ISSN 1870-4069

Evaluation of View-wise ResNet on the Digital Database for Screening Mammography

Osmar Moreno-Rivas[1], Alfonso Rojas-Domínguez[1], Matías Alvarado[2], Manuel Ornelas-Rodríguez[1]

[1] Tecnológico Nacional de México,
campus León,
Mexico

[2] Instituto Politécnico Nacional,
Centro de Investigación y Estudios Avanzados,
Mexico

m22240203@leon.tecnm.mx, alfonso.rojas@gmail.com, matias@cs.cinvestav.mx, manuel.ornelas@leon.tecnm.mx

Abstract. Screening mammography aided by deep learning classifiers has demonstrated enhanced efficacy by reducing False Positives, consequently minimizing unnecessary recalls that cause anxiety among participants. However, the development of effective classifiers necessitates substantial computational resources and a vast amount of training data. Despite these requirements, it is generally assumed that these models possess a high level of generalization, enabling them to perform well on similar datasets to the ones they were trained on. In this study, we assess the performance of a ResNet-based model for screening mammography presented by Wu et al. (2019). This model was trained on an extensive dataset of over one million images and reported an Area Under the ROC curve (AUC) of 0.88. Previous studies have fine-tuned similar models using additional data, achieving AUC values around 0.9. However, these studies had limited sample sizes in their test sets, consisting of only a few hundred images, thereby restricting the applicability of their findings and conclusions. In contrast, our evaluation utilizes the DDSM, the largest publicly available dataset for screening mammography, containing over 10,000 images. The evaluated model achieved an AUC of approximately 0.50, significantly lower than the performance reported by other authors.

Keywords: Residual networks, deep learning, breast cancer screening.

1 Introduction

Breast cancer is the most commonly diagnosed cancer among women; the World Health Organization estimates that 7.8 million women were diagnosed with breast cancer in 2021. Moreover, it was estimated that in 2022 30% of diagnosed cancers in women would be breast cancer according to the National

Breast Cancer Foundation[3]. However, the lifetime breast-cancer survival rate increases with timely diagnosis; this is why new detection and diagnostic techniques for breast cancer are constantly being developed. The most common acquisition technique employed in screening tests for detection of breast cancer is mammography, because it is a non-invasive and low cost technique compared with other techniques like MRI (magnetic resonance imaging) or CT-scan (computerized tomography). Screening tests for early detection of breast cancer through mammography can be considered a first line of defense, from which a few cases that warrant further testing can be identified (see Fig. 1). Screening tests are indicated for women who have not presented any symptoms potentially indicative of breast cancer, but who belong to an age-range in which prevalence of the disease is the highest. These studies have the ability to detect abnormalities that cannot be felt through palpation or self-examination.

Computer-aided diagnosis (CAD) systems have been successfully used to support human decision-making in radiological image analysis and precision medicine in general [4]. Traditional approaches to breast cancer CAD involve extracting manually-designed features to detect breast masses and classify them as probably benign or malignant [5, 11]. However, the outputs from these CAD systems in conjunction with radiologists' reviews result in numerous false-positives, which can increase reading times [6]. Alternative approaches involve learning features directly from the full images through deep neural networks [8]; we can list a variety of such: Convolutional Neural Networks (CNNs), Residual Networks (ResNets), Dense Networks, among others [11].

In particular, ResNets have shown favorable results on detection and classification of breast cancer. Xiang Yu et al. 2020 [14] obtained an average accuracy of 95.74% correct classification on the MINI-MIAS and InBreast datasets. On the other hand, Y. Chen et al. [1] has reported an accuracy of 93% with a CCN-based model fine-tuned with a ResNet architecture on the CBIS-DDSM [9] database. It is worth mentioning that the databases that have been employed in those previous studies contain only a few hundred images, and thus the reported results are limited. Our contribution is the evaluation of a reportedly efficient model [13] on DDSM, a public dataset with over 10,300 mammographic images.

