Using Counterfactual Information for Breast Classification Diagnosis

Miguel Cardoso, Carlos Santiago, Jacinto C. Nascimento
Institute for Systems and Robotics, LARSyS, Instituto Superior Técnico, Portugal
miguel.lopes.cardoso@tecnico.ulisboa.pt

Abstract

The last radiology report by the Royal College of Radiologists has identified the pressure that radiologists are suffering due to excessive workloads levels. This is due to the availability of a growing number of images and a short time to provide the report, making the diagnosis a difficult process. This suggests that the visualization of the radiologist should be accompanied, somehow, by an automatic "explainable" process. In this paper, we give emphasis to the breast cancer as it is one of the most common types of cancer in women. Specifically, we deal with mammography images because it is a primary step to be accomplished in an early radiological breast diagnosis. Although machine learning models are being used in medical imaging, these models still struggle to provide enough interpretability to provide reliability in the decision-making process of the radiologist. In this work, we explore solutions that improve an explainable model's performance in mammography classification. We propose the use counterfactual information for improving the breast classification task. We compare multiple approaches to the integration of counterfactual information into the training process. The experimental evaluation testifies that incorporating such counterfactual information improves both balanced accuracy and interpretability for the breast classification task.

1. Introduction

The last radiology report by the Royal College of Radiologists [1] underscores the significant strain radiologists are under due to excessive workloads levels. These pressing issues not only adversely affect the working conditions of radiologists, but also lead to an unavoidable delay and inaccurate diagnoses, potentially resulting in poor treatment outcomes for patients. One of the reasons is that the utilization of medical imaging, particularly advanced imaging techniques, has grown considerably over the past two decades [8]. Various factors have contributed to this trend. The development of modern imaging technology is notable. It has improved image acquisition by reducing imaging times, and image quality has achieved better resolution with a decreasing radiation dose. The advances above, *i.e.* in the accessibility and accuracy of medical imaging enabled radiologists to obtain larger information records to conduct the diagnosis [20]. However, and interestingly, the short time-interval to obtain the diagnosis remains unchanged, contributing for an increase pressure in the radiological workflow. This suggests that, somehow, the visualization process of the radiologist should accompany with an *explainability* or *interpretability*, and thus facilitating the diagnosis.

Although deep learning (DL) based methodologies are being successfully used to diagnose medical images [14, 26], they still struggle to provide enough explanations, but even when the explanations are available, its origin must be carefully investigated.

In the ML or DL context, interpretability can be categorized into two paradigms: (i) as an inherently interpretable model, providing not only the result but also additional information about how the result was obtained as a part of their normal functioning [3, 4, 6] and (ii) to apply post-hoc explanation methods to analyze and explain the decisions of a trained or deployed black-box model [7, 16].

In either case, interpretability may uncover a hidden problem that DL models have with confounding factors, that is, using the *wrong* information to obtain the final and eventually the *correct* outcome [3]. As such, deep learning models often resort to alternative (confounding) details in the image, with no medical relevance, to predict the correct diagnosis. Concretely, it is possible to identify (and quantify) the use of confounding information [7], and devise strategies to prevent their use in the decision process.

In this paper, we give emphasis to the breast cancer as it is one of the most common types of cancer in women, and because it surpassed lung cancer as the most common type of cancer worldwide, with the highest incidence and second highest mortality rate [8]. Specifically, we deal with mammography images, since it is a primary diagnostic method used to diagnose breast neoplasia [11]. Inspired by [4], our methodology is based on a deep network having a built-in case-based reasoning process whose explanations are not created post-hoc but used during breast classification, instead. We explore solutions that improve the explainable model's performance in mammography classification, by proposing the use of annotations of regions with lesions to generate counterfactual information for use in training.

2. Related work

The black box nature of the deep neural networks puts radiologists off right to start, thus making the explainability an imperative issue to be included in the diagnosis. As mentioned in Sec. 1, explainability can be achieved in two ways. In *post-hoc* analysis there exist works that interpret a trained deep network by fitting explanations to perform the classification task. This includes works based on activation maximization [10, 15, 22], or saliency [13, 15, 17, 18], for deep visualization. Interpretability, can also be achieved with attention, *e.g.* [5, 24, 25], where the model's output highlight the image regions when making decisions. However, such models are unable to provide a *reasoning* about *which regions* they are looking at. Our proposal is inspired in the ProtoPNet [4] to surpass the limitation above. Concretely, this network provides the *reasoning* above by exposing parts (*i.e.* prototypes or cases) to which they focus on are similar. This somehow resembles the visualization process of the radiologists. For example, radiologists compare suspected tumors in X-ray scans with *prototypical* tumor images for diagnosis of cancer [21].

