Three-blind validation of a Deep Learning model to obtain the dense tissue area in digital mammographies.

Francisco Javier Pérez-Benito [♭],[1] Andrés Larroza[♭], Juan Carlos Perez-Cortes[♭] and Rafael Llobet [♭]

(♭) Instituto Tecnológico de Informática, Universitat Politècnica de València.
Camì de Vera, S/N. 46022 València. Spain.

Abstract

Breast Cancer (BC) is the most frequent in women worldwide. Although mortality has significantly decreased in recent years, the incidence continues to increase, making screening programs crucial. In these programs, mammograms are collected from women over 40-45 years of age, which are then assessed by radiologists. The most relevant image biomarker of the risk of developing BC is the dense tissue [1], but its evaluation is time-consuming [2] and has a subjective component (inter and intra-observer) [3]. Approaches based on Machine Learning (ML) would reduce the analysis time; still, the subjective component challenges automatic methods to deal with the imperfect supervision problem. Moreover, different mammographic devices and acquisition protocols may influence the image and its analysis. In such a way, a conscious validation of automated technologies becomes critical.

We have trained an ML (Deep Learning) model, *CM-YNET* [4], to infer the dense tissue region in digital mammographies automatically. We used a dataset of images (1407 women, 2814 mammographies) from 11 centers ascribed to the Valencian screening program. Two experienced radiologists (*R1* and *R2*) segmented the dense tissue in each mammography. The model demonstrated robustness against image sources and promising performance results (overall *CM-YNET* DICE 0.82, overall *R1* vs. *R2* DICE 0.77). We also tested *CM-YNET* in a longitudinal ($2011 - 2021$) dataset of 320 mammographies from *Hospital del Mar (Barcelona)*, where acquisition devices were updated during this period, and then, the image quality changed. The same two radiologists segmented the mammographies. The results suggested the model generalization capability since the *CM-YNET* performance was comparable to the radiologists' (old devices, *R1* vs. *R2* DICE 0.71 and *CM-YNET* DICE 0.76; new devices, *R1* vs. *R2* DICE 0.76 and *CM-YNET* DICE 0.86).

After the extraction of a new set of 500 mammographies from Hospital del Mar, which were labeled by two radiologists (*L1* and *L2*), we developed the three-blind validation: (1) the images were also labeled by applying *CM-YNET (L3)* (obtaining 1500 labeled images), (2) from each labeler we randomly duplicated 100 segmented mammograms (1800 labeled images), (3) a third radiologist (R) assessed its agreement or disagreement with the 1800 labeled images randomly-showed. The experiment allowed us to measure the intraobserver variability of R, inconsistent in a total of 23/300 (7.67%) images, and check if *CM-YNET* behavior is as good as a specialist behavior for such a task (imperfect supervision). Preliminary analysis of the results shows that *CM-YNET* has an agreement with R (95.72%) as good as *L1* (96.50%) and *L2* (95.72%) have in the most current devices (those that are nowadays in clinical practice), meanwhile in older devices a slightly worse trend is observed (*L1* 72.42%, *L2* 73.66%, and *CM-YNET* 60.49%) due to the brightness of the images where the pectoral muscle appear.

[1] fjperez@iti.es

These results suggest that *CM-YNET* could be a powerful support decision tool for BC screening programs. Future work covers the application of preprocessing stages to remove undesired artifacts, such as pectoral muscle, and the repetition of the experiment to analyze the improvement of the agreement ratio observed in this study.

References

[1] Vachon, C. M., Brandt, K. R., Ghosh, K., et al., Mammographic breast density as a general marker of breast cancer risk. *Cancer Epidemiology Biomarkers & Prevention*, 16(1):43–49, 2007.

[2] Kuhl, C. K., The changing world of breast cancer: a radiologist's perspective. *Investigative Radiology*, 50(9):615, 2015.

[3] Pesce, K., Tajerian, M., Chico, M.J., et al., Interobserver and intraobserver variability in determining breast density according to the fifth edition of the BI_RADS Atlas. *Radiología*, 62(6):481–486, 2020.

[4] Larroza, A.,Pérez-Benito, F. J., Perez-Cortes, J-C., et al., Breast Dense Tissue Segmentation with Noisy Labels: A Hybrid Threshold-Based and Mask-Based Approach *Diagnostics*, 12(8): 1822, 2022.