Honig et al. 2019 [7] conducted a study about impact factors of False Positives (FPs) in recall cases; their study found that 91.6% of 1,258 recalled cases were FPs. Thus, a vast majority of women who received a recall notification had not actually developed breast cancer, despite initial screening results suggesting otherwise, which can result in unnecessary procedures and psychological effects like elevated anxiety in women [2]. This justifies further research towards the improvement of breast cancer screening systems with tools like DL (Deep Learning) models, to ensure that women receive the most accurate and reliable information about their health. On the other hand, DL-CAD systems have proved the reduction of FPs per image to 69% in comparison with traditional CAD systems that often yield a higher number of FPs [10]. Moreover, DL-CAD

[3] https://www.nationalbreastcancer.org

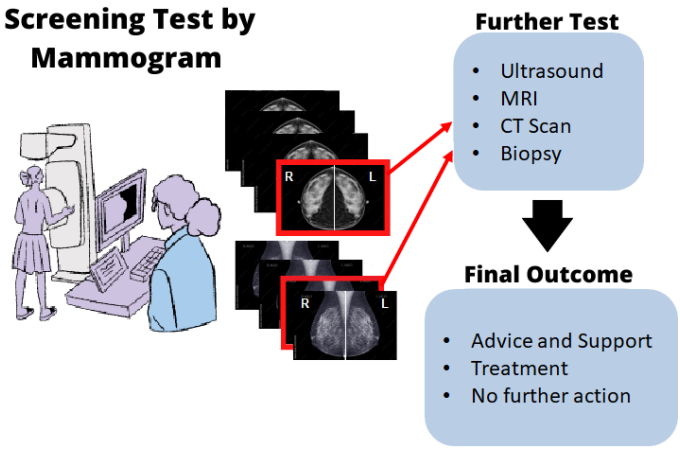

Fig. 1. Breast cancer mammography-based screening test. A very small portion of the cases (red outline) require further tests such as Ultrasound, MRI, or a biopsy. After further detection and diagnostic tests, a final outcome can be reached.

systems reduce 17% the reading time per case performed by radiologists in comparison when they used CAD systems [10], which in turn improves the benefit-cost ratio for massive studies of breast screening in women. Additionally, these systems provide more precise results and report them in less time.

2 Materials and Methods

2.1 DDSM Database and Inclusion Criteria

The Digital Database for Screening Mammography (DDSM) is a public resource that contains 2,588 exams for detection of breast cancer, including two standard anatomical views: Cranio-Caudal (CC) and Mediolateral Oblique (MLO). The images have an average size [height × width] in pixels, for CC: $[2,677 \times 1,942]$ and for MLO: $[2,974 \times 1,748]$. The optimal size stated by Wu et al. [13] to run their model is $[2,290 \times 1,890]$; thus the DDSM images mostly comply with these parameters. There are three possible outcomes for each study in the DDSM: *benign*, *malignant* and *no finding* (or *normal*).

An exam generally consists of four mammograms: L-CC (Left breast-CC view), R-CC (Right breast-CC view), L-MLO (Left breast-MLO view) and R-MLO (Right breast-MLO view). Nevertheless, five exams in the DDSM only include three images. We also found 209 studies that have more than one overlay in the same image (meaning that there is more than one abnormality present in one image). In 130 cases in which an abnormality could only be found in one of the two views (either CC or MLO, but not both). Finally, we found 5 studies with duplicated patient ID. After excluding the studies described above, we ended up with a total of 2,244 studies to be used in our evaluation.

2.2 Neural Model for Screening Mammography

In 2019 Wu et al. described a multi-view neural-network system for screening mammography based on ResNets [13] which consists of two core modules: (i) four view-specific columns that output a fixed-dimension hidden representation for each mammography view, and (ii) two fully-connected (FC) layers to map the hidden representations to the output predictions [13]. Depending on how the hidden representations are aggregated into a final prediction, four different models are produced: View-wise, Image-wise, Side-wise and Joint models.