Counterfactual explanations (CFEs) - an emerging technique under the umbrella of ML interpretability models - have shown to help the model to focus on the correct pathways towards the final decision [12]. With the use of CFEs it is possible to overpass the limitations of DL models, namely, its susceptibility to learn spurious correlation [23] and amplifying biases [19]. These counterfactual examples, commonly used in causal inference, are generated by modifying known examples with specific interventions, leading to new examples with possibly alternative expected outputs. In this work, by using an inherently interpretable model and incorporating counterfactual data into the training process, we aim to prevent the use of confounding information to obtain the diagnosis. In this process, we improve not only the explanations given by the model but also its performance in the diagnosis of mammography images.

3. Proposed Approach

In this section we start by describing the learning process based on prototypes (Sec. 3.1), followed by its reformulation to account for binary breast classification problem, as well as the incorporation of counterfactual information (Sec. 3.2).

3.1. Prototypical Learning

In prototypical learning [4], it is assumed that the network will learn a set of prototypes $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_K\}$ in a dataset $\mathcal{D}$ with K classes, where $\mathbf{P}_k = \{\mathbf{p}_j^k\}_{j=1}^J$ is the set of J prototypes in the k-class, with $\mathbf{p}_j^k$ denoting the j-th prototype in the k-class, and where $k = \{1, ..., K\}$ indexes the classes in $\mathcal{D}$. The network consists of a tree-stage architecture, containing (i) convolution network c with parameters w_c, (ii) prototype layer p, and (iii) a fully connected layer f, with parameter w_f. See top branch in Fig. 1. The training process is also characterized by the following three stages:
1) Stochastic gradient descent (SGD): In this stage, the latent space of the most meaningful patches of the input image are learned. This is achieved by clustering the patches with semantically similar prototypes of the true classes. The clustering is achieved by computing the square distances between the j-th prototype, $\mathbf{p}_j^k$, of the k-class and the convolutional output $\mathbf{z} = c(\mathbf{x})$, obtaining the distance $d_j^k = \|\mathbf{z} - \mathbf{p}_j^k\|_2^2$. Inverting the distance d_j^k, this results in an activation map of similarity scores whose value indicates how strong a prototypical part $\mathbf{z}$ is present in the image. Formally, given the convolutional output $\mathbf{z}$, the prototype unit of the k-class, $p_{\mathbf{p}_j^k}$ computes

$$p_{\mathbf{p}_j^k} = \max_{\widetilde{\mathbf{z}} \in \text{patches}(\mathbf{z})} \log\big\{(d_j^k + 1)/(d_j^k + \epsilon)\big\} \qquad (1)$$

where $\epsilon = 10^{-4}$ is a regularization constant (in the experimental evaluation we set $\epsilon = 10^{-4}$).

From the above, we conclude that in the SGD process, the most meaningful patches $\mathbf{z}$, for the classification task, are grouped around semantically similar prototypes $\mathbf{p}_j^k$ of the images' true classes. However, such classes must be well-separated, that is, the clusters that are centered at prototypes from different classes must have well defined boundaries. This means that *clustering* and *separability* must be both optimized. This is achieved by jointly optimizing the parameters of the convolutional layers w_c and the prototypes $\mathbf{P}_k$, with $k = 1, ...K$, in the prototype layer $p_{\mathbf{P}}$, and keeping the parameters w_f freeze.

