# AI Explainability in Medical Imaging: A Case Study on Breast Cancer Detection

Inas Bachiri, Fethi Azibi, Abdelkrim Bennia, Meyssa Zouambi

## Abstract

Artificial intelligence has advanced medical imaging for improved diagnosis and treatment planning, notably in Alzheimer's classification [1], lung cancer detection [2], retinal disease detection [3], and breast cancer detection [4]. However, the black-box nature of many AI systems is a concern that is challenging AI deployment for clinical use. Clinicians need to understand and trust AI predictions, in line with regulations like GDPR [5] requiring the right to explanation for people subjected to automated decision-making. Various methods have been proposed to explain these models, but evaluating them remains a challenge. Through this work, we aim to highlight some explainability techniques and explore their use in medical imaging, particularly in the breast cancer detection case, as well as evaluate and compare their performance in explaining a model that has been known to work well for this task [4].

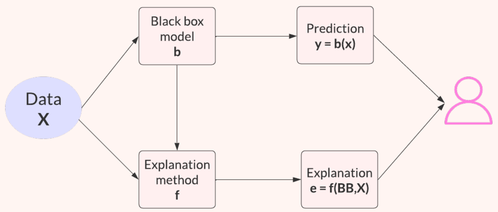
## Research Objectives

The study aims to achieve the following objectives:

- Explore the use of XAI methods to explain a model's breast cancer detection process in mammograms. [4, 6].
- Analyze the provided explanations and evaluate them.
- Compare these different approaches and draw conclusions about the future of AI in medical imaging.

## Explainability

Explainability, according to [7], refers to the details and reasons a model gives to make its functioning clear or easy to understand, given a certain audience.




There are two main types of explainable AI [7]:

- Models that are **transparent** by design (like linear regression and Bayesian Models).
- Models that are explained by a **Post-Hoc method**, i.e. those where the explanation is done on a block-box model, after designing and training it.

The latter type is explained by mainly two categories of methods [7]:

- **Model agnostic** methods, that work on any given model.
- **Model specific** methods, that are specifically designed for a particular set of models.

According to [8], given an XAI method, we can also classify it as :

- **Local**, if it explains a model's prediction on a single sample.
- **Global**, if it explains the model in general.

## XAI methods

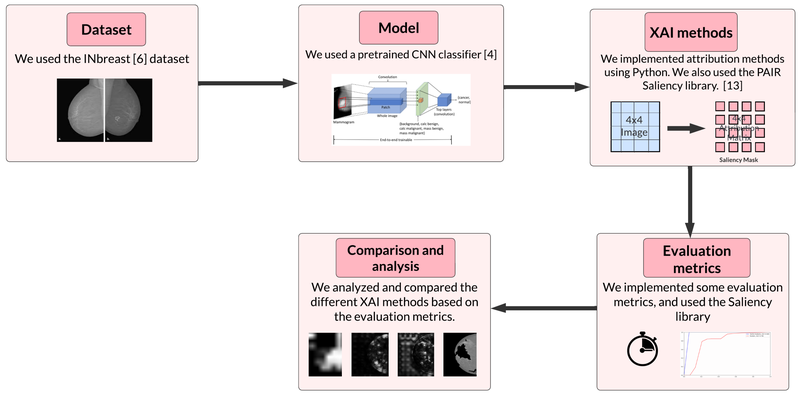
We compared the following attribution methods :

1. **Gradients** [9] involves analyzing the gradients of a model with respect to its input data to understand how specific input features influence the model's predictions.
2. **Integrated Gradients (IG)** [9] a local attribution method that produces a saliency mask that highlights the pixels in the image that contribute most to the model's prediction.
3. **XRAI** [10] a region-based attribution method that builds upon IG. It segments the input image into similar regions based on a similarity metric and then determines attribution scores for those regions.
4. **Grad-CAM** [11] works by examining the gradient information flowing through the last convolution layer of the network. It produces a localization heatmap highlighting the regions in an image most influential for the prediction.