According to Wu et al., the View-wise model obtained the best results among their models, with an Area under the ROC Curve (AUC) of approximately 88%. Consequently, this View-wise model is employed in our evaluation and is described below. A schematic representation of the model is shown in Fig. 2. In the view-wise model, the 256-dimensional hidden representations of the CC views (L-CC and R-CC) are concatenated together before going through the FC layers. Independently, the representations of the MLO views (L-MLO and R-MLO) are also concatenated together and pass through their own FC layers. This process produces independent predictions for CC and MLO views, which are averaged during inference to produce the breast-wise predictions [13]. At the top of Fig. 2 it can be seen that the model produces four numerical predictions (two for each breast) named Right-Benign: $\hat{y}_{R,b}$, Right-Malignant: $\hat{y}_{R,m}$, Left-Benign: $\hat{y}_{L,b}$ and Left-Malignant:$\hat{y}_{L,m}$. These predictions are to be compared against binary labels that correspond to the ground truth of the cases in the evaluation dataset.

2.3 Model Predictions

To evaluate the View-wise model its predictions are binarized and compared against binary labels that represent the pathology of each case in our evaluation dataset. The binary labels (two per breast, four per study) indicate the presence (1) or absence (0) of a finding, either Benign or Malignant, in the corresponding breast. Table 1 shows an example of the labels for one patient; in this example, a malignant finding is present in the left breast and a benign finding is present in the right breast. The view-wise model produces four numerical predictions between 0 and 1; binarization of the predictions is done by comparing these against a detection threshold, as illustrated in Fig. 3.

Whenever the value of a numerical prediction is equal to or above the detection threshold, it is assigned a value of 1, otherwise it is assigned a value of 0. If the values of both malignant and benign predictions of the same breast surpass the threshold, we assign 1 to the malignant prediction and 0 to the benign prediction; this is done to avoid contradictory predictions and to prioritize the detection of malignant findings over benign ones. Importantly, we test ten different threshold values, regularly spaced between 0 and 1, to generate ten different sets of results with which to evaluate the model and build a ROC curve.

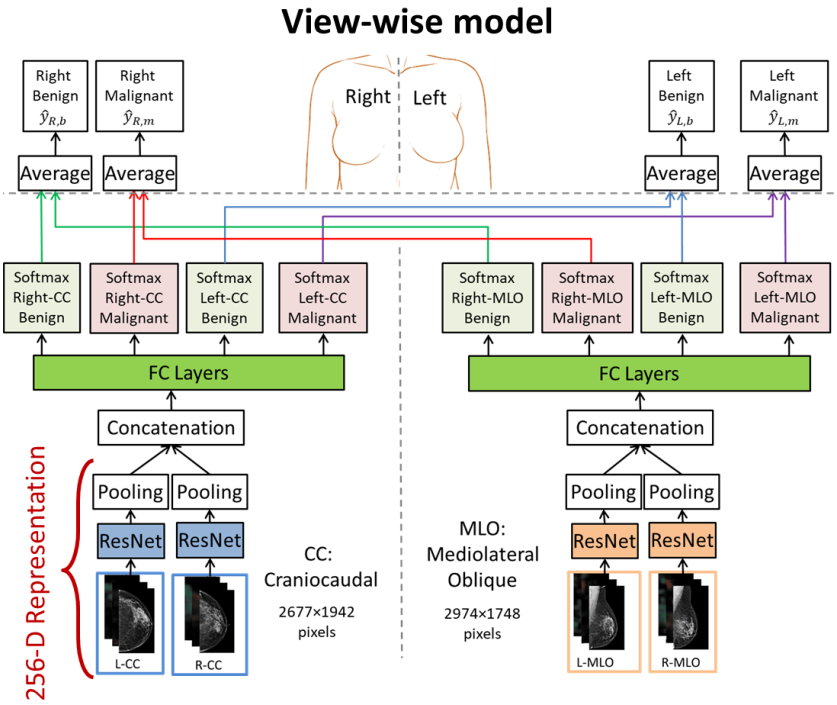

Fig. 2. Schematic representation of the View-wise model proposed by Wu et al. [13].

Table 1. Example of labels for one mammographic study.