Assuming a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we aim to minimize the three-term loss function

$$\mathcal{L} = \min_{\mathcal{P}, w_c} \frac{1}{N} \min \sum_{i=1}^{N} \mathcal{L}_{\text{CE}}(\widehat{y}_i, y_i) + \lambda_1 \mathcal{L}_{\text{Clst}} + \lambda_2 \mathcal{L}_{\text{Sep}} \quad (2)$$

where $\widehat{y}_i = (f \circ p_{\mathbf{P}} \circ c(\mathbf{x}_i))$ is the predicted label, $\mathcal{L}_{\text{Clst}}$ and $\mathcal{L}_{\text{Sep}}$ stands for the *clustering* and *separability* term losses, respectively, and given by

$$\mathcal{L}_{\text{Clst}} = \frac{1}{N} \sum_{i=1}^{N} \min_{j:\mathbf{p}_j^{k=y_i}} \min_{\mathbf{z}_i \in \text{patches}(c(\mathbf{x}_i))} \|\mathbf{z}_i - \mathbf{p}_j^{y_i}\|_2^2 \quad (3)$$

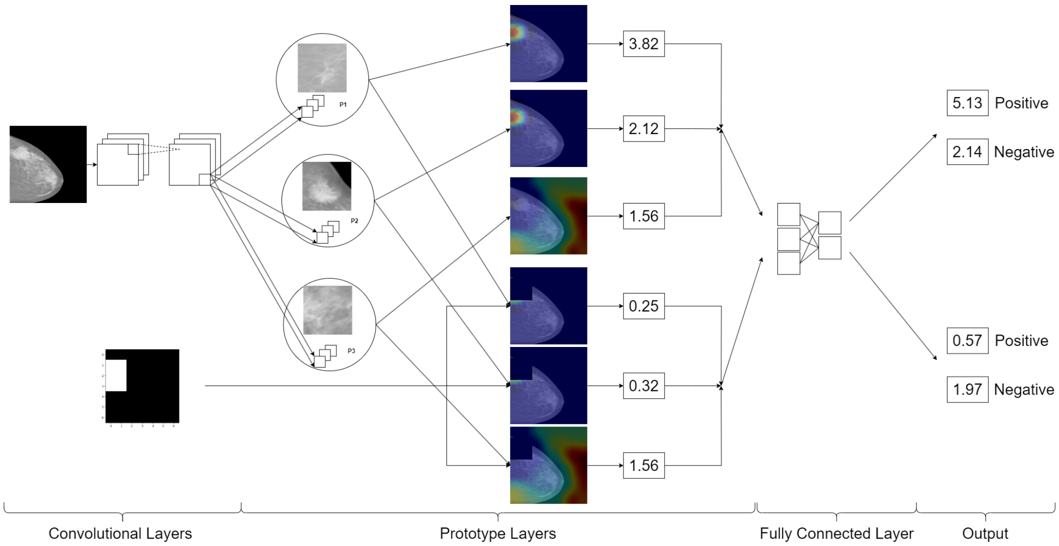

Figure 1. Diagram of the proposed model's architecture, with the classification of the whole image (top branch) and the counterfactual classification of the image without the regions of interest (bottom branch). In the network three main blocks can be seen: (i) convolution layers c with parameters w_c (left), (ii) prototype layers $p_{\mathbf{P}}$ (middle), and (iii) fully connected layers f with parameters w_f (right).

$$\mathcal{L}_{\text{Sep}} = \frac{1}{N}\sum_{i=1}^{N} \min_{j:\mathbf{p}_j^{k\neq y_i}} \; \min_{\mathbf{z}_i \in \text{patches}(c(\mathbf{x}_i))} \|\mathbf{z}_i - \mathbf{p}_j^k\|_2^2 \quad (4)$$

The minimization in (3) promotes each training image $\mathbf{x}_i$, (with label y_i) to have some latent patch $\mathbf{z}_i$ close to a prototype $\mathbf{p}_j^{y_i}$ for some $j \in \{1, ..., J\}$. The minimization in (4), promotes every latent patch $\mathbf{z}_i$ of a training image $\mathbf{x}_i$ (with label y_i) to stay away from the prototypes not of its own class y_i, that is, $\mathbf{p}_j^k$, such that $k \neq y_i$.

2) Prototype projection: This learning step aims essentially to visualize the prototypes as training images. This is performed through a projection of the prototype $\mathbf{p}_j^k$ onto the closest latent training patch $\mathbf{z}_i$, such that it belongs to the same class of the prototype $\mathbf{p}_j^k$. Formally, the update of the prototype is as follows

$$\mathbf{p}_j^k \leftarrow \arg \min_{\mathbf{z}_i \in \mathbf{Z}_j} \|\mathbf{z}_i - \mathbf{p}_j^k\|_2^2 \quad (5)$$

where $\mathbf{Z}_j = \{\mathbf{z}_i : \mathbf{z}_i \in \text{patches}(c(\mathbf{x}_i)), \text{such that} \quad k = y_i\}$.

