Improving Breast MRI Workflow Through Deep Learning–Based Exam Prioritization

Fabrício J. Barth*, Almir G. V. Bitencourt†, Vinicius C. Felipe†, Raul Ikeda*, Rodrigo Patelli*, Tiago S. da Silva*
*INSPER, São Paulo, Brazil †AC Camargo Cancer Center, São Paulo, Brazil
{fabriciojb, RaulIGS, rodrigoap10, tiagoss4}@insper.edu.br {almir.bitencourt, vinicius.felipe}@accamargo.org.br

***Abstract*—This study proposes a deep learning–based framework for automated prioritization of breast magnetic resonance imaging (MRI) examinations aimed at optimizing radiological workflow. The approach is structured as a two-stage pipeline: (i) automatic breast localization in maximum intensity projection (MIP) images using a YOLO11 detector trained on 920 publicly available images; and (ii) classification of the extracted breast regions using a Vision Transformer (ViT-384) to identify examinations requiring clinical prioritization. The classification network was trained on a private dataset comprising 1,782 annotated examples. The ViT-384 model achieved a mean F1-score of 0.769 and a mean AUC-ROC of 0.805, outperforming a VGG19 baseline. These findings indicate that the proposed framework is a viable strategy for exam-level prioritization, with potential to enhance efficiency in high-volume clinical environments.**

***Index Terms*—Breast MRI, Deep Learning, Computer Vision, Exam Prioritization**

I. Introduction

Magnetic resonance imaging (MRI) plays a key role in breast cancer diagnosis due to its high tissue contrast. However, the analysis of these exams is time-consuming, and in high-demand settings radiologists face increasing workloads. Machine learning–based tools have shown promise in supporting image interpretation and improving clinical workflow efficiency [5], [6].

This work proposes a computational pipeline for automatic prioritization of breast MRI exams, comprising: (i) automatic breast detection in MIP images; and (ii) classification of the cropped breast regions to determine whether an exam requires clinical prioritization. Crucially, the classification task is *not* to diagnose cancer, but to flag exams that should be reviewed with higher urgency. In this context, an exam is considered prioritized when the radiologist identifies findings that require expedited clinical attention, such as suspicious enhancing lesions, or imaging patterns associated with high-risk clinical profiles.

II. Breast Detection Module

A YOLO11 model [7] was trained to localize breasts in MIP images using a dataset of 920 images assembled from three public sources: 678 annotated images from Roboflow [10], 208 additional breast images [2], and 34 images without identifiable breasts from the TCIA platform [3]. The dataset was split 70/15/15 for training, validation, and testing, with data augmentation (horizontal/vertical flipping and up to 5.21% noise) applied to the training set. The final model achieved a mAP50 of 0.989 and a mAP50-95 of 0.859. This module crops each detected breast individually, enabling the downstream classifier to operate on single-breast images.

III. Classification Pipeline

A total of 1,000 breast MRI examinations were retrospectively collected. For each case, expert radiologists provided structured annotations covering the presence of suspicious enhancing lesions, and an explicit prioritization flag per breast indicating whether prompt clinical review was needed. The binary training label was derived directly from this prioritization attribute. After converting exams to MIP format and cropping individual breasts, a dataset of 1,782 examples was obtained: 519 (29.1%) labeled as prioritized and 1,263 (70.9%) as not prioritized. Images were normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$) to support transfer learning. The training set was expanded threefold via data augmentation (horizontal/vertical mirroring and Gaussian noise at 1%, 2.5%, and 5% levels).

Two architectures were evaluated. VGG19 [9] is a 19-layer convolutional network operating on 224×224 images, fine-tuned with a deep classifier head including dropout, batch normalization, and residual connections. ViT-384 [4] is a Vision Transformer pre-trained on ImageNet at 384×384 resolution, paired with a lightweight classifier head. The higher input resolution of ViT-384 is particularly relevant for detecting subtle lesion characteristics in breast MRI. Both models were initialized from ImageNet weights and fine-tuned on the private dataset using the Adam optimizer, a learning rate of 10^{-4}, and binary cross-entropy loss.

IV. Results

Model performance was evaluated using stratified 5-fold cross-validation, training each fold for up to 50 epochs with early stopping based on validation F1-score. Class imbalance was addressed by comparing random under/oversampling, cost-sensitive learning, and the original distribution. No statistically significant differences were found across strategies.

ViT-384 outperformed VGG19 across all metrics. In terms of F1-score, ViT-384 reached 0.769 ± 0.024 versus 0.689 ± 0.037 for VGG19. For AUC-ROC, the scores were 0.805 ± 0.030 and 0.640 ± 0.008, respectively. Beyond higher average performance, ViT-384 also showed lower cross-fold variability, indicating more stable generalization. At the pipeline level, an examination is classified as prioritized if at least one of its cropped breast images is predicted as requiring prioritization.

