Medical Image Analysis 68 (2021) 101908

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization☆

Yiqiu Shen[a], Nan Wu[a], Jason Phang[a], Jungkyu Park[b], Kangning Liu[a], Sudarshini Tyagi[d], Laura Heacock[b,e], S. Gene Kim[b,c,e], Linda Moy[b,c,e], Kyunghyun Cho[a,d,1], Krzysztof J. Geras[a,b,c,*]

[a] Center for Data Science, New York University, 60 5th Ave, New York, NY 10011, USA
[b] Department of Radiology, NYU School of Medicine, 530 1st Ave, New York, NY 10016, USA
[c] Center for Advanced Imaging Innovation and Research, NYU Langone Health, 660 1st Ave, New York, NY 10016, USA
[d] Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012, USA
[e] Perlmutter Cancer Center, NYU Langone Health, 160 E 34th St, New York, NY 10016, USA

ARTICLE INFO

Article history:
Received 18 February 2020
Revised 12 November 2020
Accepted 13 November 2020
Available online 16 December 2020

Keywords:
Deep learning
Breast cancer screening
Weakly supervised localization
High-resolution image classification

ABSTRACT

Medical images differ from natural images in significantly higher resolutions and smaller regions of interest. Because of these differences, neural network architectures that work well for natural images might not be applicable to medical image analysis. In this work, we propose a novel neural network model to address these unique properties of medical images. This model first uses a low-capacity, yet memory-efficient, network on the whole image to identify the most informative regions. It then applies another higher-capacity network to collect details from chosen regions. Finally, it employs a fusion module that aggregates global and local information to make a prediction. While existing methods often require lesion segmentation during training, our model is trained with only image-level labels and can generate pixel-level saliency maps indicating possible malignant findings. We apply the model to screening mammography interpretation: predicting the presence or absence of benign and malignant lesions. On the NYU Breast Cancer Screening Dataset, our model outperforms (AUC = 0.93) ResNet-34 and Faster R-CNN in classifying breasts with malignant findings. On the CBIS-DDSM dataset, our model achieves performance (AUC = 0.858) on par with state-of-the-art approaches. Compared to ResNet-34, our model is 4.1x faster for inference while using 78.4% less GPU memory. Furthermore, we demonstrate, in a reader study, that our model surpasses radiologist-level AUC by a margin of 0.11.

1. Introduction

Breast cancer is the second leading cause of cancer-related death among women in the United States (DeSantis et al., 2017). It is estimated that 276,480 women would be diagnosed with breast cancer and 42,170 would die in 2020 (Siegel et al., 2020). Screening mammography, a low-dose X-ray examination, is a major tool for early detection of breast cancer. A standard screening mammogram consists of two high-resolution X-rays of each breast, taken from the side (the mediolateral oblique or MLO view) and from above (the craniocaudal or CC view) for a total of four images. Radiologists, physicians specialized in the interpretation of medical images, analyze screening mammograms for tissue abnormalities that may indicate breast cancer. Any detected abnormality leads to additional diagnostic imaging and possible tissue biopsy. A radiologist assigns a standardized assessment to each screening mammogram per the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS), with specific follow-up recommendations for each category (Liberman and Menell, 2002).

Screening mammography interpretation is a particularly challenging task because mammograms are in very high resolutions while most asymptomatic cancer lesions are small, sparsely distributed over the breast and may present as subtle changes in the breast tissue pattern. While randomized clinical trials have shown

☆ This paper is an extension of work originally presented at the 10th International Workshop on Machine Learning in Medical Imaging Shen et al. (2019b).

* Corresponding author.
E-mail address: k.j.geras@nyu.edu (K.J. Geras).