3) Convex optimization in f layer: As it is mentioned in the SGD training stage, the w_f parameters are kept frozen. Denoting $w_f^{k,j}$ as the parameter weights associated to the connection between the prototype unit $p_{\mathbf{p}_j}$ and the logit class $k = c\ell$. Such freezing mechanism is straightforwardly achieved by fixing $w_f^{k,j} = 1$, $\forall j \in 1, ..., J$, with $\mathbf{p}_j^{k=c\ell}$ and $w_f^{k,j} = -0.5$, $\forall j \in 1, ..., J$, with $\mathbf{p}_j^{k\neq c\ell}$.

The goal of this stage is to provide sparsity to the final model, that is $w_f^{k,j} \approx 0$, instead of the initial value of -0.5. Such optimization is performed as follows

$$\min_{w_f} \frac{1}{N}\text{CE}(\widehat{y}_i, y_i) + \lambda_3 \sum_{k=1}^{K} \sum_{j:\mathbf{p}_j^{k\neq c\ell}} |w_f^{k,j}| \quad (6)$$

where $\widehat{y}_i = (f \circ p_{\mathbf{P}} \circ c(\mathbf{x}_i))$.

3.2. Counterfactual Learning

To achieve a proper binary task classification for the breast diagnosis, and to account for the counterfactual information in the prototype learning, the architecture network must be reformulated. Now, given an image $\mathbf{x}_i$ with the correspondent image label y_i (*i.e.*, $y_i = 1$ if $\mathbf{x}_i$ contains lesion(s), $y_i = 0$, otherwise), the model must have two outputs: (i) the predicted classification of the breast image $\widehat{y}_i$, and (ii) the *counterfactual classification* of the image $\mathbf{x_i}$, that is, the classification of the image $\mathbf{x}_i$, if the region(s) of interest (ROI) (*i.e.*, the region(s) containing the lesion(s)) were removed. Formally, let $\mathbf{x}_i : \Omega \mapsto \mathbb{R}$, with Ω the image lattice, the counterfactual classification function is given as:

$$\mathcal{F}_c : \Omega \setminus \mathcal{R} \mapsto \{0, 1\} \quad (7)$$

with the positive regions $\mathcal{R} = \{R_1, ..., R_n\} \in$ ROI in $\mathbf{x}_i$. We denote $\mathbf{x}_i'$ as the image $\mathbf{x}_i$ removing the positive region(s) in $\mathcal{R}$, and y_i' the corresponding label, that is, the output of the function $\mathcal{F}_c$ above. Particularly, in our case, the regions can be obtained by either using breast lesion detector (*i.e.*, the masses in X-ray) [2], or be defined manually.

In order to obtain a second classification y'_i, we introduce a second branch (see Fig. 1 bottom branch) taking advantage of the prototype layer $p_\mathbf{p}$. From Sec. 3.1, we see that for each convolutional output patch in c and for each prototype $\mathbf{p}^k_j$, the prototype layer has the following two stages: (1) computation of the distances d^k_j (recall (1)), and (2) an inversion step into similarity scores. Thus, we proceed as follows: first, we make a copy of the distances d^k_j from the top to the bottom branch before the stage (2) above. Second, we apply a mask M to remove the positive regions in $\mathcal{R}$, (*i.e.* the ROI), obtaining the *counterfactual distance*

$$d'^{k}_{j} = d^k_j \odot M \tag{8}$$

where $\odot$ represents the element-wise multiplication operator. The rationale behind (8) is that after inversion, the distance d'^{k}_{j} should be as minimal as possible not having an impact in the classification.

Notice that M is a binary mask containing the regions to be removed from the input image. This is illustrated in Fig. 1 (bottom-left), where it can be seen the white bounding box masking the mass lesion present in the image. With the modification above, we can then train the model using this second output as counterfactual in positive images.

