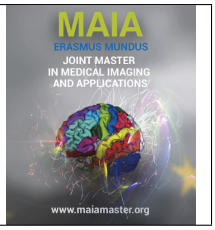
Medical Imaging and Applications

Master Thesis, June 2023



Deep Learning Explainability for Breast Cancer Detection in Mammography

Karla Guadalupe Sam Millan, Prof. Robert Martí

ViCOROB, Universitat de Girona

Abstract

Deep Learning's recent popularity in the field of Medical Imaging and CADx has generated concerns over the need to understand the decision-making process of the models, which has caused the development of different explainability methods (XAI). This study applied several of these XAI algorithms, namely saliency maps, Occlusion, Integrated Gradients, Guided GradCAM, LIME, SHAP, and DeepLIFT, to evaluate the performance of two CNNs trained on two different classification tasks: patch-based breast cancer classification and whole mammogram classification. The attribution maps obtained from the first task were qualitatively evaluated, and it was found that true positive predictions tended to highlight the lesion's area, while true negative predictions had more spread out highlighted regions. The attribution maps from the second task showed that the CNN highlighted the area of the mammogram where the lesion was located. The lesions were also highlighted in false negative and true positive mammograms with low probability scores, which implies that the model underwent correct training and learned the relevant features. Considering the idea that there should be a connection between explainability and the position of the lesions, IOU scores were computed to quantitatively evaluate the different XAI approaches. Integrated Gradients performed best at locating the lesions, while SHAP and LIME were the worst-performing.

Keywords: Deep Learning, explainability, XAI, mammography, attribution maps, breast cancer, classification, occlusion, SHAP, saliency maps, integrated gradients, DeepLIFT, Guided GradCAM

1. Introduction

Deep Learning has risen in popularity in recent years in the field of Computer Aided Diagnosis (CAD), with current state of the art implementing different deep learning models for solving a wide-range of tasks in medical imaging (Singh et al., 2020). One such task is breast cancer classification. As one of the leading causes of mortality amongst women, breast cancer has spurred significant interest in developing improved detection methodologies. Due to the vast amount of breast cancer cases, CAD systems seek to alleviate radiologists' workload, aiming to reduce work hours and improve detection accuracy (Balkenende et al., 2022). Convolutional Neural Networks (CNNs), in particular, have been extensively used for the purpose of classifying mammographies as benign or malignant (Loizidou et al., 2023). Arevalo et al. (2015) were among the first to use CNNs for classifying breast lesions, achieving an AUC value of 0.86 on the BCDR-F03 dataset. Since then, various other CNN architectures have been applied, like ResNet-50 and InceptionResNet-V2, where the latter achieved a 97.5% and 95.3% accuracy on the DDSM and INbreast datasets respectively in a study by Al-antari et al. (2020). Research findings have demonstrated that these CNN-models exhibit comparable performance to that of radiologists, and in certain instances, can increase the proficiency of radiologists when employed as a supplementary tool (Ou et al., 2021). Although DL methods have continuously been proven to perform adequately as CAD methods for Breast Cancer, there have been growing concerns as to the use of this technology. Deep Learning models are essentially considered as "black boxes", due to the sheer number of layers and weights inside, rendering it practically impossible to completely understand the inner workings and mechanisms of the neural network. This is critical in the medical field, where a decision made based on the results obtained from a neural network could have a huge direct impact on a patient's life. Moreover,

DL should comply with regulations like the European Union's General Data Protection Regulation (GDPR) (van der Velden et al., 2022). Explainable AI (XAI) aims to alleviate these previously stated concerns by developing various techniques that seek to understand deep learning algorithms. Some of these XAI methods include Saliency Maps, Integrated Gradients, Occlusion, Guided GradCAM, LIME, SHAP and DeepLIFT.

The objective of this work is to visualize and evaluate the performance of a trained classifier for 2D mammographies with the different XAI algorithms , and to compare the differences between them. The study was performed as follows:

1. Patch based XAI: To train standard CNNs on mamography with normal and malign patches, and assess the results of the different XAI methods on the model.
2. Whole mammogram XAI: To train models on full mammography images to classify between malignant and healthy images and assess the results with several XAI methods.
3. Evaluation of XAI methods: Qualitative and quantitative evaluation of the XAI attribution maps in terms of robustness and localization of findings with bounding boxes and IOU scores.

2. State of the art

2.0.1. XAI Methods

Saliency maps are considered as the baseline approach in XAI for medical imaging, and provide insights to the parts of an image that contributed the most towards a prediction from a neural network (van der Velden et al., 2022). They work by computing the gradients of the target class score with respect to the input. The output is a matrix of similar shape to the input image, where values close to 0 correspond to pixels which have a smaller impact on the output. High values, either positive or negative, on the other hand, correspond to pixels which majorly affect the output score (Simonyan et al., 2014).

Guided GradCAM is a point-wise multiplication between Guided Backpropagation and GradCAM, which makes it able to function with any type of CNN. It works by computing the gradient of the score for a given class with respect to a feature map, and it then applies global average pooling. Then, it obtains a weighted combination of the feature maps, and a RELU is subsequently applied to only acquire the features which influence positively the result. The resulting heatmap has the same size as the feature maps, and thus needs to be resized (Selvaraju et al., 2019).

Integrated Gradients compute the gradients of the output from the model for each step along a linear path between a baseline, which can be a blank image in the case of images, and the input. The gradients are then integrated, which accumulates the changes as the input goes from the baseline to the input image. A characteristic of this method is that it doesn't need to modify the network (Sundararajan et al., 2017).

Occlusion perturbs the image by applying a gray sliding window along the image, noting the changes in the output for each window position. Thus, when the correct class is occluded by the window, the output probability is expected to drop significantly (Zeiler and Fergus, 2013).

Like Occlusion, LIME (Local Interpretable Model Agnostic Explanations) perturbs the image and computes its corresponding output score for each of the perturbed images. This perturbation occurs at a pixel or superpixel level. The latter is done by providing LIME with a segmentation mask that divides the image into different regions. Then, a simpler interpretable model is trained on the perturbed images and their predictions, in order to learn the relationship between the changes to the original image and the original model's predictions. The higher the weights assigned to a pixel or superpixel, the higher its impact on the prediction (Ribeiro et al., 2016).

Based on cooperative game theory, SHapley Additive exPlanations (SHAP) uses Shapley values to compute the magnitude of the contribution of each feature to the output of the model. It achieves this by perturbing the image and altering the pixels (or superpixels). To compute the contributions, SHAP considers the different subsets of features and calculates the output with each given subset until it considers all possible combinations. However, this is computationally intensive, which has spawned some variations to approximate Shapley Values, such as Kernel SHAP and Deep SHAP (van der Velden et al., 2022)

DeepLIFT addresses the saturation problem by introducing a reference input and its reference activations in the network. The reference activations are compared to new activations to compute the activation changes and, therefore, the contributions of each neuron (de Vries et al., 2023) (Shrikumar et al., 2019).

The following section will list some of the state of the art applications of XAI in Mammographies.

2.0.2. XAI in Mammography

XAI methods have been used in the medical imaging field used to alleviate the concerns surrounding the black-box paradigm, as previously stated. Several visual explainability approaches have been used to study and understand a wide range of image modalities from various anatomical locations, like: brain, breast, cardiovascular, chest, prostate, eye, skin, among others (van der Velden et al., 2022). In the case of breast applications, researchers have obtained visual explanations for X-rays, MRI, ultrasound, and histological images.

Regarding XAI in 2D mammographies specifically, XAI has been used to evaluate the performance of the

networks on various problems. Huang et al. (2020) proposed a hybrid neural network comprised by two parts: a modified PCANet and a DenseNet. They compared their proposed HybridNet with other popular models like PCANet, ResNet and DenseNet, and found that HybridNet outperformed all of them. To confirm its correct performance, CAM was applied, and it was observed that the resulting attribution maps focused on the abnormal parts (i.e. the lesions) of the mammographies, indicating that HybridNet had successfully learned the important features for the classification problem.

Akserlod-Ballin et al. (2019) developed an ML-DL model that combined mammographies and clinical data to detect breast cancer, resulting in an AUC of 0.91, a specificity of 77.3% and a sensitivity of 87%. The combination with clinical data provided a level of interpretability to the model, and this was further explored by computing the impact of each of the data's features via SHAP.

Xi et al. (2020) tackled the problem of having high-resolution mammographies with meaningful information located in very small regions of the image. Instead of resizing the full images, which entails a loss of information, they trained several CNNs like AlexNet, VGGNet, GoogLeNet, and ResNet with the cropped ROIs, and and then applied created abnormality detectors by integrating the CNNs with either CAM or other region proposal networks. ResNet was the network selected for integration with CAM, and found that the detection results obtained from the heatmap aligned with the ground truth for the lesion localizations.

A paper by Kobayashi et al. (2022) utilized generative contribution mapping (GCM). GCM is a classification model proposed by Arai and Nagao (2017) that uses XAI to explain its classification predictions by creating class contribution maps and class weight maps. Kobayashi et al. used this method to classify the existence of calcifications on mammographies. They found that GCM was more efficiently explainable when combined with class contribution maps and class weight maps than with GradCAM. It also found that GCM, when used with the maps, could provide important visual information even in the case of a false negative, due to the highlighting of the microcalcification localizations even with an incorrect prediction.

Yi et al. (2019) aimed to develop a CNN to classify mammography images according the view, the breast laterality and the breast density. The model architecture selected for the three tasks was ResNet-50, modifying its last layer and assigning it either two or four output neurons, according to the task. CAM was applied to visualize the network's decision when predicting for each of the three objectives. The model showed an AUC of 1 for mammographic view and 0.93 laterality classification, but it displayed a 68% accuracy when classifying breast tissue density. CAM's heatmaps displayed the network's focus on the superior part of the image generally corresponding to the pectoral muscle for the mammographic view task. As for the laterality task, the heatmap highlighted the region located towards the left or right, depending on the laterality. Even though the third task achieved such a low accuracy, the heatmaps consistently highlighted the regions corresponding to the breast, even if the network predicted an incorrect breast density, indicating that it correctly based its decision on the breast tissue.

Prodan et al. (2023) developed both CNN and Visual Transformers for breast cancer detection. They employed a data augmentation technique involving synthetic images to reach better performance. After training the models, they made use of GradCAM and bounding boxes to gain insight into their models' decision making procedure and behavior.

3. Material and methods

3.1. Dataset

The dataset used for the breast patches classification task is the Iceberg Selection, a subset of the OPTIMAM Mammography Image Database (OMI-DB) (Halling-Brown et al., 2020), consisting of patches centered on breast lesions and patches with normal breast tissue. There are a total of 3808 full-images acquired from three different scanners: Hologic, Siemens, and GE, with each image having a patch centered around the lesion, and a normal patch, thus yielding a total number of 7616 image patches. The dataset was divided into training (80%) and validation (20%) subsets, ensuring no overlap of patients between them to avoid bias.