[1] CIFAR Associate Fellow

https://doi.org/10.1016/j.media.2020.101908

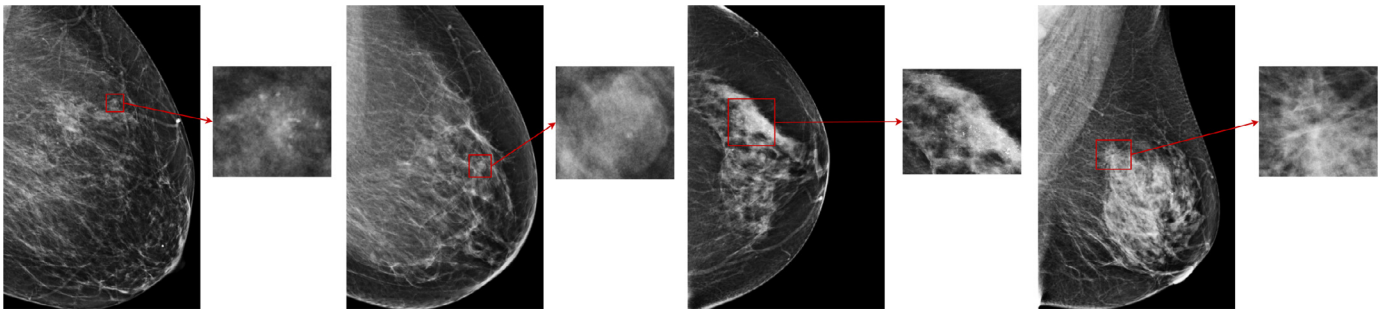

Fig. 1. Four examples of breasts that were biopsied along with the annotated findings. The breasts (from left to right) were diagnosed with benign calcifications, a benign mass, malignant calcifications, and malignant architectural distortion. While microcalcifications are common in both benign and malignant findings, their presence in a ductal distribution, such as in the third example, is a strong indicator of malignancy.

that screening mammography has significantly reduced breast cancer mortality (Duffy et al., 2002; Kopans, 2002), it is associated with limitations such as false positive recalls for additional imaging and subsequent false positive biopsies which result in benign, non-cancerous findings. About 10% to 20% of women who have an abnormal screening mammogram are recommended to undergo a biopsy. Only 20% to 40% of these biopsies yield a diagnosis of cancer (Kopans, 2015).

To tackle these limitations, convolutional neural networks (CNN) have been applied to assist radiologists in the analysis of screening mammography (Kim et al., 2018; McKinney et al., 2020; Ribli et al., 2018; Zhu et al., 2017; Kyono, Gilbert, van der Schaar; Wu, Phang, Park, Shen, Huang, Zorin, Jastrzebski, Févry, Katsnelson, Kim, et al.). An overwhelming majority of existing studies on this task utilize models that were originally designed for natural images. For instance, VGGNet (Simonyan and Zisserman, 2014), designed for object classification on ImageNet (Deng et al., 2009), has been applied to breast density classification (Wu et al., 2018) and Faster R-CNN (Ren et al., 2015) has been adapted to localize suspicious findings in mammograms (Ribli et al., 2018; Févry, Phang, Wu, Kim, Moy, Cho, Geras).

Screening mammography is inherently different from typical natural images from a few perspectives. First of all, as illustrated in Fig. 1, regions of interest (ROI) in mammography images, such as masses, asymmetries, and microcalcifications, are often smaller in comparison to the salient objects in natural images. Moreover, as suggested in multiple clinical studies (Van Gils et al., 1998; Pereira et al., 2009; Wei et al., 2011), both the local details, such as lesion shape, and global structure, such as overall breast fibroglandular tissue density and pattern, are essential for accurate diagnosis. For instance, while microcalcifications are common in both benign and malignant findings, their presence in a ductal distribution, such as in the third example of Fig. 1, is a strong indicator of malignancy. This is in contrast to typical natural images where objects outside the most salient regions provide little information towards predicting the label of the image. In addition, mammography images are usually of much higher resolutions than typical natural images. The most accurate deep CNN architectures for natural images are not applicable to mammography images due to the limited size of GPU memory.