V. Qualitative Interpretation of Results

Explainability techniques were applied to audit model decisions and guide dataset refinement. For VGG19, Grad-CAM [8] produced spatial activation maps highlighting the regions most influential to each prediction. For ViT-384, Attention Rollout [1] accumulated attention weights across transformer layers, revealing which image patches drove the classification. Figure 1 illustrates examples of these visualizations applied to validation samples.

To identify systematic failure modes, K-Means clustering was applied to the extracted visual features of misclassified examples. Two dominant error patterns emerged: (i) images with prominent vasculature led to false positives, as the model confused vessel structures with lesion-like features; and (ii) exams containing small or scattered nodules produced false negatives due to low lesion saliency. These findings guided the targeted acquisition of additional training examples matching each error profile, progressively improving model robustness.

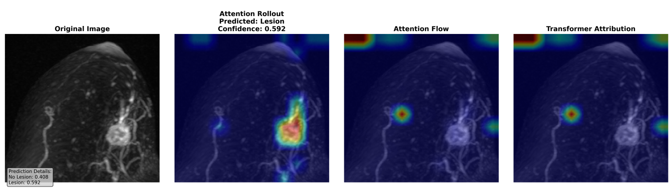

Fig. 1. Attention Rollout and related visualizations generated by ViT-384, showing which image regions most influenced the prioritization decision.

VI. Discussion

This study presents the development of a computational framework for the automatic prioritization of breast MRI examinations, integrating computer vision and machine learning techniques. The proposed solution was structured as a pipeline comprising two main components: (i) automatic breast detection in MIP images using a YOLO11 model trained on public datasets and validated on real-world images from a cancer institute; and (ii) classification of the cropped breast regions using a ViT-384 model trained on a private dataset of 1,782 samples to identify cases requiring clinical prioritization. The breast detection module achieved mAP50 of 0.989 and mAP50-95 of 0.859, whereas the prioritization module obtained an F1-score of 0.769 ± 0.024. As a decision rule applied at the end of the pipeline, an examination was classified as prioritized if at least one breast was predicted as requiring prioritization.

The data collection strategy initially relied on examinations from publicly available databases and was progressively refined throughout development based on model-driven analysis. The use of explainability techniques enabled the identification of predictive patterns, dataset gaps, and ambiguous cases, thereby informing the targeted selection of additional examinations. This iterative feedback process enhanced the representativeness of the dataset and consequently improved the model's performance and robustness.

Future work should focus on validating the proposed solution in a real-world clinical setting to assess its practical applicability and impact on the radiological workflow. Additionally, expanding the training dataset to include a larger number of examinations may lead to further improvements in model performance and generalization. Finally, although beyond the scope of the present study, future research could investigate the clinical utility of explanations generated by attention rollout, attention flow, and transformer attribution methods, evaluating their potential to support clinical interpretation and enhance confidence in the adoption of such tools in radiology practice.

Acknowledgment

The authors acknowledge the financial support from Philips under the Brazilian Informatics Law (Law No. 8.248/91 and updates). The authors also thank Renato Falcão, Eric Possato, Rafael Lima, Laura Pontiroli, Rafaela Oliveira, Leonardo Paloschi and João Sarti for their contributions to this project.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4190–4197, 2020.

[2] BreastDM. Breast1 dataset. https://universe.roboflow.com/breastdm/breast1, may 2024. visited on 2025-05-22.

[3] C. E. Comstock, C. Gatsonis, G. M. Newstead, B. S. Snyder, I. F. Gareen, J. T. Bergin, H. Rahbar, J. S. Sung, C. Jacobs, J. A. Harvey, M. H. Nicholson, R. C. Ward, J. Holt, A. Prather, K. D. Miller, M. D. Schnall, and C. K. & Kuhl. Abbreviated breast mri and digital tomosynthesis mammography in screening women with dense breasts (ea1141) the cancer imaging archive, 2023.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Lukas Hirsch, Yu Huang, Shaojun Luo, Carolina Rossi Saccarelli, Roberto Lo Gullo, Isaac Daimiel Naranjo, Almir G. V. Bitencourt, Natsuko Onishi, Eun Sook Ko, Doris Leithner, Daly Avendano, Sarah Eskreis-Winkler, Mary Hughes, Danny F. Martinez, Katja Pinker, Krishna Juluru, Amin E. El-Rowmeim, Pierre Elnajjar, Elizabeth A. Morris, Hernan A. Makse, Lucas C. Parra, and Elizabeth J. Sutton. Radiologist-level performance using deep learning for segmentation of breast cancers on MRI. *Radiology: Artificial Intelligence*, 4(1):e200231, 2021.

[6] Payam Jannatdoust, Parya Valizadeh, Nikoo Saeedi, Gelareh Valizadeh, Hanieh Mobarak Salari, Hamidreza Saligheh Rad, and Masoumeh Gity. Computer-aided detection (CADe) and segmentation methods for breast cancer using magnetic resonance imaging (MRI). *Journal of Magnetic Resonance Imaging*, 61(6):2376–2390, 2025.

[7] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[10] New Workspace. Breastcancer dataset. https://universe.roboflow.com/new-workspace-qysn1/breastcancer-bl1ys, oct 2022. visited on 2025-04-26.