For the full-mammography classification task, two datasets were used. The first dataset is a more balanced subset of the training data of the RSNA22 challenge (Carr et al., 2022). It has a total of 2767 images, of which 1647 were negative cases, and 562 were positive (malignant) cases. The images in the dataset were all preprocessed, flipping the images to have the same laterality (left), cropping the background, and saving them as PNG files. Like in the previous task, the dataset was divided into 80% training and 20% validation, avoiding patient overlap. The second dataset contained the malignant full-images from the OMI-DB subset described previously that were acquired with the Hologic scanner, due to the higher similarity of the images from the RSNA22 dataset in comparison to those acquired with the Siemens or GE scanners. As for the benign images, an equal amount as the malignant images were selected from the OMI-DB dataset, which were also subsequently flipped, cropped and saved as PNG files. Thus, the total number was 7229 full-images. Training and validation were divided 80%:20% with no patient overlap. Table 1 summarizes the datasets used.

		Train	Validation	Total
OMIDB Subset (Iceberg Selection)	**Non-malignant**	3045	763	3808
	Malignant	3045	763	3808
	Total	6090	1526	**7616**
RSNA22 Subset	**Non-malignant**	1647	411	2058
	Malignant	562	147	709
	Total	2209	558	**2767**
OMIDB Hologic Subset (Full-Images)	**Non-malignant**	2896	719	3615
	Malignant	2892	722	3614
	Total	5788	1441	**7229**

Table 1: Datasets used for the breast classification tasks.

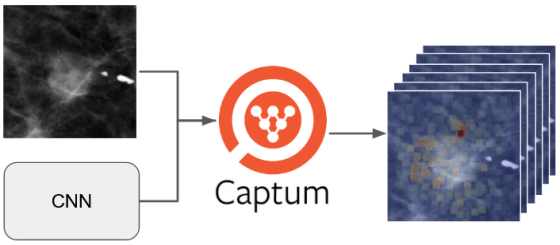


Figure 1: General diagram showing the generation of the heatmaps with Captum. Each of the explainability methods takes the CNN (either the ResNet-50 or EfficientNetB0) and an image (patch or whole mammogram) as inputs, and generates an attribution map containing the contribution scores for each pixel or superpixel.

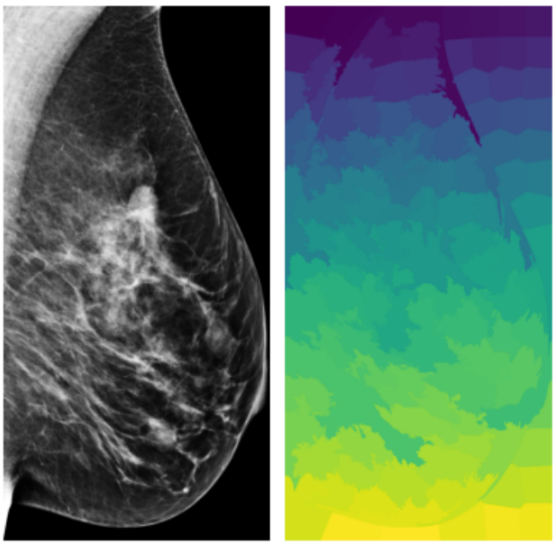


Figure 2: An example of segmentation performed with Scikit-Image's SLIC. The segmented image is fed as a feature mask for the generation of SHAP and LIME's attribution maps.

3.2. Methods

3.2.1. Preprocessing

The whole mammograms in from the Iceberg Selection were previously flipped, cropped and saved as PNG files. In order to create the Normal + Malignant OMI-DB Hologic database, the normal images were preprocessed in the same way. The DICOM files were read, thresholded, and a bounding box was computed with OpenCV's ConnectedComponents and findContours. The images were then cropped according to the computed bounding box to remove the background. Finally, the mammograms were saved as PNG files.

3.2.2. Breast Patches Classification

MobileNetV2 and ResNet-50 are two popular CNN architectures, and were therefore selected for this problem. Both were trained on the Iceberg Selection Dataset, with the train set image transformations consisting of a horizontal flip, a vertical flip, and a random rotation of 30 degrees for data augmentation, and a 224x224 resizing and normalization. The validation set transformations consisted only of the 224x224 resizing and normalization. The loss used was Cross Entropy Loss, the optimizer was Adam with a learning rate of 0.001, and a ReduceLROnPlateau scheduler with a patience equal to 5.

Out of both of them, ResNet-50 performed the best, with an accuracy of 97% compared to MobileNetV2's 93%. ResNet-50's high performance was in line with what was expected, as the classification problem was relatively simple due to the very different appearances of a normal patch vs one with a malignant lesion.

3.2.3. Full-Image Classification

Due to ResNet-50's good performance in the previous problem, several experiments were initially performed with it for Full-Image Classification with the RSNA22 subset dataset. Since it was an unbalanced problem with around three times the amount of negative cases vs positive cases, the weight of the positive examples was set to 3. Several attempts with varying learning rate values were made. However, ResNet-50 failed to yield satisfactory results.

The next network attempted was EfficientNetB0. Cantone et al. used this architecture, among others, for classifying whole mammograms. They used SGD with a momentum of 0.9 and a Cosine Annealing scheduler with warm restart, and varied the input image resolution Cantone et al. (2023). Thus, the same hyperparameters were tested in this study, in addition to the focal loss. The selected learning rate was set to 0.01, and the input size was set to 1024x512, as higher resolutions did not perform significantly better and its computational costs were substantially higher. This approach with the RSNA22 subset reached an AUC of 0.72.

Aiming to further improve the performance of the network, EfficientNet-B0 was then trained on the Normal + Malignant OMI-DB Hologic subset previously described. The hyperparameters selected were Cross Entropy Loss, SGD with momentum equal to 0.9 and a learning rate of 0.01, and Cosine Annealing with warm restart as a scheduler. The input size was kept at 1024x512. The AUC reached was higher than in the previous step, with a value of 0.83.

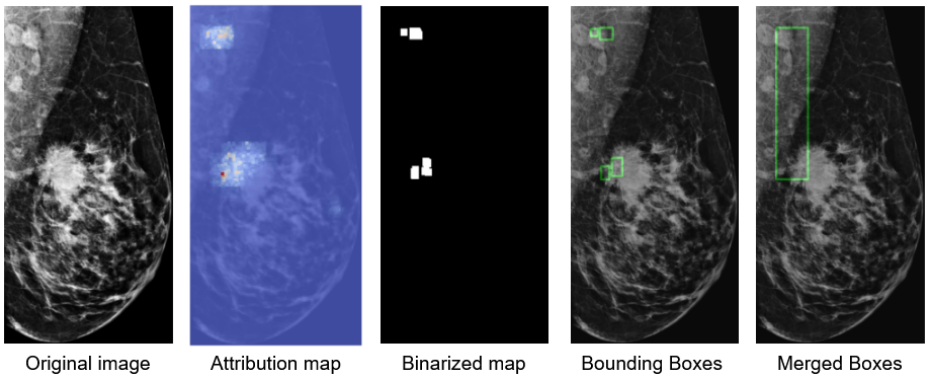


Figure 3: Steps to acquire the bounding boxes from the attribution maps. The attribution maps are obtained with a XAI technique, and they are subsequently binarized by keeping only the pixels with a score above a selected quantile, which are then cleaned by keeping only the larger regions via morphological operations. The bounding boxes are then acquired with the contours of the binarized maps, and they are finally merged into a large bounding box.

3.2.4. Explainability

Once the ResNet-50 and EfficientNetB0 models were trained and selected for each of the tasks, several explainability methods were applied on them to observe the inner workings and performance of the models. Captum for PyTorch is an open-source library that includes many explainability methods (Kokhlikyan et al., 2020). As such, the following attribution methods were computed utilizing this library. Fig. 1 summarizes the generation of the attribution maps by evaluating the model's contributions with an input image.

Computing the vanilla saliency maps and the DeepLIFT attribution maps require solely the trained model and the target class for which the gradients are computed. The attribution maps obtained initially with both of these methods are hard to visualize, given the large size of the input image and the small highlighted regions. To solve this, the attribution maps were dilated with a rectangular kernel of size 9x9 in a single iteration.

Integrated gradients, like the previous two, require only the target class. The number of steps performed by the approximation method for the integrals was set to 200. This was selected as a compromise between computational cost and attribution map resolution. As with vanilla saliency and DeepLIFT, the acquired maps were dilated for better viewing.

Guided GradCAM requires the specific layer for which the attributions are to be computed. The last layers for both models were specified and the attribution maps were also dilated.

As a perturbation-based approach, Occlusion required the shape of the patch with which to occlude the input image, and, optionally, the strides that the patch should take in each direction after every iteration. The sliding window shape was set to (3, 60, 60) and the strides to (3, 30, 30). The relatively large window size and strides were a compromise for the large computational times and the resolution of the attribution maps.

Captum's LIME approach takes in an optional feature mask argument, which groups the image's pixels into superpixels, and treats each group as a single interpretable feature. If a feature mask is not provided, then LIME considers each pixel as an individual interpretable feature, which largely increases their number, resulting in very slow attribution map computations. Thus, the feature mask is obtained by dividing the input images into 150 segments using Scikit-Image's SLIC method. Fig. 2 shows an example of a mammogram segmented with SLIC. Afterwards, the feature mask is fed into LIME with a number of steps equal to 200. The resulting attribution maps were not dilated, as the superpixels are easily observable. SHAP, much like LIME, was fed the same feature mask, on account of the same reasons. The attributions were not dilated either because of the superpixel grouping. An example of a feature mask obtained with SLIC can be observed in Fig.2.

3.2.5. Quantitative Evaluation: IOU

Using OMI-DB's subset that solely includes images with lesions, it was hypothesized that the explainability results should show a certain relationship with the position of the lesions. To this effect, and to evaluate the different explainability methods' attributions in the whole mammogram classification task, the Intersection Over Union (IOU) was calculated with respect to the ground truth bounding boxes available for the malignant full-images from the OMI-DB subset. Bounding boxes generated from the attribution maps were therefore needed. To achieve this, the following steps were taken:

1. The dilated attribution maps were selected (except for LIME and SHAP) because they generated larger connected regions corresponding to the mass' location which were not as affected by the morphological operations from the next steps. These dilated attribution maps were binarized by thresholding them according to a given quantile. If