To address the aforementioned issues, in this work, we extended and comprehensively evaluated the globally-aware multiple instance classifier (GMIC), whose preliminary version we proposed in Shen et al. (2019). GMIC first applies a low-capacity, yet memory-efficient, global module on the whole image to generate saliency maps that provide coarse localization of possible benign/malignant findings. As a result, GMIC is able to process screening mammography images in their original resolutions while keeping GPU memory manageable. In order to capture subtle patterns contained in small ROIs, GMIC then identifies the most informative regions in the image and utilizes a high-capacity local module to extract fine-grained visual details from these regions. Finally, it employs a fusion module that aggregates information from both global context and local details to predict the presence or absence of benign and malignant lesions in a breast. The specific contributions of this work are the following:

- We extended the original architecture (Shen et al., 2019) with a fusion module which combines information from both global and local features. We applied the improved model to the task of screening mammography interpretation: predicting the presence or absence of benign and malignant lesions. On the NYU Breast Cancer Screening Dataset (NYUBCS) (Wu et al., 2019c), consisting of more than one million images, GMIC achieves an AUC of 0.93 in identifying breasts with malignant findings, outperforming baselines including ResNet-34 (He et al., 2016a), Faster R-CNN (Févry et al., 2019), and DMV-CNN (Wu et al., 2019b). To demonstrate its generalizability, we trained and evaluated GMIC on the CBIS-DDSM dataset (Lee et al., 2017). We showed that GMIC achieved slightly stronger performance (AUC = 0.858) than the state-of-the-art approaches (Zhu et al., 2017; Shu et al., 2020). In addition, GMIC is computationally efficient. Compared to ResNet-34, GMIC has 28.8% fewer parameters, uses 78.4% less GPU memory, is 5.6x faster during training and 4.1x faster during inference.
- We demonstrate the clinical potential of the GMIC by comparing the improved model to human experts. In the reader study, we show that it surpasses a radiologist-level classification performance: the AUC for the proposed model was greater than the average AUC for radiologists by a margin of 0.11, reducing the error approximately by half. In addition, we experimented with hybrid models that combine predictions from both GMIC and each of the radiologists separately. At radiologists' sensitivity (62.1%), the hybrid models achieve an average specificity of 91.9% improving radiologists' average specificity by 6.3%.
- An advantage of GMIC over networks, such as Faster R-CNN (Ren et al., 2015) and its derivatives (Ribli et al., 2018), is that GMIC only needs image-level labels (e.g. presence of cancer) to learn to localize lesions, so it does not reply on manual segmentation (e.g. pixel-level location of caner lesions) which is often expensive to obtain for medical images. In Section 3.5, we demonstrate that the regions highlighted by the saliency maps indeed correlate with the objects of interest.

2. Methods

We frame the task of screening mammography interpretation as a multi-label classification problem: given a grayscale image $\mathbf{x} \in \mathbb{R}^{H,W}$, we predict the image-level label $\mathbf{y} = \begin{bmatrix} y^b \\ y^m \end{bmatrix}$, where $y^b, y^m \in \{0, 1\}$ indicate whether any benign/malignant lesion is present in $\mathbf{x}$.