## Evaluation metrics

- **Execution time:** Measures how fast each XAI method is at explaining the model's predictions.
- **Accuracy Information Curves (AICs) [11]:** Starts with a completely blurred image and gradually sharpens the image areas that are deemed important by a given saliency method. Gradually sharpening the image areas increases the information content of the image. We compare the XAI methods by measuring the model's accuracy at each blur level.
- **Softmax Information Curves (SICs) [11]:** Similar to the AICs, but measures the model's certainty of the prediction (target class probability) instead of the accuracy.
- **Sanity Checks[12]:** Help determine whether a saliency method's results meaningfully correspond to a model's learned parameters. It is based on two statistical randomization test comparing the natural experiment with an artificially randomized experiment: model parameters randomization and data randomization.

## Experimental work




## Comparison Results

| XAI method | Execution time (s) | Sanity | Correctly Predicted at % Unblur | Confidently* Predicted at % Unblur |
|---|---|---|---|---|
| Gradients | 32.49 | Yes | 75% | 95% |
| IG | 708.90 | Yes | 10% | 30% |
| XRAI | 5753.74 | No | 68% | 87% |
| GradCAM | 32.15 | Yes | 42% | 69% |

*Confidence $\equiv$ softmax value $> 70\%$ .

- **Sanity:** A method is considered 'sane' if it passes the randomization tests proposed by Sanity Checks.
- **Correctly Predicted at % Unblur:** Indicates how effective the salient region provided by the method is in helping the model to correctly predict the class.
- **Correctly Predicted at % Unblur:** Indicates how effective the salient region, as provided by the method, is in ensuring the model's confidence in its prediction of the correct class.

## Conclusion and Future Work

From the results we have observed, a necessary trade-off emerges between execution time and the quality of explanations provided by various XAI methods. Some of them generate explanations within seconds, but often these might attempt to rationalize a model with random behavior or on randomized data. Moreover, it is evident that certain XAI methods, previously successful in computer vision tasks, may not perform as effectively on atypical images such as medical imagery. This is due to their unique characteristics. However, it is worth noting that IG, despite its time-consuming nature, gives good insights into each pixel's contribution to the model's predictions.

Moving forward, several more experiments can be done. One promising direction is to explore variants of IG, as these may offer solutions to its limitations. Additionally, we should investigate the applicability of other XAI methods to medical imaging tasks. It is also imperative to explore more evaluation metrics, because measuring XAI methods' performance is a challenging task.

## References

[1] Taeho Jo, Kwangsik Nho, and Andrew J Saykin. "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data". In: *Frontiers in aging neuroscience* 11 (2019), p. 220.

[2] Kai-Lung Hua et al. "Computer-aided classification of lung nodules on computed tomography images via deep learning technique". In: *OncoTargets and therapy* (2015), pp. 2015–2022.

[3] Zhiguang Wang and Jianbo Yang. "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation". In: *Workshops at the thirty-second AAAI conference on artificial intelligence*. 2018.

[4] Li Shen et al. "Deep learning to improve breast cancer detection on screening mammography". In: *Scientific reports* 9.1 (2019), p. 12495.

[5] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI magazine* 38.3 (2017), pp. 50–57.

[6] Inês C Moreira et al. "Inbreast: toward a full-field digital mammographic database". In: *Academic radiology* 19.2 (2012), pp. 236–248.

[7] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[8] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. "Explainable deep learning models in medical image analysis". In: *Journal of imaging* 6.6 (2020), p. 52.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[10] Andrei Kapishnikov et al. "Xrai: Better attributions through regions". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4948–4957.

[11] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[12] Julius Adebayo et al. "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31 (2018).

[13] Daniel Smilkov et al. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017).

[14] PAIR-code. *Saliency*. Available at: `https://github.com/PAIR-code/saliency`. 2023. URL: `https://github.com/PAIR-code/saliency`.



This work is made possible by the support of School of AI Algiers, Ecole Nationale Supérieure d'Informatique, Algeria

fm_zouambi@esi.dz