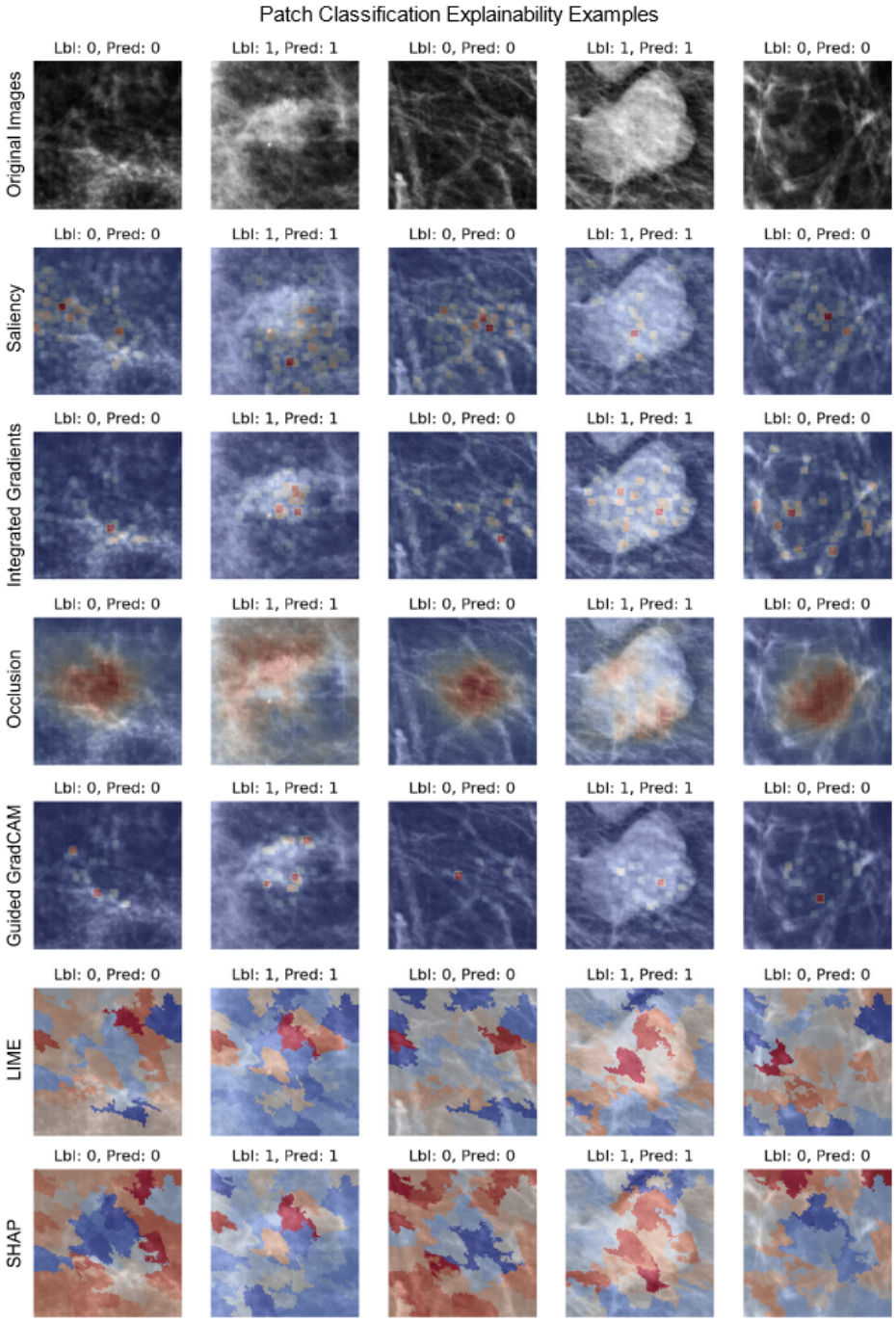


Figure 4: Examples of mammography patches that are either non-malignant or malignant. The top row corresponds to the original images, and the following rows show the attribution maps generated with a different explainability technique. Red areas correspond to higher attribution scores and higher impact on the model's prediction, while blue areas represent low scores.

the attribution score for a pixel was higher than that quantile, the score would be set to 1. Otherwise, it was set to 0.

2. The binarized attributions were eroded and then dilated, to remove the very small regions.
3. The bounding boxes for each separate region were obtained by finding their contours and generating a box for each contour.
4. In the case of multiple bounding boxes in a single image, they were combined into a single big bounding box with the minimum and maximum xy coordinate values found among the multiple boxes and setting them as the upper left and lower right coordinates for the combined bounding box, respectively.

5. Finally, the IOU scores for each image for each of the explainability methods were computed.

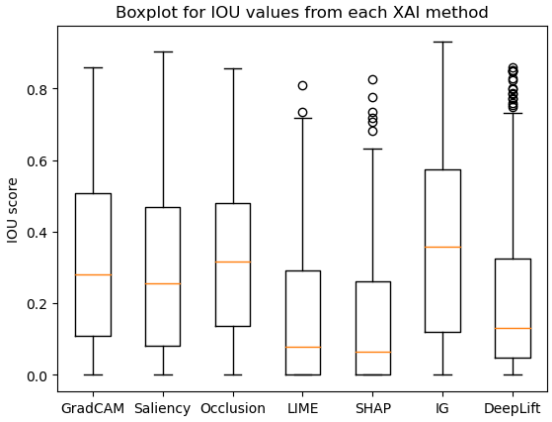


Figure 5: Boxplots computed with the IOU scores of the bounding boxes generated with the attribution maps of all the images. Higher IOU scores indicate a better overlap with the ground truth boxplots for lesion localization.

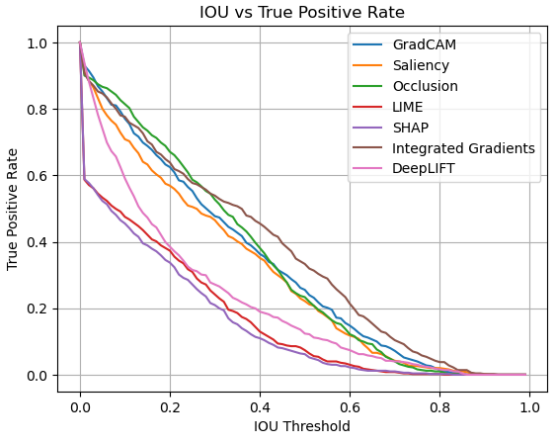


Figure 6: IOU vs TPR curves.

Fig. 3 illustrates each of the steps to generate the bounding boxes from the attribution maps

4. Results

4.1. Patch Classification

Attribution maps for several instances of the validation dataset were acquired in order to observe ResNet-50's performance and attempt to visualize its learned features. The randomly selected images indicate the prediction and the label, and each attribution map shows the attribution scores for each region of the image. The higher the score, the greater the impact of that region in the probability score for the target class. The attribution maps are color coded such that the redder a certain pixel, the higher its importance for the prediction, and the bluer, the lower its importance.

Fig. 4 shows the five original images plus their attribution map computed with each explainability method: Saliency, Integrated Gradients, Occlusion, Guided GradCAM, LIME, and SHAP. The maps differ greatly from one another, although some of them do seem to focus on areas that correspond to the lesion in the true positive cases. The true negative examples, however, visually vary more among themselves. On many cases, the focus of the model seems to be on regions which appear to be denser.

4.2. Full Image Classification

As with explainability for the patch classification task, the attribution maps for several images from the validation set were acquired. The true positive images with the highest probability scores were retrieved and their corresponding attribution maps for each of the methods were obtained. Figures 8 and 9 show the five images with the highest probability scores for malignancy, as well as the attribution maps generated with the explainability methods. All of the methods seem to highlight the areas corresponding to the lesions. Vanilla saliency appears to highlight areas outside the lesions more than the rest of the methods, but it still focuses mostly on the lesion.

The bounding boxes for each attribution map for each image per method were also generated. Fig. 10 and Fig. 11 show the corresponding bounding boxes generated from the attribution maps from Figs. 8 & 9, as well as the ground truth bounding boxes.

Examples showing the images with the lowest true positive scores and false negative scores were also generated, and are added in the appendix. In the case of the lowest true positive score examples, even though the network was not very confident at classifying the images, it does highlight the area corresponding to the lesion in most of the cases. The highest false negative score examples displayed this same behavior, focusing in the lesion area or highlighting it along with other regions in many cases.

Once having generated these examples for visual appraisal, the IOU for all the validation images for all the methods were computed as described in the previous section. Bounding boxes were computed to compare the overall IOU scores among the different explainability techniques, and are shown in Fig. 5. The best performing XAI techniques in terms of IOU scores were Integrated Gradients and Occlusion, while LIME and SHAP were found to be the lowest performing ones.

An IOU threshold vs True Positive Rate graph was also generated for further comparison of the seven different methods by varying the IOU thresholds. A TP would mean the bounding boxes from the attribution maps overlap significantly with the ground truth, and are therefore able to locate the lesions. The graph is shown in Fig. 6. As before, the best curves correspond to Integrated Gradients and Occlusion, while the worst belong to LIME and SHAP.

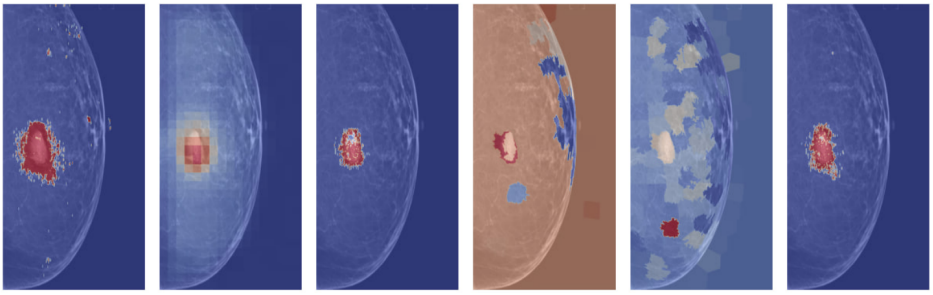


Figure 7: Attribution maps for the synthetic lesion generated over a real mammogram via stable diffusion.

4.3. Synthetic Lesions with Stable Diffusion

We have developed an additional experiment to further test attribution maps in synthetically generated lesions in normal mammograms. This experiment has been done in collaboration with Montoya (2023), whose master thesis work focused on the generation of high-resolution synthethic mammographies with diffusion models, and is featured in this year's proceedings. A sample image of a real mammogram with a synthetic lesion generated during his work was selected to extract attribution information with some explainability methods, as shown in Fig. 7. Most of the methods assign the synthetic lesion's area with the highest attribution scores. SHAP was the only method where this did not happen, although the synthetic lesion was counted as being of moderate impact.

5. Discussion

In this study, several explainability methods were tested on two networks, each trained on one of two breast cancer classification tasks. The explainability maps generated from the patch classifier exhibit an apparent overlap with the area of the patch corresponding to the mass, in the case of patches with masses. In general, the attribution maps seem to highlight the denser areas of the patches. The attribution maps look quite different from one another. This was to be expected for non-malignant patches, as, with the absence of a lesion, there is nothing that the classifier should focus on in particular. For the patches containing a lesion, it was observed that most of the attribution methods higher scores were often positioned in the middle of the patch. All in all, patches with no lesions seem to indicate that the network's attention seems to be dispersed throughout the image, whereas lesion patches' attribution maps mostly coincide with the lesion's position. This suggests that the network apparently learned to identify the masses properly, although it is difficult to ascertain when many lesions span almost the entirety or the majority of the patch, thus generating rather disperse attribution maps. Since the dataset consisted of patches centered around the lesion, this behavior is reasonable, as almost the entire image is visually different from normal patches and most of the the image could provide important information.

In the whole mammogram classification task, the attribution maps for all of the methods were successfully acquired for the validation dataset. When compared to the attribution maps from the previous task, the focus on specific regions of the image was much more apparent.