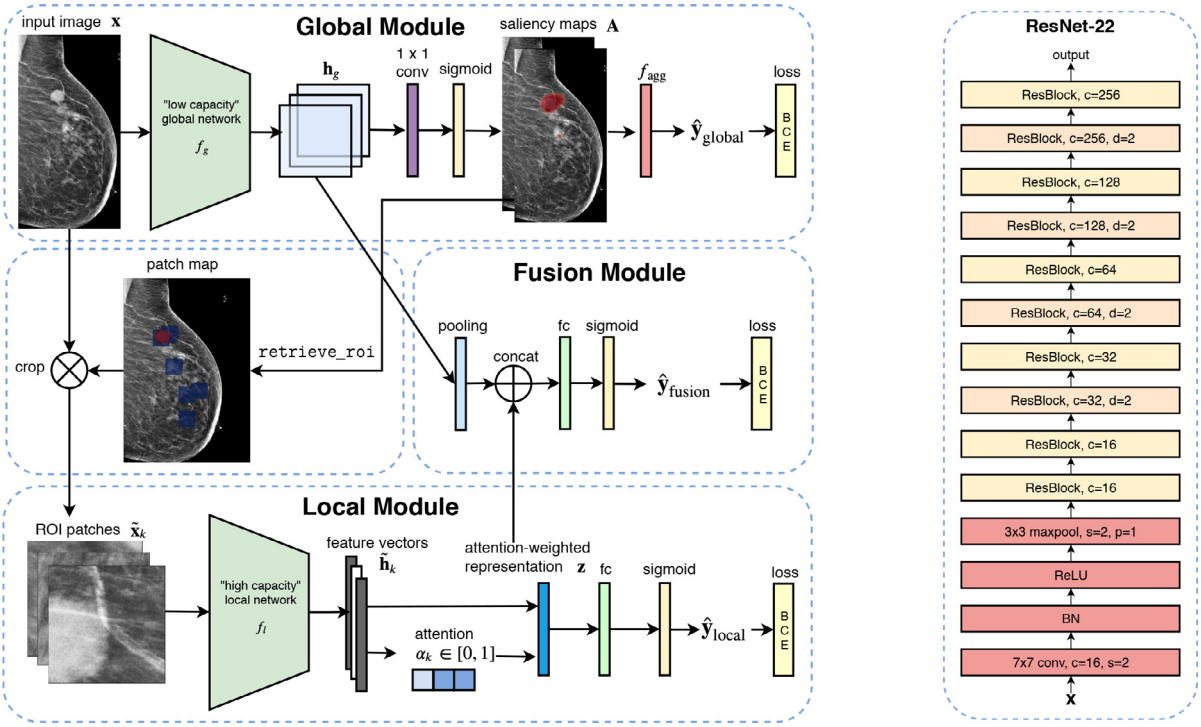

Fig. 2. Overall architecture of GMIC (left) and architecture of ResNet-22 (right).[2] The patch map indicates positions of ROI patches (blue squares) on the input. In ResNet-22, we use c, s, and p to denote number of output channels, strides and size of padding. "ResBlock, c = 32, d = 2" denotes a vanilla ResBlock proposed in He et al. (2016b) with 32 output channels and a downsample skip connection that reduces the resolution with a factor of 2. In comparison to canonical ResNet architectures (He et al., 2016a), ResNet-22 has one more residual block and only a quarter of the filters in each convolution layer. Narrowing network width decreases the total number of hidden units which reduces GPU memory consumption. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.1. Globally-aware classification framework

As shown in Fig. 2, we propose a classification framework that resembles the diagnostic procedure of a radiologist. We first use a global network f_g to extract a feature map $\mathbf{h}_g$ from the input image $\mathbf{x}$, i.e. we compute

$$\mathbf{h}_g = f_g(\mathbf{x}), \tag{1}$$

which is analogous to a radiologist roughly scanning through the entire image to obtain a holistic view.

We then apply a 1×1 convolution layer with sigmoid non-linearity to transform $\mathbf{h}_g$ into two saliency maps $\mathbf{A}^b, \mathbf{A}^m \in \mathbb{R}^{h,w}$ indicating approximate locations of benign and malignant lesions. Each element $\mathbf{A}^c_{i,j} \in [0, 1]$ where $c \in \{b, m\}$, denotes the contribution of spatial location (i, j) towards predicting the presence of benign/malignant lesions. Let $\mathbf{A}$ denote the concatenation of $\mathbf{A}^b$ and $\mathbf{A}^m$. That is, we compute $\mathbf{A}$ as