Left Benign	Left Malignant	Right Benign	Right Malignant
0	1	1	0

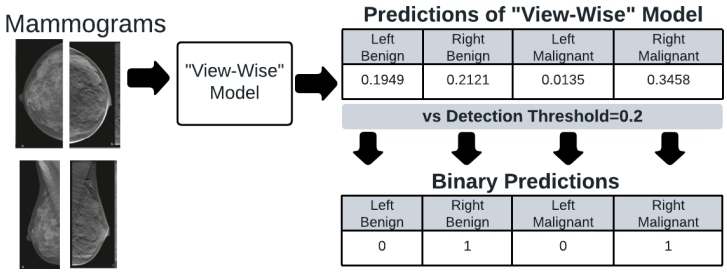

Fig. 3. Example of binary predictions generated by thresholding; a binary prediction is obtained for each side and type of finding. Notice that the threshold=0.2 is only used to illustrate the process; in the actual evaluation ten threshold values were employed.

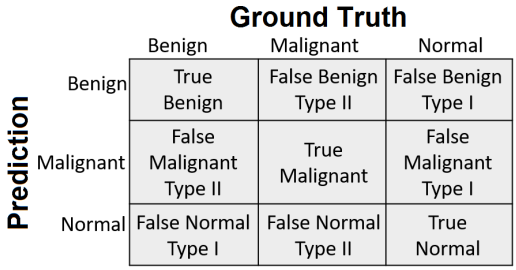

Fig. 4. Multi-class confusion matrix employed to evaluate the model; notice the different possible errors for each of the three classes: *Benign*, *Malignant* and *Normal*.

2.4 Model Evaluation

To evaluate the classification performance of the model, a multi-class confusion matrix is employed. There are three classes in our test dataset: *Benign*, *Malignant*, and *Normal* (see Fig. 4). For each class the model's prediction can be correct, or it can be one of two types of misclassification, depending on the Ground Truth (GT). For instance, if the model produces a *Benign* prediction, this may be a True Benign, a False Benign of Type I (GT indicates that the true class is *Normal*), or a False Benign of Type II (the actual class is *Malignant*).

To correctly compute each of the values in a confusion matrix, the binary predictions produced by the model need to be compared against the corresponding GT labels. Figure 5 shows a flowchart of the different comparisons that must be carried out to reach one (and only one) of the nine possible outcomes contained in a confusion matrix. Notice that the predictions and GT labels correspond to individual mammograms, while each confusion matrix corresponds to one of the patient's sides (Left or Right). In this work we tallied the classification results independently, for the Left side and for the Right side.

Complementarily to the multi-class evaluation, we also performed an evaluation of the model in which only the Malignant and Normal classes are considered (Benign cases were treated as Normal). Thus we can examine if there is a difference in the performance of the model when considering only two classes instead of three. In the work of Wu et al. [13] there is not sufficiently detailed information regarding how numerical predictions are binarized and on how the Benign predictions were treated to obtain their final results. Our best assumption is that Benign predictions were ignored in the computation of their ROC curves.

3 Results and Discussion

Table 2 shows a small portion of the results of the model obtained with different values of the detection threshold. As can be seen, the performance of the View-wise model is higher for the smaller values of the detection threshold and

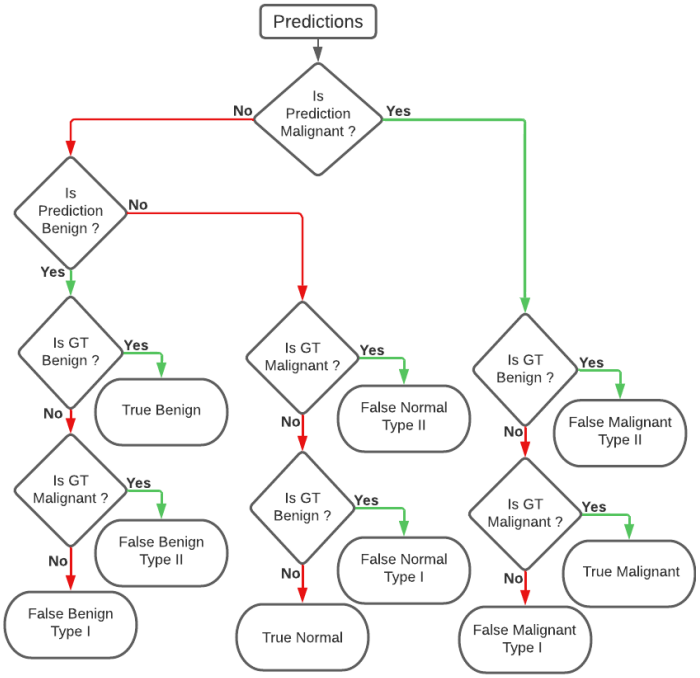

Fig. 5. Flowchart to generated the classifications of each type finding.

gradually decreases as the model is tested with larger threshold values. However, even for the smallest threshold value, the model shows a very low performance.