In true positive images, the highlighted regions mostly coincided with the position of the lesion in all the methods. Vanilla saliency presented more disperse attribution maps, presenting generally more clusters of highlighted pixels than the rest. This could be cause by vanilla saliency not being class discriminative, thus highlighting areas that contribute negatively towards the malignant target class. SHAP and LIME, for their part, although correctly assigning the highest attribution scores to the superpixels that overlapped with the lesion, occasionally allocated relatively high scores to superpixels that were positioned elsewhere in the image. A reason for this could be that those superpixels overlapped with several pixels that got high attribution scores, which, by themselves, would not be as noticeable, but by congregating in a superpixel the latter's attribution scores would rise and be more visually apparent across a larger area.

The bounding boxes generated from the attribution maps were overall able to center on the region indicated by the ground truth, although the large majority of them were quite bigger than their ground truth counterparts. This was expected, as they were acquired from the previously dilated attribution maps, which enlarged the highlighted regions. Integrated gradients achieved the highest scores, but generating their attribution maps was the computationally expensive and took the longest out of the seven methods. In contrast, even though Occlusion and GradCAM did not achieve IOU scores as high as Integrated Gradients, they were much faster, with GradCAM's attribution maps generating almost instantly. LIME and SHAP's much lower IOU scores were mostly caused by their bounding boxes being affected by the image's initial segmentation. Segmenting the mammogram into smaller regions could possibly help mitigate this, although it would increase the computationally resources when acquiring the attribu-

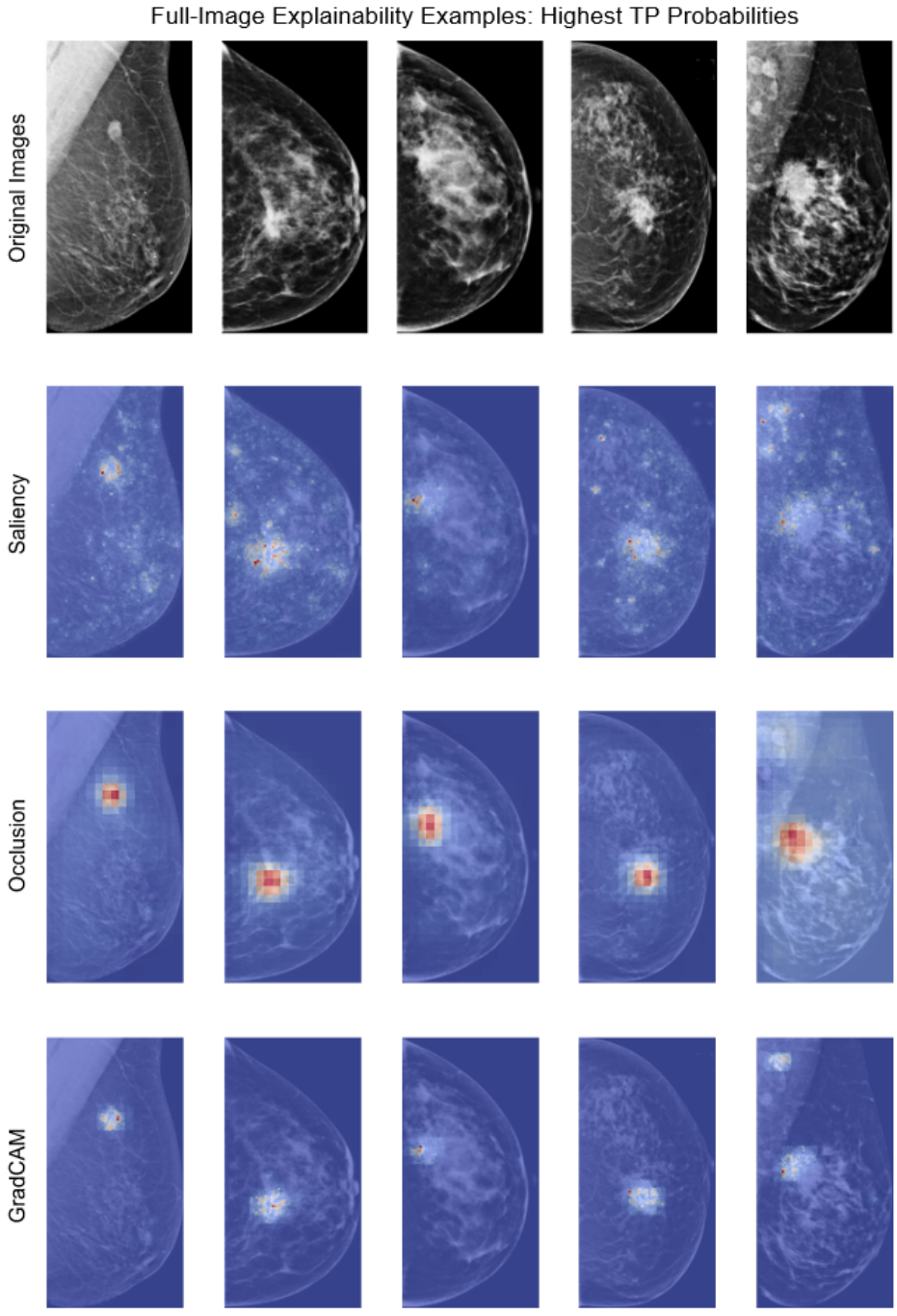


Figure 8: Examples for Saliency, Occlusion, and GradCAMs' attribution maps on images with high probability scores. Red and yellow regions correspond to higher attribution scores.

tion scores, as LIME and SHAP would have now many more image segments to consider.

True positive images with low probability scores and false negative images generated attribution maps where most of them contained the lesion area among its highlighted regions, indicating that while the model was not as confident classifying it as a malignant image or even misclassified it altogether, it was still focusing in the lesion. This supports the idea that the network successfully learned the masses' features and effectively bases its decision on the relevant data. The bounding boxes in this cases were expectedly not as accurate as the ones generated from the true positive images with higher probability scores, as even though the mass' position was highlighted, in many instances this was only one of several highlighted regions, which effectively created more bounding boxes throughout the mammogram, This resulted in a much bigger final bounding box when combining them and in predictably much lower IOU scores.

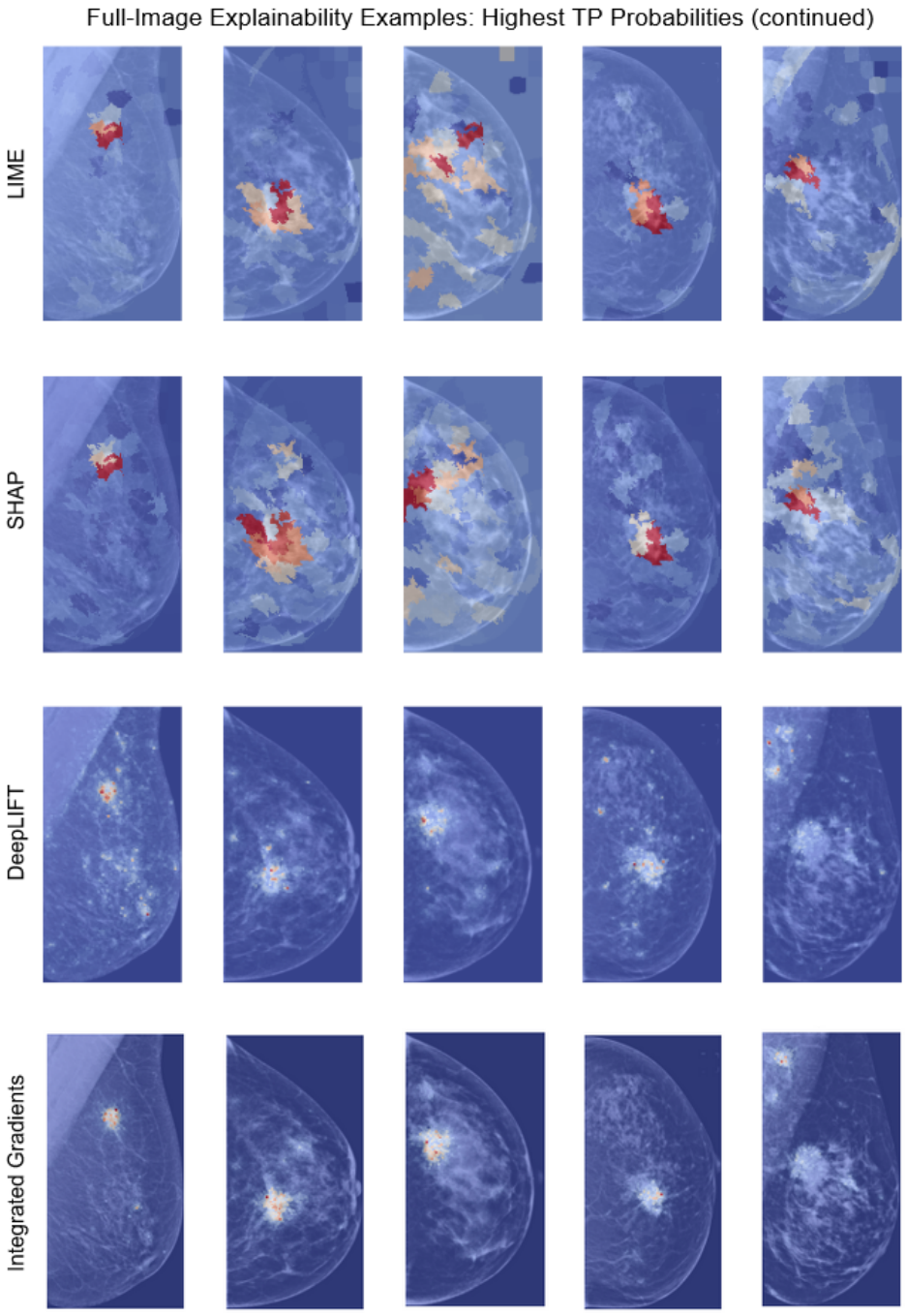


Figure 9: Examples for LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on images with high probability scores. Red and yellow regions correspond to higher attribution scores.

Finally, the attribution maps from the synthetic lesion generated with stable diffusion highlighted the lesion's area. This means that even while being synthetic, the explainability methods indicate that the network was focusing on it when evaluating for malignancy. Therefore, explainability algorithms could provide some insights for evaluating stable diffusion results.

