29

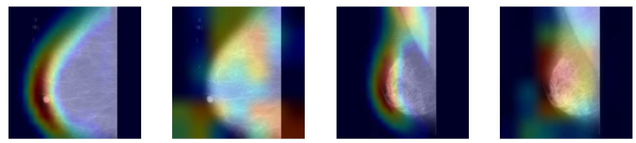
*Thousands(K) **Average time (milliseconds) per test image

The proposed CNNs have the highest test accuracy with the lowest test loss and the best AUC for both classification problems, showing a better generalization in both test



(a) (b) (c) (d)

Figure 2: Grad-CAM visualization comparison between PNet (a), (c) and MNetV2 (b) (d) for CC/MLO classification.



(a) (b) (c) (d)

Figure 3: Grad-CAM visualization comparison between LRNet (a), (c) and MNetV2 (b) (d) for L/R classification.

datasets [4, 12]. The MNetV2 has the second-one performance. The proposed CNNs show a significant difference in the number of trainable parameters with a range of factors between 14.5-136.4 for LRNet and 117.2-1102.5 for PNet, while inference time is improved by a range of 3.4-5.7 for LRNet and 5.1-13.7 for PNet. This finding confirms that the proposed CNNs have the best trade-off between classification performance and inference time, using smaller and simpler architectures. Additionally, we made a more detailed comparison with MNetV2 trained with 100 epochs (12 hours), for CC/MLO classification it achieves a test accuracy of 0.986 and test loss of 0.185. For L/R classification, MNetV2 obtains 0.956, 0.325 of test accuracy and test loss respectively. These results are slightly better than the 20 epoch MNetV2. However the proposed CNNs have better performance. In Figures 2 and 3 we present the gradient-weighted class activation mapping (Grad-CAM) proposed by [11] of each classification task, where the red color represents the most important features. It is possible to note that localization maps of the proposed CNNs are highlighting the anatomical regions of the breast (pectoral muscle, nipple, edges, perimeter). In contrast, the MobileNetV2 gradients focuses mainly in broad areas with empty space.

6. Conclusion

In this work, we address a fundamental step for preparing mammography screening images by classifying the CC/MLO and L/R projections. The proposed CNNs trained have an exceptional classification performance and inference time tested against three state-of-the-art DL models. The results demonstrate that smaller and simpler architectures such as the proposed CNNs can be better to solve the

[1] 2GB NVIDIA GeForce GTX 950M, i7-6700HQ CPU @2.60GHz 2.59 GHz, RAM: 16 GB

[2] If we train from scratch these models, the classification performance is significatively reduced under the same train conditions.

specific classification tasks that we addressed. The value of this work lies in the fact that these models could help to reduce the time-intensive data preparation-curation, increasing data access. As a result, this work will help to facilitate the access to more mammography data from unlabeled projections, even with any DICOM information. As future work, we would like to explore the deployment of these models in the cloud due to their quick inference time and also to embedded systems because of their low development cost and portability.

Acknowledgements

We wish to acknowledge the help provided by the following institutions: Secretaría de Salud Jalisco, Clínica de la mama del Instituto Mexicano del Seguro Social (IMSS), Coordinación General de Innovación Gubernamental del estado de Jalisco, Postgrado en Ciencias Computacionales de la Universidad Autónoma de Guadalajara and CONACYT.

References

[1] Joann G Elmore, Katrina Armstrong, Constance D Lehman, and Suzanne W Fletcher. Screening for breast cancer. *Jama*, 293(10):1245–1256, 2005. 1

[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, and Jemal A. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries., 2020. https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf. 1

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2, 3

[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, 2017. 2, 3

[5] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. 1

[6] Louise C. Miller. Mammography Positioning, 2018. https://www.sbi-online.org/Portals/. 2

[7] Manju Bala Popli, Rahul Teotia, Meenakshi Narang, and Hare Krishna. Breast positioning during mammography: mistakes to be avoided. *Breast cancer: basic and clinical research*, 8:BCBCR–S17617, 2014. 2

[8] Christine Renshaw, Ruth H Jack, Steve Dixon, Henrik Møller, and Elizabeth A Davies. Estimating attendance for breast cancer screening in ethnic groups in london. *BMC Public Health*, 10(1):1–8, 2010. 1

[9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2

[10] ISBN 970-721-092-3 Secretaria de Salud. Manual de Control de Calidad en Mastografia, 2002. http://cnegsr.salud.gob.mx/contenidos/descargas/cama/manualCtrlCal.pdf. 1

[11] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 3

[12] P SUCKLING J. The mammographic image analysis society digital mammogram database. *Digital Mammo*, pages 375–386, 1994. 2, 3

[13] Rhonda-Joy I Sweeney, Sarah J Lewis, Peter Hogg, and Mark F McEntee. A review of mammographic positioning image quality criteria for the craniocaudal projection. *The British journal of radiology*, 91(1082):20170611, 2018. 2

[14] Christian Szegedy, 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*, pages 2818–2826, 2016. 2, 3

[15] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020. 1

[16] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzkebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 2019. 1