This low performance is visualized for both sides: in Table 2-a) and 2-d) we can see that class Benign has more instances correctly classified than classes Malignant and Normal. In contrast, Table 2-c) and 2-f) show that class Normal has more instances correctly classified than classes Malignant and Benign. This indicates that with larger threshold values the correct predictions fall much more into the Normal class, because the model does not detect as many abnormalities as with smaller threshold values. On the other hand, correct classifications for class Malignant are always very low, regardless of the threshold value applied. Specially it should be noticed that for larger threshold values there are no hits for instances on the Left side (Table 2-b) and 2-c)) and there is only four instances correctly classified among instances of the Right side (Table 2-e) and 2-f)).

Based on the confusion matrices obtained for the two-class evaluation (on classes Malignant and Normal) for different threshold values, which effectively represent a set of operating points, an ROC curve was obtained for the View-wise model on the DDSM dataset, with an AUC of about 0.5. Clearly this (around 50% correct classification) is not a desirable performance, as it is indicative that the model does not posses any predictive ability. Also, this was not the expected result, given that Wu et al. reported an ROC curve of this same model on their dataset with about 88% correct classification.

Table 2. Example confusion matrices; a) and d): threshold=0.1, b) and e): threshold=0.6, c) and f): threshold=0.9.

			Ground Truth				
	Left Side				**Right Side**		
Class	Benign	Malignant	Normal		Benign	Malignant	Normal
a)	405	338	128	d)	359	302	1232
Pred.	20	37	112		26	55	236
	33	37	1134		3	5	26
b)	0	5	10	e)	9	22	86
Pred.	0	0	1		0	3	2
	458	407	1363		379	337	1406
c)	0	0	0	f)	0	3	2
Pred.	0	0	0		0	1	0
	458	412	1374		388	358	1492

As can be observed, the evaluation on a different dataset (other than that with which it was trained) is not favorable to this View-wise model, as we obtained approximately 33% lower performance on the DDSM than what was previously reported on the NYU dataset [13]. Moreover, we can observe that in both views (CC and MLO) the DDSM images possess close to optimal sizes to be processed by the model, but we suspect that there are other properties that may affect the performance, such as poor contrast, different range of the intensity values, noise in the images etc. Although it is not very probable, image pre-processing could also affect the performance of the model.

Recent investigations have made similar observations. Frazer et al. [3] used models pretrained with the NYU dataset of Wu et al. [13] as the base for their whole-image classifier, observing poor performance (around 55% correct classification). Then the models were retrained with a small subset of DDSM and up to 87% correct classification was obtained. Similarly, Shen et al. [12] found that when applying transfer learning with around 239 images on a model pretrained with CBIS-DDSM, classification of the InBreast dataset improved. Because of this, we attempted to classify the DDSM database using the model proposed by Shen et al. [12]. This model is open source and has an architecture that enables it to classify small patches and extend the patch classifier to the entire image. We obtained similar results with the model proposed by Shen et al. as with the model of Wu et al., achieving an AUC of approximately 0.5. Subsequently, we applied transfer learning with a small subset of the DDSM database. Fig. 7 illustrates the training and validation curves. We observed that the model of Shen et al. exhibited good retraining; however, it did not improve the classification performance as expected. We suspect that this behavior is due to the heterogeneity of the images in the DDSM database, where some images have excessive contrast, while others have a noisy background, as illustrated in Fig. 6.