5.1. Limitations and Future Work

The models trained for these classification tasks were not the best performing ones, which most certainly impacted the attribution maps. It could be possible to obtain attribution maps which highlight the lesions even more accurately if applied to better performing models. The bounding boxes were generated following a simple and rudimentary technique, causing them to be larger and less accurate. A more refined method to generate the bounding boxes could be developed, which could

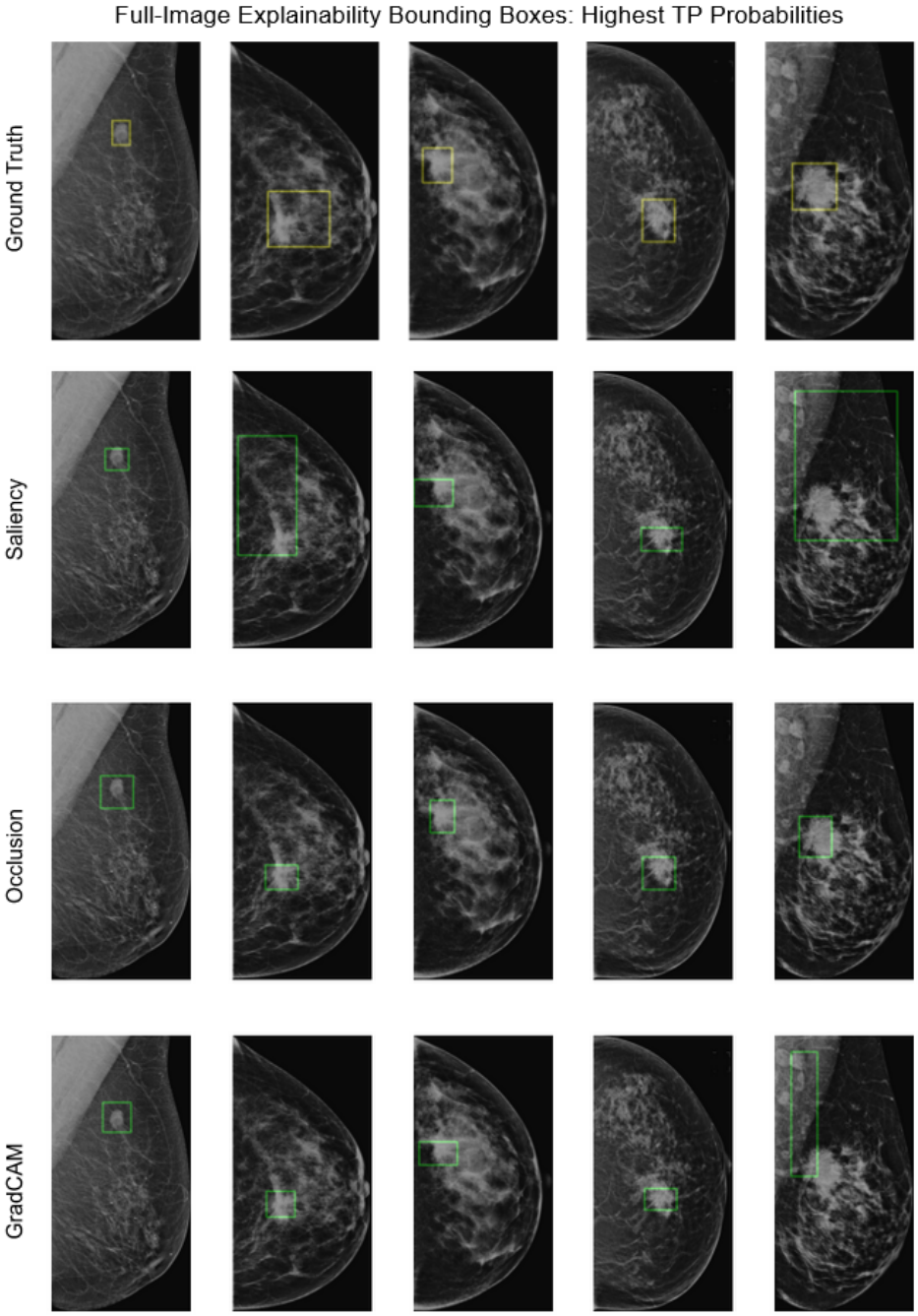


Figure 10: Examples of bounding boxes obtained with Saliency, Occlusion, and GradCAMs' attribution maps on images with high probability scores.

in turn improve the IOU results. Finally, a decently accurate model that classifies between different breast cancer subtypes could provide useful information in the attribution maps. At present, it is not possible to identify these subtypes without a more invasive procedure, but XAI could potentially provide useful morphological information for detecting these subtypes in mammograms.

6. Conclusions

In this paper, some XAI techniques were applied on two breast cancer classification tasks. The results obtained indicate that these techniques could provide users with useful information for understanding the decision-making processes of a neural network in medical imaging. When correctly trained, the methods should highlight the areas with clinical relevance, which in this case translated to the lesions in malignant mammograms. They can also show possible areas of improve-

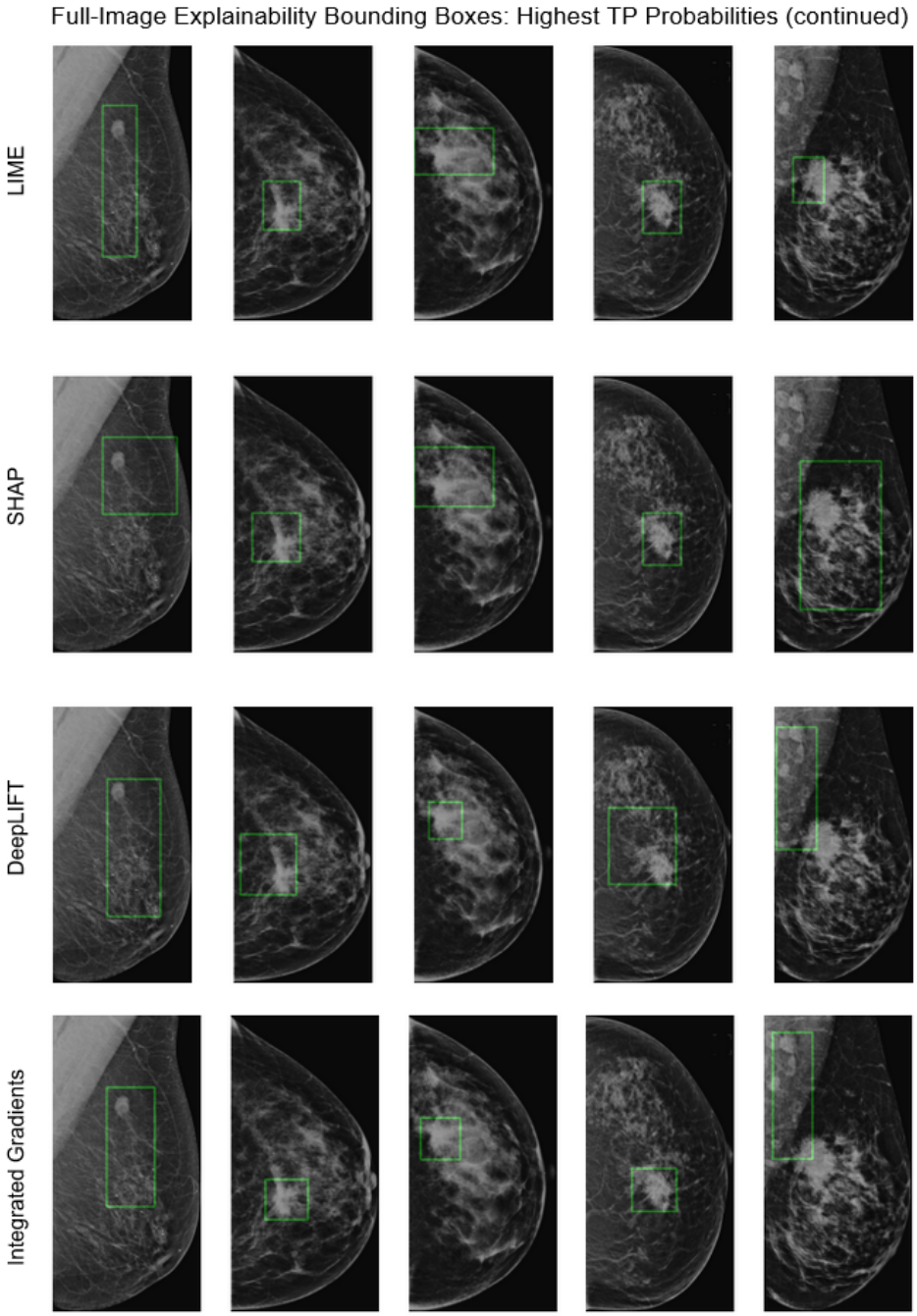


Figure 11: Examples of bounding boxes obtained with LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on images with high probability scores.

ment by indicating the regions which are causing the network to misclassify the images, or to detect possible biases. IOU scores with the lesion location were low but could potentially increase by improving the bounding box generation method from the attribution maps. Integrated gradients possessed the best IOU scores, but it is a rather computationally expensive technique. Grad-CAM and Occlusion could be used instead with potentially slightly worse results. Additionally, the various explainability techniques possess a potential for providing insights for evaluating and subsequently improving the models for the generation of synthetic images via stable diffusion. Future work could improve on this by refining the bounding box generation from the attribution maps and by applying XAI methods for breast cancer subtype classification.

Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, Professor Robert Martí, for his invaluable

guidance and unwavering patience throughout the entire process of developing this thesis. Additionally, I extend my heartfelt appreciation to my fellow MAIA classmates for their continued assistance over the course of these two years, and for enriching this program with priceless experiences that will be remembered fondly. Lastly, I am deeply grateful to my family and friends, whose inconditional support made this journey possible.

References

Akserlod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., Koren, G., Goldschmidt, Y., Shalev, V., Rosen-Zvi, M., Guindy, M., 2019. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 292, 182622. doi:10.1148/radiol.2019182622.

Al-antari, M.A., Han, S.M., Kim, T.S., 2020. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms. Computer Methods and Programs in Biomedicine 196, 105584. doi:https://doi.org/10.1016/j.cmpb.2020.105584.

Arai, S., Nagao, T., 2017. Intuitive visualization method for image classification using convolutional neural networks. Information Processing Society of Japan. Transactions on mathematical modeling and its applications 10, 1–13.

Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Guevara Lopez, M.A., 2015. Convolutional neural networks for mammography mass lesion classification, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 797–800. doi:10.1109/EMBC.2015.7318482.

Balkenende, L., Teuwen, J., Mann, R.M., 2022. Application of deep learning in breast cancer imaging. Seminars in Nuclear Medicine 52, 584–596. doi:https://doi.org/10.1053/j.semnuclmed.2022.02.003. breast Cancer.

Cantone, M., Marrocco, C., Tortorella, F., Bria, A., 2023. Convolutional networks and transformers for mammography classification: An experimental study. Sensors (Basel) 23.

Carr, C., Kitamura, F., Partridge, G., inversion, Kalpathy-Cramer, J., Mongan, J., Lavender, K.A., Vazirabad, M., Riopel, M., Ball, R., Dane, S., Chen, Y., 2022. Rsna screening mammography breast cancer detection. URL: https://kaggle.com/competitions/rsna-breast-cancer-detection.

Halling-Brown, M.D., Warren, L.M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M.G., Wilkinson, L., Given-Wilson, R.M., McAvinchey, R., Young, K.C., 2020. Optimam mammography image database: a large scale resource of mammography images and clinical data.

Huang, Z., Zhu, X., Ding, M., Zhang, X., 2020. Medical image classification using a light-weighted hybrid neural network based on pcanet and densenet. IEEE Access 8, 24697–24712. doi:10.1109/ACCESS.2020.2971225.

Kobayashi, T., Haraguchi, T., Nagao, T., 2022. Classifying presence or absence of calcifications on mammography using generative contribution mapping. Radiological Physics and Technology 15. doi:10.1007/s12194-022-00673-3.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O., 2020. Captum: A unified and generic model interpretability library for pytorch.

Loizidou, K., Elia, R., Pitris, C., 2023. Computer-aided breast cancer detection and classification in mammography: A comprehensive review. Computers in Biology and Medicine 153, 106554. doi:https://doi.org/10.1016/j.compbiomed.2023.106554.