To integrate the new counterfactual output into training, we modify the loss in (2). Specifically, we introduce the following three terms:

$$\begin{aligned} S1 &= \mathcal{L}_{\text{CE}}(\widehat{y}'_i, y'_i) \\ S2 &= \mathcal{L}_{\text{Clst}}(\mathbf{z}'_i, y'_i) \\ S3 &= \mathcal{L}_{\text{Sep}}(\mathbf{z}'_i, y'_i) \end{aligned} \tag{9}$$

where $\widehat{y}'_i$ are the model's counterfactual predictions, y'_i are conterfactual labels of $\mathbf{x}'_i$, (*i.e.*, the label of $\mathbf{x}_i$ after removing $\mathcal{R}$ from $\mathbf{x}_i$), and $\mathbf{z}'_i$ are the patches of convolutional output, *i.e.* $\mathbf{z}'_i \in \text{patches}(c(\mathbf{x}'_i))$.

For the strategy of "$\mathcal{J}$"oining the standard and counterfactual outputs, we can use $J1$, $J2$ and $J3$ as follows:

$$\begin{aligned} J1 &= \mathcal{L}_{\text{CE}}(\widehat{y}_i \oplus \widehat{y}'_i, y_i \oplus y'_i) \\ J2 &= \mathcal{L}_{\text{Clst}}(\mathbf{z}_i \oplus \mathbf{z}'_i, y_i \oplus y'_i) \\ J3 &= \mathcal{L}_{\text{Sep}}(\mathbf{z}_i \oplus \mathbf{z}'_i, y_i \oplus y'_i) \end{aligned} \tag{10}$$

where $\oplus$ represents concatenation operator. In (10), we aim to explore several combinations and evaluating the impact of the two branches of the network, for each cross-entropy, separability and clustering terms, as we thoroughly detail in Sec. 4.

4. Experimental Setup

For this work we used the INBreast dataset [9], which is a public benchmark breast dataset. It consists in 410 images in DICOM format with 3328 × 4084 or 2560 × 3328 pixels, depending on the breast size of the patient. The BI-RADS[1] labels for each image and segmentation masks for the masses are also provided. Before the data is given as input to the model for training, the images were downsampled to 224 × 224, while the segmentation masks resized to 7 × 7.

Table 1. Balanced accuracy and interpretability of different loss strategies with *fully augmented no lesion vs has lesion* dataset. The scores in bold highlights better performance against the baseline (ProtoPNet).

Model	B. acc.	Interpretability
ProtoPNet	0.55	0.12
ProtoPNet with S1	0.49	0.08
ProtoPNet with S1+S2	**0.72**	**0.20**
ProtoPNet with S1+S2+S3	0.53	0.06
ProtoPNet with J1+J2	**0.62**	**0.24**
ProtoPNet with J1+J2+J3	**0.61**	**0.37**
ProtoPNet with J1+J2+S3	**0.63**	**0.21**

We tested the model in two binary classification scenarios: (i) *no lesion* vs *has lesion*, and (ii) *benign* vs *malignant*. These are the most used scenarios in radiology. The first one, can be interpreted a selection or dividing positive cases from negative cases. The second one, corresponds to refine the positive cases into one of the benign or malignant classes. In the first scenario, the image is considered as *malignant* (positive class) if it has a BI-RADS > 3, otherwise it is *benign* (negative class). In the second scenario, only the images with a label of BI-RADS = 1 are considered *no lesion* (negative class).

Two different data augmentation processes were compared: *"full augmentation"* comprising horizontal flipping, random rotation (up to 15°), random skew (0.2 magnitude) and random shear (up to 10°); and *"semi augmentation"* with just the horizontal flipping and random rotation. These two different type of regularization processes come from the fact that, during the experiments, the accuracy performance depended on the transformation that had been used. Thus, these two types of regularization serve to better systematize the results and to discover what is the best way to perform the ProtoPNet regularization. Also, this allow us to compare the results and to discover if the "semi augmentation" process can achieve similar results to "full augmentation" as the smaller dataset would result in faster training of the network.

To evaluate the classifier's performance, two metrics were used, one for the accuracy of the model and other for the corresponding interpretability. For measuring the accuracy of the model we used the balanced accuracy, *i.e.* $\mathbf{B}.\text{acc}. = (\text{Sensivity} + \text{Specificity})/2$. This metric is

[1] **B**reast **I**maging-**R**eporting and **D**ata **S**ystem, is a quality assurance tool originally designed for use with mammography to grade the lesion severity.