To address this issue, we clustered the mammograms to obtain image sets with reduced heterogeneity. We first performed segmentation to separate the breast from the background. This was done by applying a manually defined threshold of 128 (half of the grayscale range from 0 to 255). Alternatively we also utilized Otsu's method to determine the optimal threshold value for each

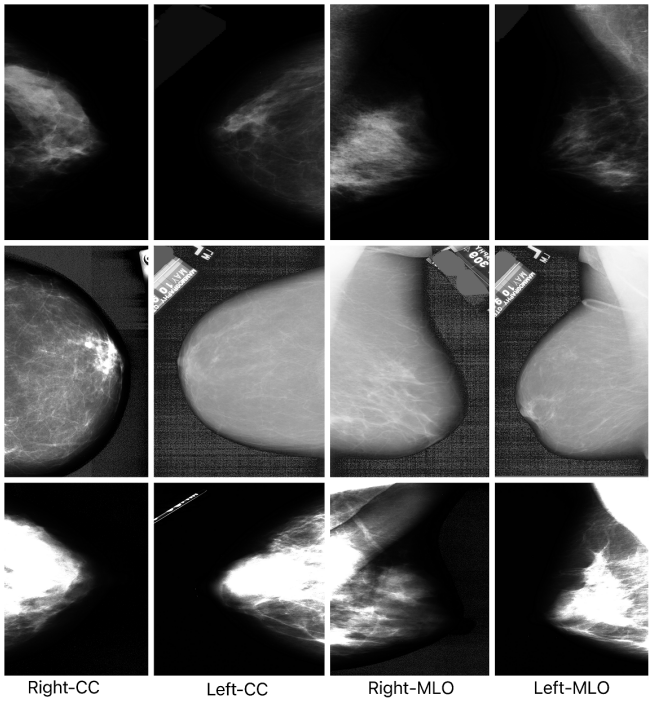

Fig. 6. A few images of DDSM to illustrate the heterogenity in the dataset.

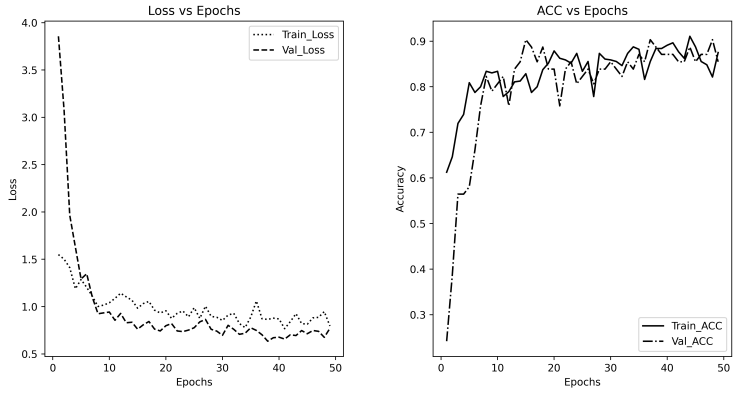

Fig. 7. Results of fine-tuning the model by Shen et al. with 600 images of DDSM.

image, resulting in more accurate segmentation (Fig. 8). Next, background and breast tissue features (mean and standard deviation of the pixel values) were obtained from the images and the K-means algorithm was employed to partition the DDSM images into subsets with similar properties.

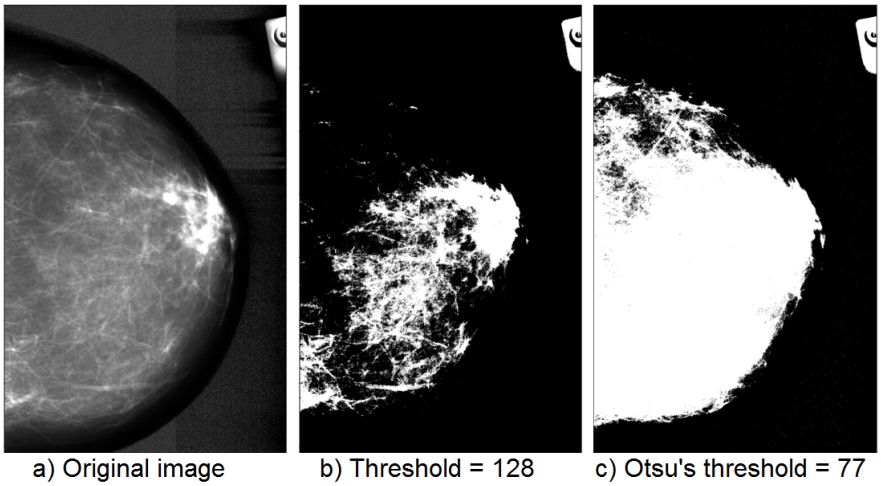

Fig. 8. Image segmentation with manual threshold and Otsu's threshold.