Ou, W.C., Polat, D., Dogan, B.E., 2021. Deep learning in breast radiology: current progress and future directions. European radiology 31, 4872—4885. doi:10.1007/s00330-020-07640-9.

Prodan, M., Paraschiv, E., Stanciu, A., 2023. Applying deep learning methods for mammography analysis and breast cancer detection. Applied Sciences 13, 4272. URL: http://dx.doi.org/10.3390/app13074272, doi:10.3390/app13074272.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128, 336–359. doi:10.1007/s11263-019-01228-7.

Shrikumar, A., Greenside, P., Kundaje, A., 2019. Learning important features through propagating activation differences.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. Journal of Imaging 6.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks.

van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A., 2022. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis 79, 102470. doi:https://doi.org/10.1016/j.media.2022.102470.

de Vries, B.M., Zwezerijnen, G.J.C., Burchell, G.L., van Velden, F.H.P., Menke-van der Houven van Oordt, C.W., Boellaard, R., 2023. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Front Med (Lausanne) 10, 1180773.

Xi, P., Guan, H., Shu, C., Borgeat, L., Goubran, R., 2020. An integrated approach for medical abnormality detection using deep patch convolutional neural networks. The Visual Computer 36. doi:10.1007/s00371-019-01775-7.

Yi, P., Lin, A., Wei, J., Yu, A., Sair, H., Hui, F., Hager, G., Harvey, S., 2019. Deep-learning-based semantic labeling for 2d mammography and comparison of complexity for machine learning tasks. Journal of Digital Imaging 32. doi:10.1007/s10278-019-00244-w.

Zeiler, M.D., Fergus, R., 2013. Visualizing and understanding convolutional networks. arXiv:1311.2901.

Appendix A. Attribution maps for whole mammograms

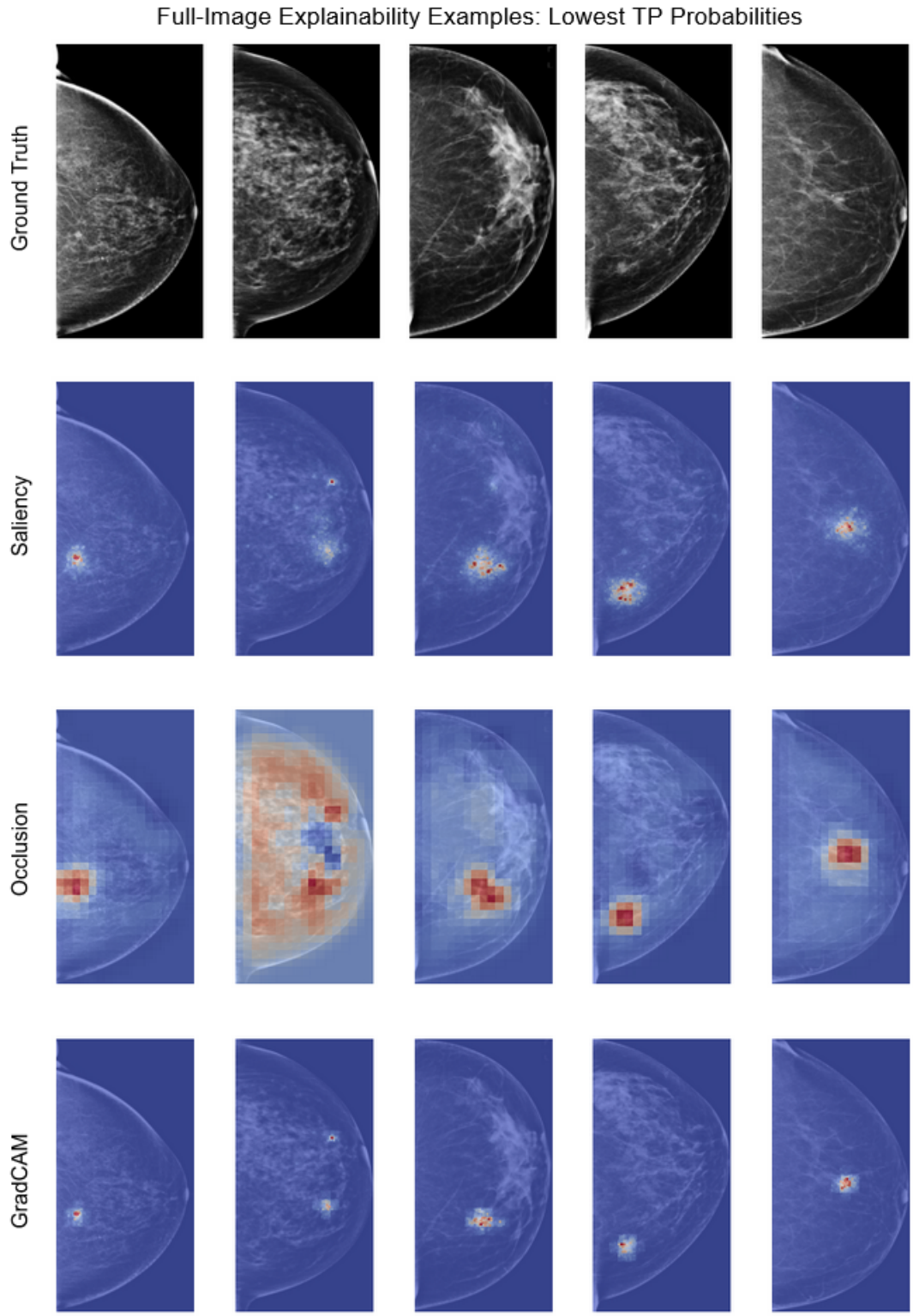


Figure A.12: Examples for Saliency, Occlusion, and GradCAMs' attribution maps on TP images with low probability scores. Red and yellow regions correspond to higher attribution scores.

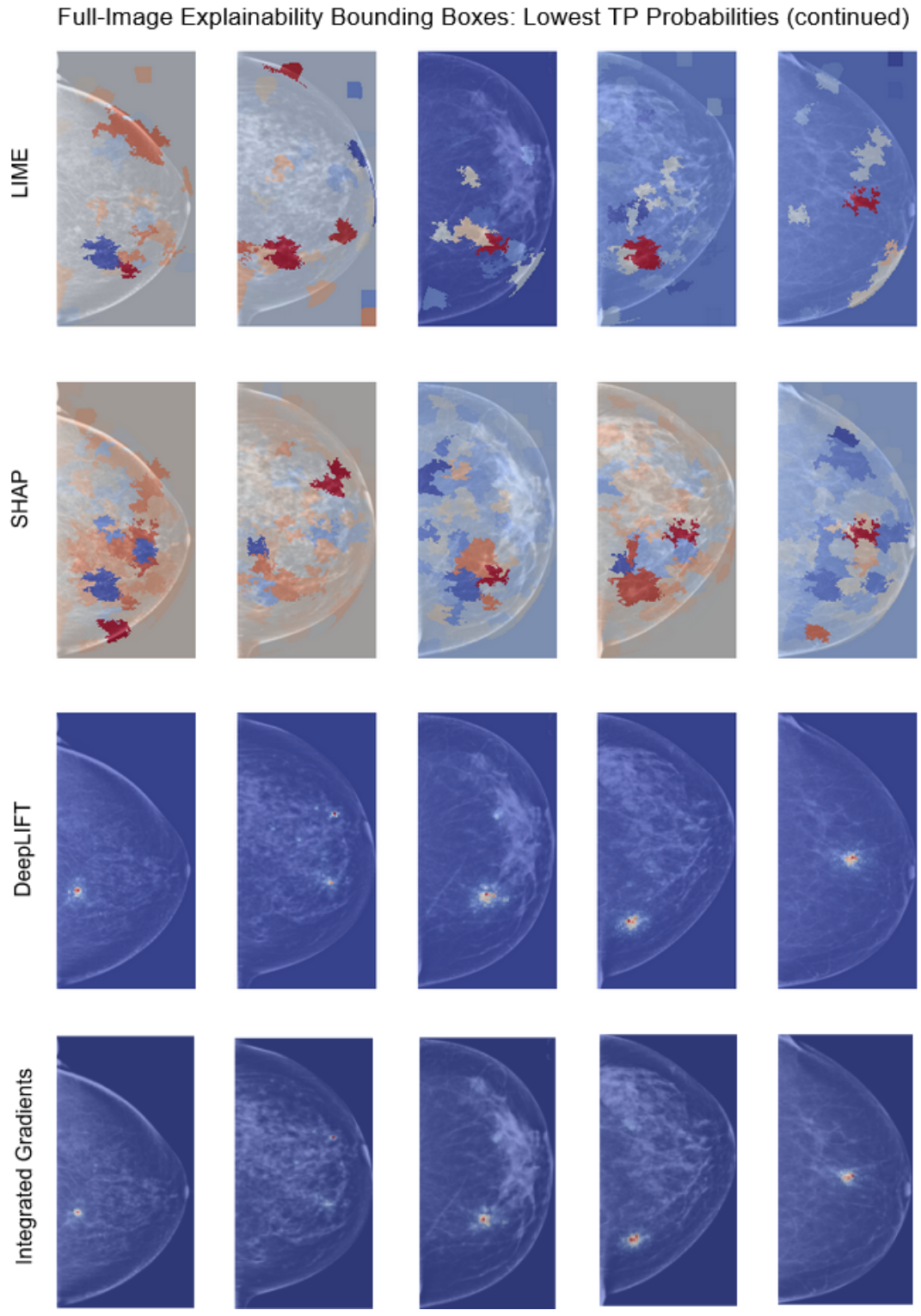


Figure A.13: Examples for LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on TP images with low probability scores. Red and yellow regions correspond to higher attribution scores.