$$\mathbf{A} = \text{sigm}(\text{conv}_{1\times1}(\mathbf{h}_g)). \tag{2}$$

Due to limited GPU memory, in prior work, input images $\mathbf{x}$ are usually down-sampled (Guan et al., 2018; Yao et al., 2018; Zhong et al., 2019). For mammography images, however, down-sampling distorts important visual details such as lesion margins and blurs small microcalcifications. Instead of sacrificing the input resolution, we control memory consumption by reducing the complexity of the *global* network f_g. Because of its constrained capacity, f_g may not be able to capture all subtle patterns contained in the images at all scales. To compensate for this, we utilize a high-capacity *local* network f_l to extract fine-grained details from a set of informative regions. In the second stage, we use $\mathbf{A}$ to retrieve K most informative patches from $\mathbf{x}$:

$$\{\tilde{\mathbf{x}}_k\} = \texttt{retrieve\_roi}(\mathbf{A}), \tag{3}$$

where `retrieve_roi` denotes a heuristic patch-selection procedure described later. This procedure can be seen as an analogue to a radiologist concentrating on areas that might correspond to lesions. The fine-grained visual features $\{\tilde{\mathbf{h}}_k\}$ contained in all chosen patches $\{\tilde{\mathbf{x}}_k\}$ are then processed using f_l and are aggregated into a vector $\mathbf{z}$ by an aggregator f_a. That is,

$$\tilde{\mathbf{h}}_k = f_l(\tilde{\mathbf{x}}_k) \quad \text{and} \quad \mathbf{z} = f_a(\{\tilde{\mathbf{h}}_k\}). \tag{4}$$

Finally, a fusion network f_{fusion} combines information from both global structure $\mathbf{h}_g$ and local details $\mathbf{z}$ to produce a prediction $\hat{\mathbf{y}}$. This is analogous to modelling a radiologist comprehensively considering the global and local information to render a full diagnosis as

$$\hat{\mathbf{y}} = f_{\text{fusion}}(\mathbf{h}_g, \mathbf{z}). \tag{5}$$

2.2. Model parameterizaiton

Generating the saliency maps

To process high-resolution images while keeping GPU memory consumption manageable, we parameterize f_g as a ResNet-22 (Wu et al., 2019b) whose architecture is shown in Fig. 2. In comparison to canonical ResNet architectures (He et al., 2016a), ResNet-22 has one more residual block and only a quarter of the filters in each convolution layer. As suggested by Tan and Le (2019), a deeper CNN has larger receptive fields and can capture richer and more complex features in high-resolution images. Narrowing network width can decrease the total number of hidden units which reduces GPU memory consumption.

[2] In our experiments, each input image $\mathbf{x}$ has a resolution of 2944 × 1920 pixels and each ROI patch $\tilde{\mathbf{x}}_k$ has a resolution of 256 × 256 pixels. The dimensions of the intermediate representations depend on the implementation of f_g and f_l. With f_g parameterized as ResNet-22 and f_l parameterized as ResNet-18, we have the following dimensions: $\mathbf{h}_g \in \mathbb{R}^{46,30,256}$, $\mathbf{A} \in \mathbb{R}^{46,30,2}$, $\tilde{\mathbf{h}}_k \in \mathbb{R}^{512}$, and $\mathbf{z} \in \mathbb{R}^{512}$.

It is difficult to define a loss function that directly compares saliency maps $\mathbf{A}$ and the cancer label $\mathbf{y}$, since $\mathbf{y}$ does not contain localization information. In order to train f_g, we use an aggregation function $f_{\text{agg}}(\mathbf{A}^c) : \mathbb{R}^{h,w} \mapsto [0,1]$ to transform a saliency map into an image-level class prediction:

$$\hat{\mathbf{y}}^c_{\text{global}} = f_{\text{agg}}(\mathbf{A}^c). \tag{6}$$

With f_{agg} we can train f_g by backpropagating the gradient of the classification loss between $\mathbf{y}$ and $\hat{\mathbf{y}}_{\text{global}}$. The design of $f_{\text{agg}}(\mathbf{A}^c)$ has been extensively studied (Durand et al., 2017). Global average pooling (GAP) would dilute the prediction as most of the spatial locations in $\mathbf{A}^c$ correspond to background and provide little training signal. On the other hand, with global max pooling (GMP), the gradient is backpropagated through a single spatial location, which makes the learning process slow and unstable. In our work, we propose, *top t% pooling*, which is a soft balance between GAP and GMP. Namely, we define the aggregation function as