We applied different methods to determine the correct number of clusters: gap statistic, elbow method, and silhouette. All of these methods converge to the most suitable number of clusters, which is 4 (Fig. 9-a). The cluster points represented by the features extracted from the images using the Otsu's threshold are shown in Fig. 9-b.

Analysis of the data clusters in Figure 9-b reveals that cluster 1 comprises images with dark background (small background mean) and bright tissue (large tissue mean). Moreover, the grayscale values of the tissue exhibit significant variation (large standard deviation). Overall, these characteristics indicate well-equalized images.

Finally, the model was fine-tuned separately with 627 images per cluster, in proportions of 90% and 10% for training and validation, respectively, according to the methodology followed by Shen et al. [12], classification results reported in Table 3. The clusters generated by features extracted through Otsu's method display highly similar AUC scores, all above 0.60. Notably, cluster 3 exhibits the best performance in training, with an AUC=0.92. Conversely, the clusters generated via the single threshold method show greater discrepancies in AUC scores on the test set, ranging from 0.38 to 0.82. The lowest score is observed for cluster 2, while cluster 3 achieves the highest score among both methods, reaching an AUC=0.82. Additionally, cluster 3 when generated from a single threshold showcases the highest score in training, AUC=0.83, compared to the other clusters from the same method. The closest to this performance is cluster 1, with an AUC=0.82.

In conclusion, our analysis indicated that using four clusters yields the most appropriate split of the DDSM data. Features extracted through the Otsu's method demonstrate consistent performance across the clusters, while the single

Table 3. AUC results of fine-tuning the subsets on DDSM.

	with Otsu's threshold			**with threshold=128**		
Cluster	Images	Training	Test	Images	Training	Test
0	198	0.69	0.64	313	0.78	0.54
1	462	0.84	0.68	332	0.82	0.66
2	316	0.79	0.69	209	0.60	0.38
3	143	0.92	0.66	259	0.83	0.82

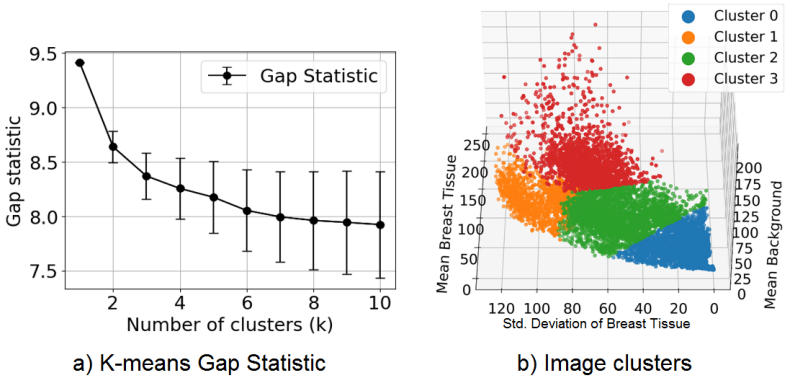

Fig. 9. Determinate the K-value and features representation for Otsu Threshold A) Gap statistic method applied to extracted features to determined best K-value B) extracted features from images, grouped into four clusters through K-means algorithm.

threshold method displays more variability. Specifically, cluster 3 consistently exhibits a good performance in both training and testing, regardless of the threshold method used. These results highlight the value of cluster analysis and careful feature selection for images to reduce heterogeneity in the DDSM database for classification tasks.