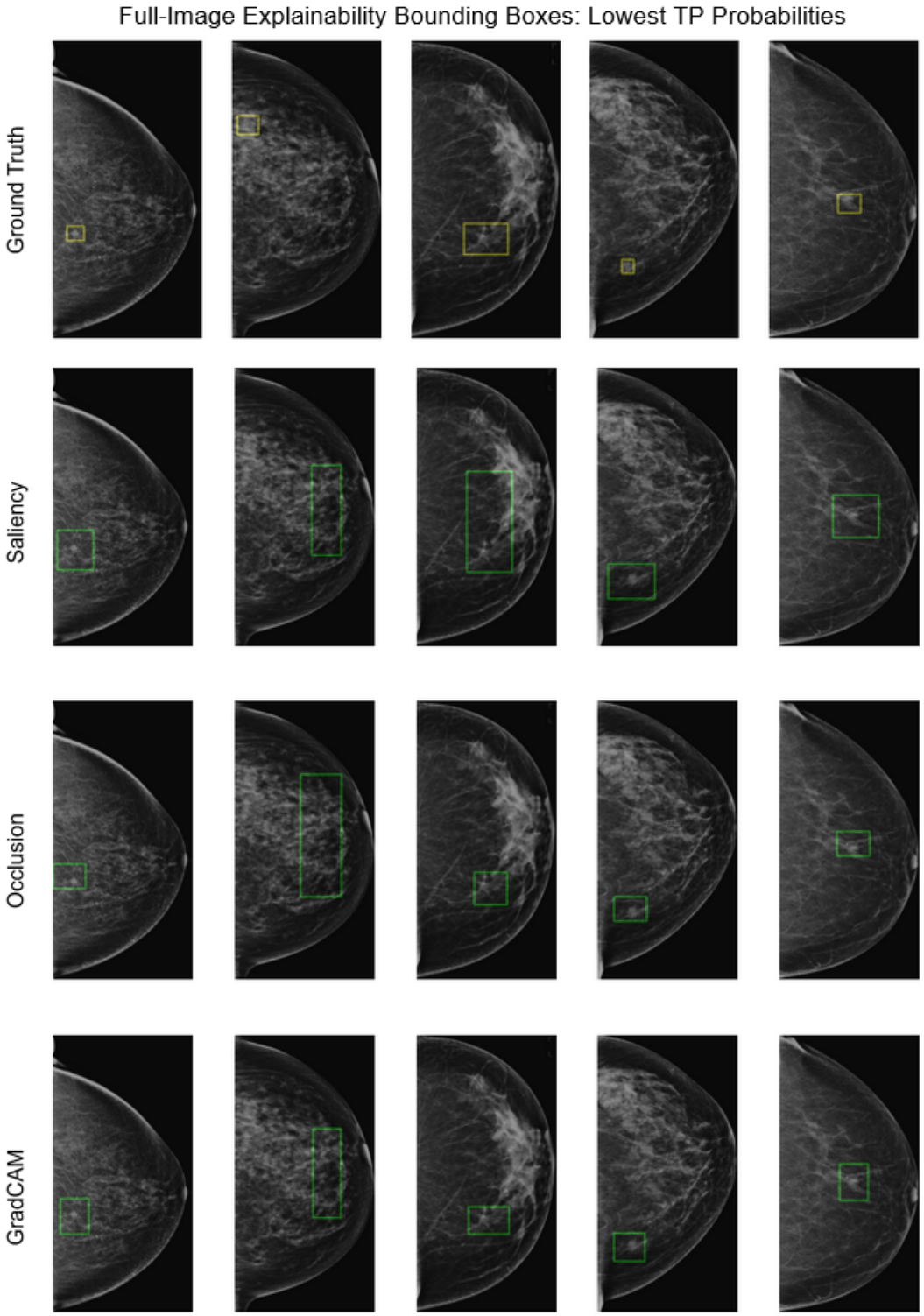


Figure A.14: Examples of bounding boxes obtained with Saliency, Occlusion, and GradCAMs' attribution maps on TP images with low probability scores.

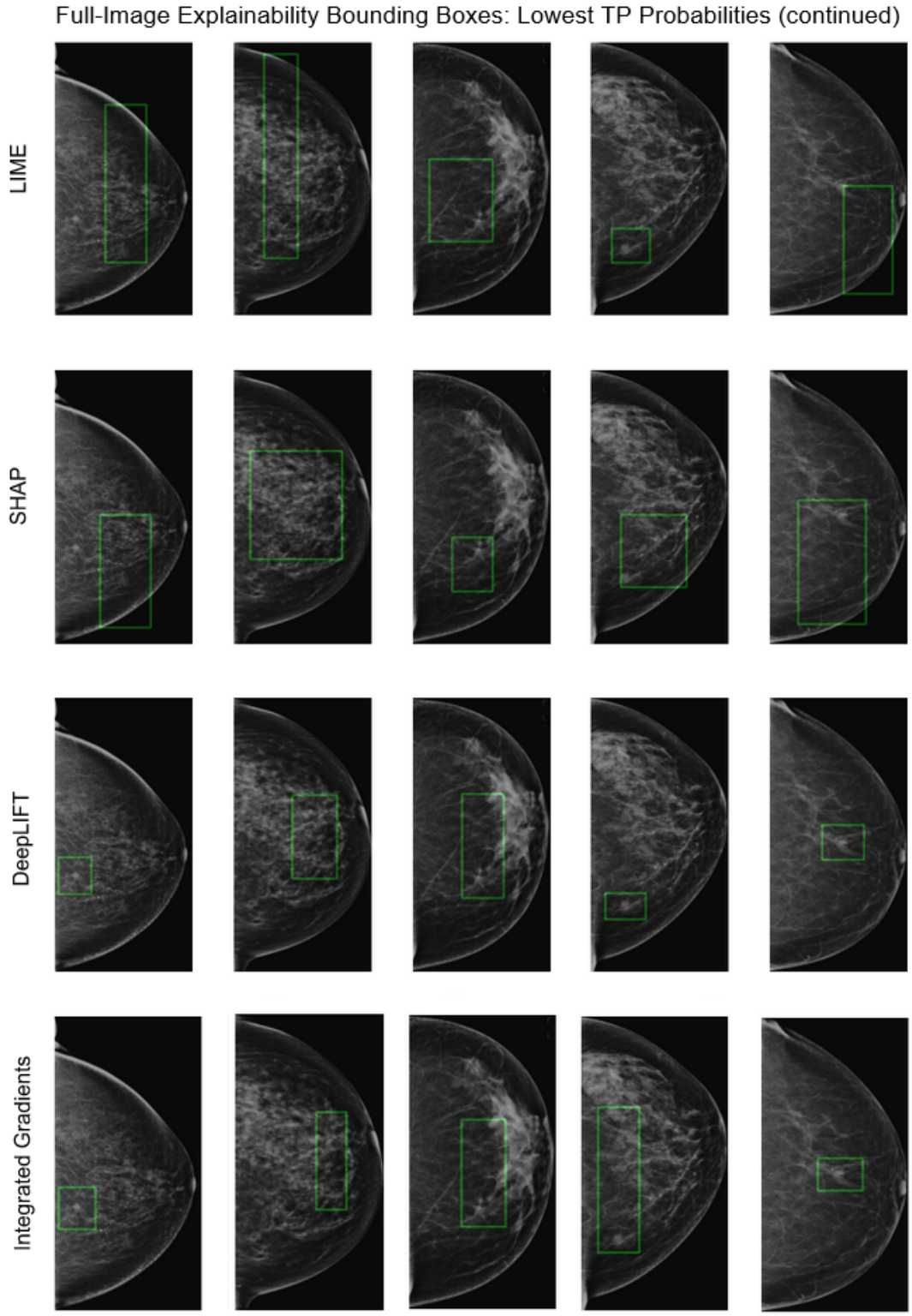


Figure A.15: Examples of bounding boxes obtained with LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on TP images with low probability scores.

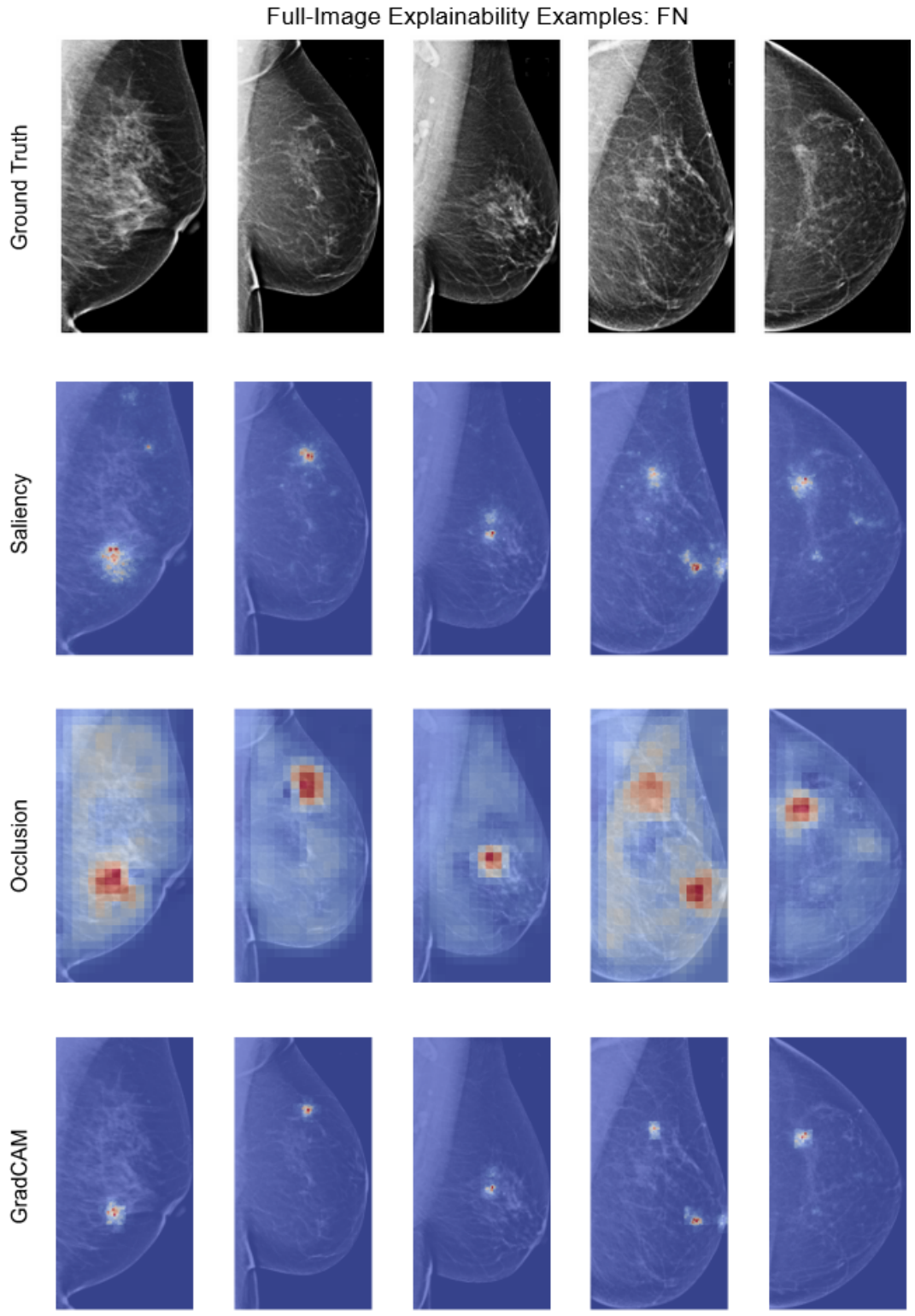


Figure A.16: Examples for Saliency, Occlusion, and GradCAMs' attribution maps on FN images. Red and yellow regions correspond to higher attribution scores.