$$f_{\text{agg}}(\mathbf{A}^c) = \frac{1}{|H^+|} \sum_{(i,j)\in H^+} \mathbf{A}^c_{i,j}, \tag{7}$$

where H^+ denotes the set containing locations of top t% values in $\mathbf{A}^c$, where t is a hyperparameter. In all experiments, we tune t using a procedure described in Section 3.3. In fact, GAP and GMP can be viewed as two extremes of *top t% pooling*. GMP is equivalent to setting $t = \frac{1}{h \times w}$ and GAP is equivalent to setting $t = 100\%$. In Section 3.6, we study the impact of t and empirically demonstrate that our parameterization of f_{agg} achieves performance superior to GAP and GMP.

Acquiring ROI patches We designed a greedy algorithm (Algorithm 1) to retrieve K patches as proposals for ROIs, $\tilde{\mathbf{x}}_k \in \mathbb{R}^{h_c,w_c}$, from the input $\mathbf{x}$, where $w_c = h_c = 256$ in all experiments. In each iteration, `retrieve_roi` greedily selects the rectangular bounding box that maximizes the criterion defined in line 7. The algorithm then maps each selected bounding box to its corresponding location on the input image. The reset rule in line 12 explicitly ensures that extracted ROI patches do not significantly overlap with each other. In Section 3.6, we show how the classification performance is impacted by K.

Algorithm 1 retrieve_roi

Input: $\mathbf{x} \in \mathbb{R}^{H,W}$, $\mathbf{A} \in \mathbb{R}^{h,w,2}$, K
Output: $O = \{\tilde{\mathbf{x}}_k | \tilde{\mathbf{x}}_k \in \mathbb{R}^{h_c,w_c}\}$
1: $O = \emptyset$
2: **for** each class $c \in$ {benign, malignant} **do**
3: $\quad \tilde{\mathbf{A}}^c = \text{min-max-normalization}(\mathbf{A}^c)$
4: **end for**
5: $\mathbf{A}^* = \sum_c \tilde{\mathbf{A}}^c$
6: l denotes an arbitrary $h_c \frac{h}{H} \times w_c \frac{w}{W}$ rectangular patch on $\mathbf{A}^*$
7: $\text{criterion}(l, \mathbf{A}^*) = \sum_{(i,j)\in l} \mathbf{A}^*[i,j]$
8: **for** each $1, 2, ..., K$ **do**
9: $\quad l^* = \text{argmax}_l \text{criterion}(l, \mathbf{A}^*)$
10: $\quad L =$ position of l^* in $\mathbf{x}$
11: $\quad O = O \cup \{L\}$
12: $\quad \forall (i,j) \in l^*$, set $\mathbf{A}^*[i,j] = 0$
13: **end for**
14: **return** O

Utilizing information from patches

With `retrieve_roi`, we can focus learning on a selected set of small yet informative patches $\{\tilde{\mathbf{x}}_k\}$. We can now apply a local network f_l with higher capacity (wider or deeper) that is able to utilize fine-grained visual features, to extract a vector representation $\tilde{\mathbf{h}}_k \in \mathbb{R}^L$ from every patch $\tilde{\mathbf{x}}_k$. We experiment with several parameterizations of f_l including ResNet-18, ResNet-34 and ResNet-50.