4 Conclusions

We conclude that evaluation of the model described by Wu et al. on the DDSM has not been favorable. As we have previously mentioned, we believe that there exist several properties of the images in the DDSM that may be negatively affecting the performance of the model. We also observed that evaluating only two types of classes (i.e. Malignant vs. Normal) does not produce any performance improvement when compared against the evaluation with three classes (Malignant, Benign and Normal). However, as future work we will test another public dataset (for instance, InBreast) to determine if the model's performance changes or not. We will also modify the preprocessing of the images to try to obtain better results. Our purpose is to evaluate the feasibility of employing a pretrained model directly on datasets of the same nature as that in which the model was trained originally (in this case, screening mammograms),

that is, to evaluate in practice the generality of such models. Moreover, it is necessary to experiment with more extensive feature selection for the DDSM database in order to improve the classification performance of the model of Shen et al., in this way we hope to obtain better results when applying fine-tuning and reduce the heterogeneity of the training images.

Acknowledgments. This work was partially supported by the National Council of Humanities, Science and Technology (CONAHCYT) of Mexico, via Postgraduate Scholarship 813768 (O. Moreno) and Research Grant CÁTEDRAS-2598 (A. Rojas).

References

1. Chen, Y., Zhang, Q., Wu, Y., Liu, B., Wang, M., Lin, Y.: Fine-tuning resnet for breast cancer classification from mammography. In: Proceedings of the 2nd International Conference on Healthcare Science and Engineering 2nd. pp. 83–96. Springer (2019)
2. El Hachem, Z., Zoghbi, M., Hallit, S.: Psychosocial consequences of false-positive results in screening mammography. Journal of Family Medicine and Primary Care, vol. 8, no. 2, pp. 419 (2019)
3. Frazer, H. M., Qin, A. K., Pan, H., Brotchie, P.: Evaluation of deep learning-based artificial intelligence techniques for breast cancer detection on mammograms: Results from a retrospective study using a breastscreen victoria dataset. Journal of medical imaging and radiation oncology, vol. 65, no. 5, pp. 529–537 (2021)
4. Giger, M. L., Chan, H.-P., Boone, J.: Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm. Medical physics, vol. 35, no. 12, pp. 5799–5820 (2008)
5. Giger, M. L., Karssemeijer, N., Schnabel, J. A.: Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. Annual review of biomedical engineering, vol. 15, pp. 327–357 (2013)
6. Harvey, H., Karpati, E., Khara, G., Korkinof, D., Ng, A., Austin, C., Rijken, T., Kecskemethy, P.: The role of deep learning in breast screening. Current Breast Cancer Reports, vol. 11, pp. 17–22 (2019)
7. Honig, E. L., Mullen, L. A., Amir, T., Alvin, M. D., Jones, M. K., Ambinder, E. B., Falomo, E. T., Harvey, S. C.: Factors impacting false positive recall in screening mammography. Academic radiology, vol. 26, no. 11, pp. 1505–1512 (2019)
8. Huynh, B. Q., Li, H., Giger, M. L.: Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of Medical Imaging, vol. 3, no. 3, pp. 034501–034501 (2016)
9. Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., Rubin, D. L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data, vol. 4, no. 1, pp. 1–9 (2017)
10. Mayo, R. C., Kent, D., Sen, L. C., Kapoor, M., Leung, J. W., Watanabe, A. T.: Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based cad. Journal of digital imaging, vol. 32, pp. 618–624 (2019)
11. Rojas-Domínguez, A., Puga, H., Rodríguez, M. O., Guerrero-Gasca, I.: Cad of breast cancer: A decade-long review of techniques for mammography analysis. Advances in Artificial Intelligence, vol. 115 (2020)

12. Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., Sieh, W.: Deep learning to improve breast cancer detection on screening mammography. Scientific reports, vol. 9, no. 1, pp. 12495 (2019)
13. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Févry, T., Katsnelson, J., Kim, E., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE transactions on medical imaging, vol. 39, no. 4, pp. 1184–1194 (2019)
14. Yu, X., Kang, C., Guttery, D. S., Kadry, S., Chen, Y., Zhang, Y.-D.: Resnet-scda-50 for breast abnormality classification. IEEE/ACM transactions on computational biology and bioinformatics, vol. 18, no. 1, pp. 94–102 (2020)