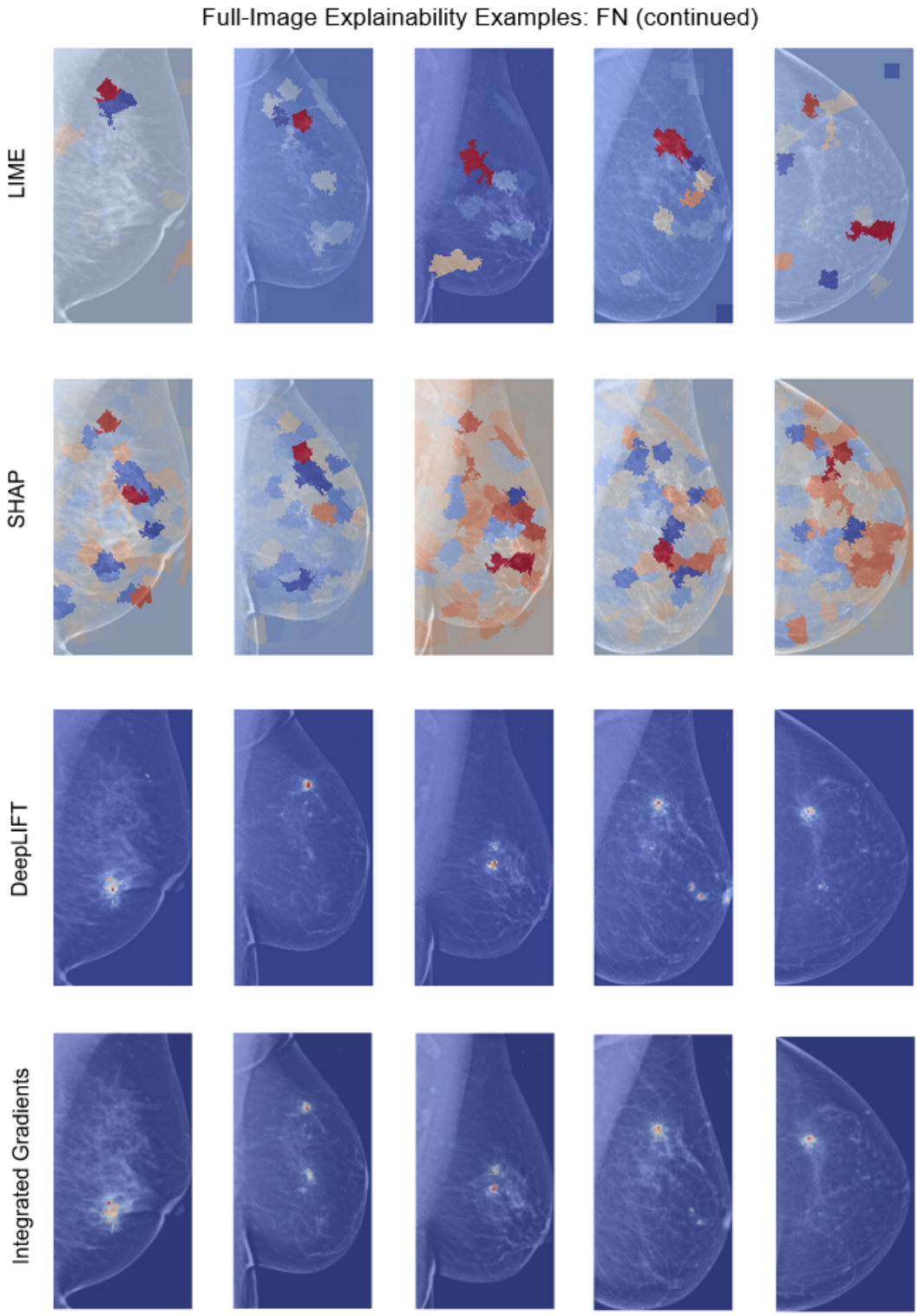


Figure A.17: Examples for LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on FN images. Red and yellow regions correspond to higher attribution scores.

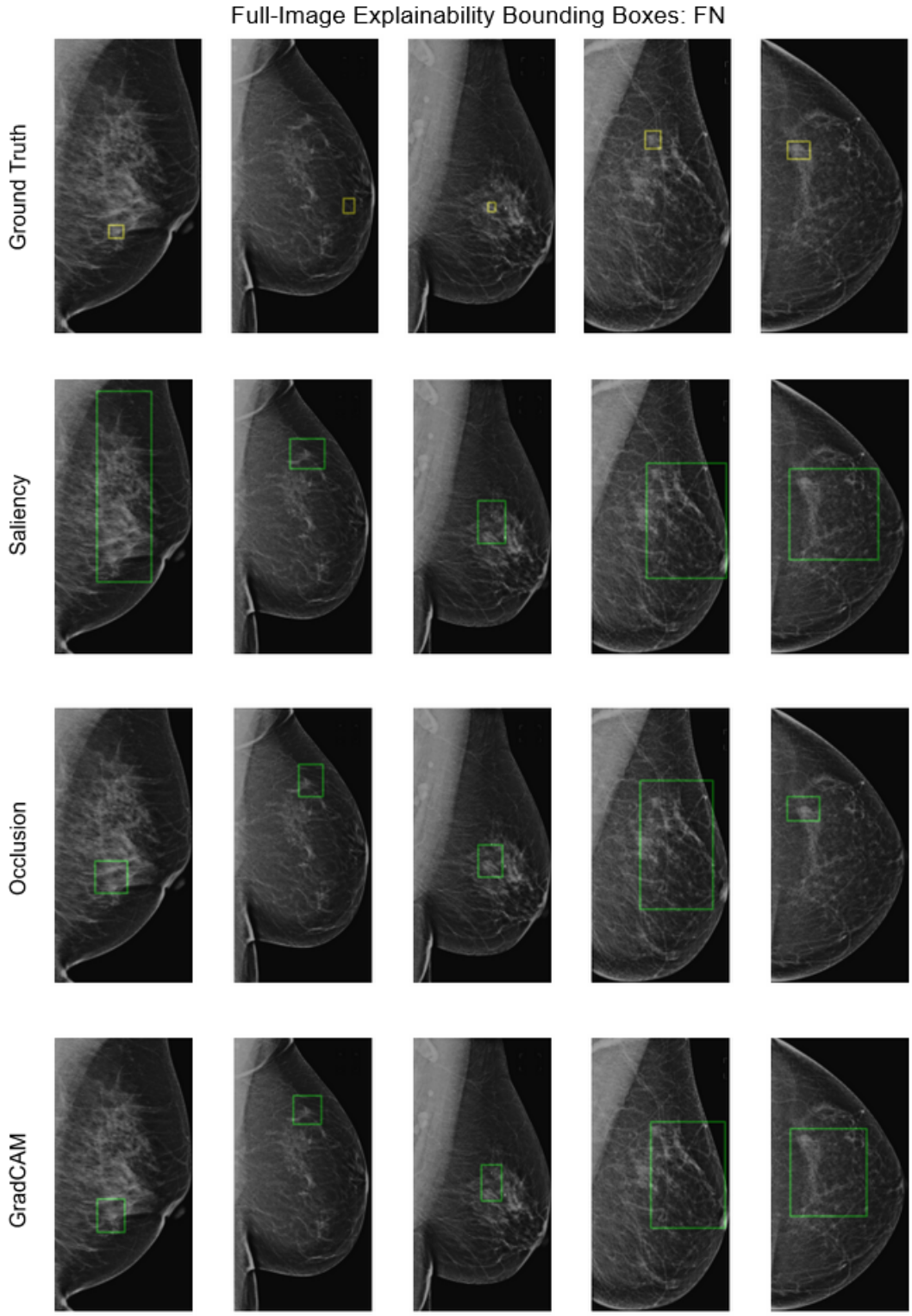


Figure A.18: Examples of bounding boxes obtained with Saliency, Occlusion, and GradCAMs' attribution maps on FN images.

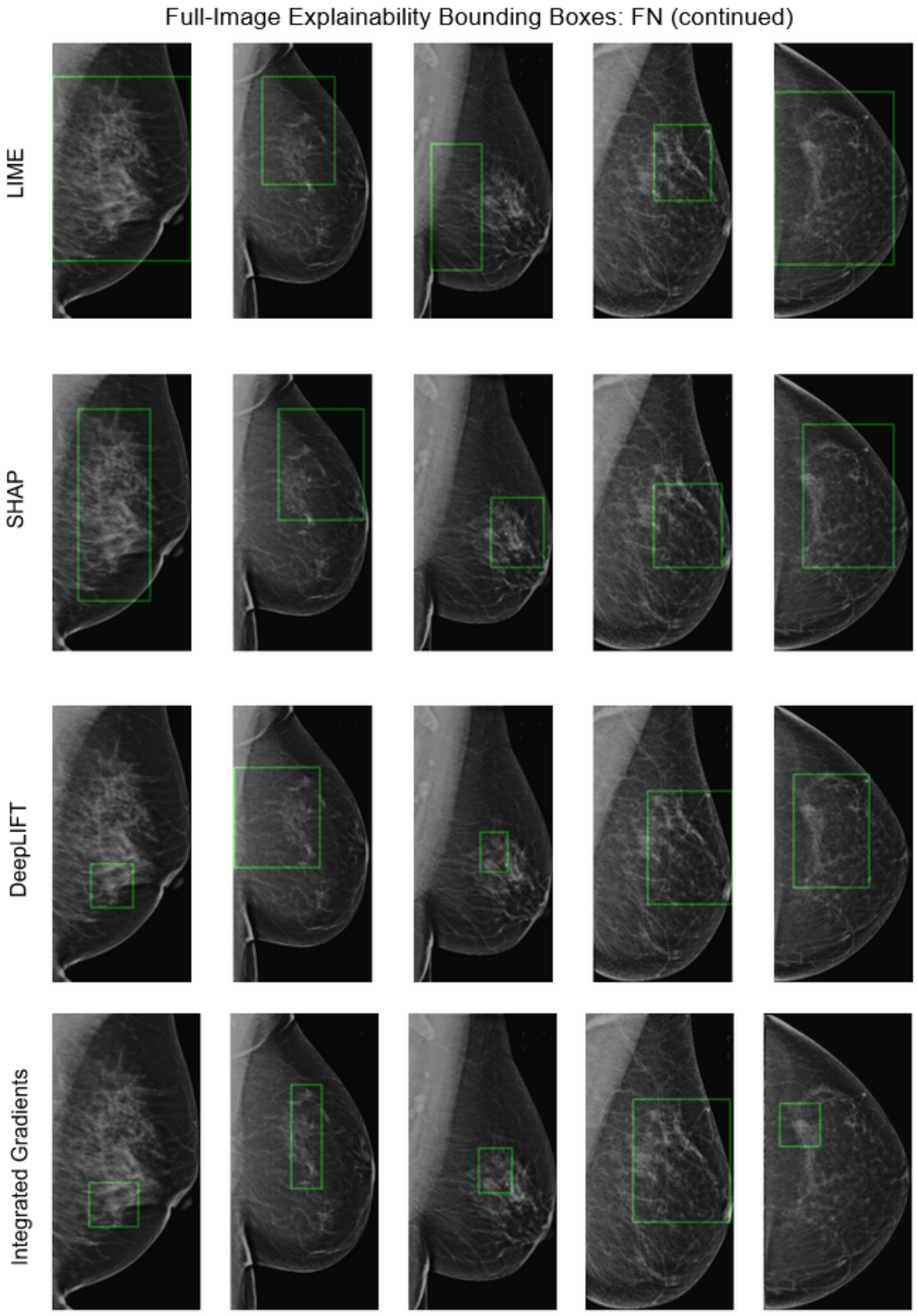


Figure A.19: Examples of bounding boxes obtained with LIME, SHAP, DeepLIFT, and Integrated Gradients' attribution maps on FN images.