Table 2. Balanced accuracy and interpretability of different loss strategies with *fully augmented benign vs malign dataset*. The scores in bold highlights better performance against the baseline (ProtoPNet).

Model	B. acc.	Interpretability
ProtoPNet	0.67	0.19
ProtoPNet with S1	**0.78**	**0.22**
ProtoPNet with S1+S2	**0.75**	**0.44**
ProtoPNet with S1+S2+S3	**0.72**	**0.31**
ProtoPNet with J1+J2	**0.73**	**0.35**
ProtoPNet with J1+J2+J3	**0.77**	**0.37**
ProtoPNet with J1+J2+S3	**0.72**	**0.46**

preferable over accuracy because the INbreast dataset is imbalanced. In concrete, the *benign* vs *malignant* case, we have 310 negative images and 100 positive and in the *no lesion* vs *has lesion* case, we have 67 negative images and 343 positive. For the model's interpretability, we measured how much of the activation of the positive prototypes was in the region of interest, since it quantifies how good that prototype is at detecting relevant features to the classification. The interpretability measure is given as

$$\text{Interpretability} = \frac{\sum_{j:\mathbf{p}_j^{k=y_i}} (M \cap A(\mathbf{p}_j^k))}{\sum_{j:\mathbf{p}_j^{k=y_i}} A(\mathbf{p}_j^k)}, \qquad (11)$$

where M is the binary mask for the ROI and $A(\mathbf{p}_j^k)$ is activation map of the the prototype $\mathbf{p}_j^k$ for a given class y_i of the image $\mathbf{x}_i$.

5. Results

We perform the experimental evaluation exploring diverse loss strategies evaluating the inclusion of the counterfactual information into each of the cross-entropy, separability and clustering terms as in eq. (10), and also, how the two different data augmentation strategies impact on both the classification scenarios (*i.e.* *no lesion* vs *has lesion*, and *benign* vs *malignant*).

First, we have evaluated the fully data augmentation for the two classification scenarios. Tables 1, 2 report these results. It can be seen that from the 12 experiments, 10 provided better results when compared to the baseline (*i.e.* without the use of counterfactual information), exhibiting an increased in both the balanced accuracy and interpretability. These experiments also show that the proposed solution is a viable way to reduce the use of confounding information, as shown in Fig. 2 and to reduce the accuracy gap between ProtoPNet and black-box models.

In the second set of experiments, we tested some of the best performing loss combinations of the previous set of experiments in the semi augmented dataset on both classification tasks. The obtained results are shown in Tabs. 3, 4.

Table 3. Balanced accuracy and interpretability of different loss strategies with *semi augmented no lesion vs has lesion dataset*. The scores in bold highlights better performance against the baseline (ProtoPNet).

Model	B. acc.	Interpretability
ProtoPNet	0.57	0.10
ProtoPNet with S1+S2	0.51	0.06
ProtoPNet with J1+J2+J3	0.53	**0.20**

Table 4. Balanced accuracy and interpretability of different loss strategies with *semi augmented benign vs malign* dataset. The scores in bold highlights better performance against the baseline (ProtoPNet).

Model	B. acc.	Interpretability
ProtoPNet	0.48	0.11
ProtoPNet with S1+S2	**0.67**	**0.16**
ProtoPNet with J1+J2+J3	**0.56**	0.08

In these experiments we are able to observe that, although the results are not massively better in the both metrics as in the previous experiment (in Tabs. 1, 2), we should highlight that in Tab. 3 the configuration "J1+J2+J3" doubles the interpretability performance regarding the baseline. Concerning the Tab. 4, we also observe that a significant improvement in the balance accuracy is achieved. Despite the obtained reduced training time with more restricted augmentation process, these results are important to highlight, as they help to design the regularization methodology to be adopted. In particular, the ProtoPNet requires extensive data augmentation for medical imaging applications.

6. Conclusions

The main goal of this work is to demonstrate that the incorporation of counterfactual information improves the classification accuracy in the radiological diagnosis. In particular, and from our experiments, we see that there is an improvement in both the balanced accuracy and interpretabilty in most of loss strategies, notably in the separate cross entropy and cluster losses, as well as all losses joined. Further work, will include more precise counterfactual information. The radiologist may be interested in removing benign lesion from the image, and concentrate the attention only in the presence of malignant lesions. It this way, can be beneficial the inclusion of a text prompt detailing and refining what is the best counterfactual information for the reliability of the diagnosis.

7. Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was *not* required.

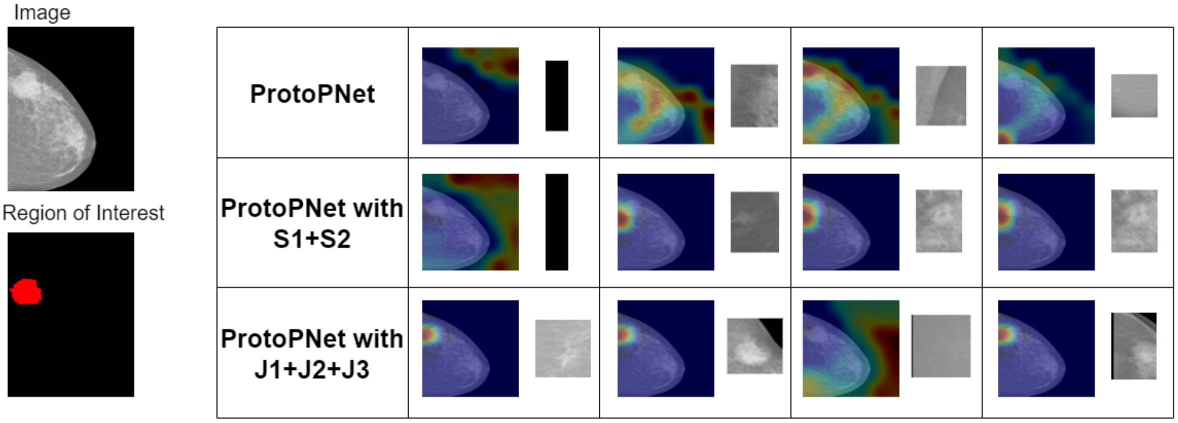

Figure 2. Qualitative comparison of an example test image in fully augmented benign vs malign dataset.

Acknowledgements

This work was supported by LARSyS funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020) and projects 10.54499/2022.07849.CEECIND/CP1713/CT0001, MIA-BREAST [10.54499/2022.04485.PTDC], PT SmartRetail [PRR - C645440011-00000062].

References

[1] Clinical radiology workforce census 2022. https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-workforce-census-2022, 2023. 1

[2] Richa Agarwal, Oliver Diaz, Xavier Lladó, Moi Hoon Yap, and Robert Martí. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3):031409–031409, 2019. 3

[3] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021. 1

[4] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 1, 2

[5] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017. 2

[6] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1

[7] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1

[8] Christine Dan Lantsman, Yiftach Barash, Eyal Klang, Larisa Guranda, Eli Konen, and Noam Tau. Trend in radiologist workload compared to number of admissions in the emergency department. *European Journal of Radiology*, 149: 110195, 2022. 1

[9] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012. 4

[10] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016. 2

[11] Luca Nicosia, Giulia Gnocchi, Ilaria Gorini, Massimo Venturini, Federico Fontana, Filippo Pesapane, Ida Abiuso, Anna Carla Bozzini, Maria Pizzamiglio, Antuono Latronico, et al. History of mammography: analysis of breast imaging diagnostic achievements over the last century. In *Healthcare*, page 1596. MDPI, 2023. 1

[12] Judea Pearl et al. Causality: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000. 2

[13] 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*, pages 618–626, 2017. 2

[14] Yiqiu Shen, Nan Wu, Jason Phang, Jungkyu Park, Kangning Liu, Sudarshini Tyagi, Laura Heacock, S Gene Kim, Linda Moy, Kyunghyun Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68: 101908, 2021. 1

[15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image

classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[16] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 519–528. Springer, 2021. 1

[17] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 2

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[20] James H Thrall. Trends and developments shaping the future of diagnostic medical imaging: 2015 annual oration in diagnostic radiology. *Radiology*, 279(3):660–666, 2016. 1

[21] Jeremy M et al. Wolfe. How do radiologists use the human search engine? *Radiation Protection Dosimetry*, 169(1-4): 24–31, 2016. 2

[22] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 2

[23] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017. 2

[24] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017. 2

[25] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 2

[26] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International conference on medical image computing and computer-assisted intervention*, pages 603–611. Springer, 2017. 1