To combine information from all ROI patches, we utilize the aggregator f_a which computes an attention-weighted average of vector representations $\tilde{\mathbf{h}}_k$, as formalized in Eq. (9). Since ROI patches are retrieved using coarse saliency maps, the information relevant for classification carried in each patch varies significantly. To address this issue, we use the Gated Attention Mechanism (GA) (Ilse et al., 2018), allowing the model to selectively incorporate information from all patches. Compared to other common attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015), GA uses the sigmoid function to provide a learnable non-linearity which increases model flexibility. An attention score α_k is computed on each patch:

$$\alpha_k = \frac{\exp\{\mathbf{w}^\intercal(\tanh(\mathbf{V}\tilde{\mathbf{h}}_k^\intercal) \odot \text{sigm}(\mathbf{U}\tilde{\mathbf{h}}_k^\intercal))\}}{\sum_{j=1}^K \exp\{\mathbf{w}^\intercal(\tanh(\mathbf{V}\tilde{\mathbf{h}}_j^\intercal) \odot \text{sigm}(\mathbf{U}\tilde{\mathbf{h}}_j^\intercal))\}}, \tag{8}$$

where $\odot$ denotes an element-wise multiplication and $\mathbf{w} \in \mathbb{R}^L$, $\mathbf{V} \in \mathbb{R}^{L\times M}$, $\mathbf{U} \in \mathbb{R}^{L\times M}$ are learnable parameters. In all experiments, we set $L = 512$ and $M = 128$. This process yields an attention-weighted representation

$$\mathbf{z} = \sum_{k=1}^K \alpha_k \tilde{\mathbf{h}}_k, \tag{9}$$

where the attention score $\alpha_k \in [0,1]$ indicates the relevance of each patch $\tilde{\mathbf{x}}_k$. The representation $\mathbf{z}$ is then passed to a fully connected layer with sigmoid activation to generate a prediction

$$\hat{\mathbf{y}}_{\text{local}} = \text{sigm}(\mathbf{w}_{\text{local}}{}^\intercal \mathbf{z}), \tag{10}$$

where $\mathbf{w}_{\text{local}} \in \mathbb{R}^{L\times 2}$ are learnable parameters.

Information fusion To combine information from both saliency maps and ROI patches, we apply a global max pooling on $\mathbf{h}_g$ and concatenate it with $\mathbf{z}$. The concatenated representation is then fed into a fully connected layer with sigmoid activation to produce the final prediction:

$$\hat{\mathbf{y}}_{\text{fusion}} = \text{sigm}(\mathbf{w}_f[\text{GMP}(\mathbf{h}_g), \mathbf{z}]^\intercal) \tag{11}$$

where GMP denotes the global max pooling operator and $\mathbf{w}_f$ are learnable parameters.

2.3. Learning the parameters of GMIC

In order to constrain the saliency maps to only highlight important regions, we impose the $L1$ regularization on $\mathbf{A}^c$ to make the saliency maps sparser:

$$L_{\text{reg}}(\mathbf{A}^c) = \sum_{(i,j)} |\mathbf{A}^c_{i,j}|. \tag{12}$$

Despite the relative complexity of our proposed framework, the model can be trained end-to-end using stochastic gradient descent with following loss function, defined for a single training example as:

$$\begin{aligned} L(\mathbf{y}, \hat{\mathbf{y}}) = &\textstyle\sum_{c\in\{b,m\}} \text{BCE}(\mathbf{y}^c, \hat{\mathbf{y}}^c_{\text{local}}) + \text{BCE}(\mathbf{y}^c, \hat{\mathbf{y}}^c_{\text{global}}) \\ &+\text{BCE}(\mathbf{y}^c, \hat{\mathbf{y}}^c_{\text{fusion}}) + \beta L_{\text{reg}}(\mathbf{A}^c), \end{aligned} \tag{13}$$

where BCE is the binary cross-entropy and β is a hyperparameter. While all of $\hat{\mathbf{y}}_{\text{global}}$, $\hat{\mathbf{y}}_{\text{local}}$, and $\hat{\mathbf{y}}_{\text{fusion}}$ are used in the loss calculation during training, we use $\hat{\mathbf{y}}_{\text{fusion}}$ as the predictions of the model at test time since $\hat{\mathbf{y}}_{\text{fusion}}$ already contains information that is used to derive both $\hat{\mathbf{y}}_{\text{local}}$ and $\hat{\mathbf{y}}_{\text{fusion}}$.

3. Experiments and results

To demonstrate the effectiveness of GMIC on high-resolution image classification, we evaluated it on the task of screening mammography interpretation: predicting the presence or absence of benign and malignant findings in a breast. On NYUBCS, we

compared GMIC to a ResNet-like network dedicated to mammography (Wu et al., 2019b) as well as to the standard ResNet-34 (He et al., 2016a) and Faster-RCNN (Ren et al., 2015; Févry et al., 2019) in terms of classification accuracy, number of parameters, computation time, and GPU memory consumption. On the CBIS-DDSM dataset, we compared GMIC to two state-of-the-art models designed for whole-mammogram classification (Zhu et al., 2017; Shu et al., 2020). In addition, we also evaluated the localization performance of GMIC on NYUBCS by qualitatively and quantitatively comparing the saliency maps produced by GMIC with the ground truth segmentation provided by the radiologists.

3.1. Data

NYU breast cancer screening dataset The NYU Breast Cancer Screening Dataset (Wu et al., 2019c) includes 229,426 exams (1,001,093 images) from 141,472 patients.[3] Each exam contains at least four images which correspond to the four standard views used in screening mammography: R-CC (right craniocaudal), L-CC (left craniocaudal), R-MLO (right mediolateral oblique) and L-MLO (left mediolateral oblique). An example is shown in Fig. 3.

Across the entire dataset (458,852 breasts), malignant findings were present in 985 breasts (0.21%) and benign findings in 5,556 breasts (1.22%). All findings were confirmed by at least one biopsy performed within 120 days of the screening mammogram. For the remaining screening exams that were not matched with a biopsy, we assigned labels corresponding to the absence of malignant and benign findings in both breasts. In each exam, the two views of the same breast share the same label.

For all exams matched with biopsies, we asked a group of radiologists (provided with the corresponding pathology reports) to retrospectively indicate the location of the biopsied lesions. This way we obtained the segmentation labels: $\mathbf{M}^b, \mathbf{M}^m \in \{0,1\}^{H \times W}$ where $\mathbf{M}^{b/m}_{i,j} = 1$ if pixel i, j belongs to the benign/malignant findings. An example of such a segmentation is shown in Fig. 3. In all experiments (except for experiments in Section 3.6 that assess the benefits of utilizing segmentation labels), segmentation labels are only used for evaluation. We found that, according to the radiologists, approximately 32.8% of exams were mammographically occult, i.e., the lesions that were biopsied were not visible on mammography, even retrospectively, and were identified using other imaging modalities: ultrasound or MRI.

We split NYUBCS into training, validation and test sets in accordance to the Checklist for Artificial Intelligence in Medical Imaging (Mongan and Moy, 2020). We first sorted the patients according to the date of their latest exam and divided them into disjoint training (first 80%), validation (next 10%) and test (last 10%) sets. This step ensures that each patient only appears in one of the training, validation, and test set. We then retrieved the corresponding exams associated with all patients. For patients in the training and validation sets we utilized all the exams available for each patient; for test patients we dropped all but the latest exam for each test patient. After this procedure there were 186,816, 28,462 and 14,148 exams in the training, validation and test sets respectively.

All images were cropped to 2944×1920 pixels and normalized to have zero mean and unit standard deviation. We adopted the same pre-processing and augmentation (random cropping, size noise) as Wu et al. (2019b). During test phase, we similarly apply data augmentation and average predictions over 10 random augmentations to compute the prediction for a given image. No data augmentation is used during validation. Since the classes of the images in the dataset are imbalanced, we adopted the following sampling strategy during training. In each epoch, we trained the model using all exams that contain at least one benign or malignant finding and an equal number of randomly sampled negative exams. During the training phase, we also randomly rotated the selected ROI patches by $\{0, 90, 180, 270\}$ degrees with equal probability. No rotation to the patches was applied during validation and test phases.

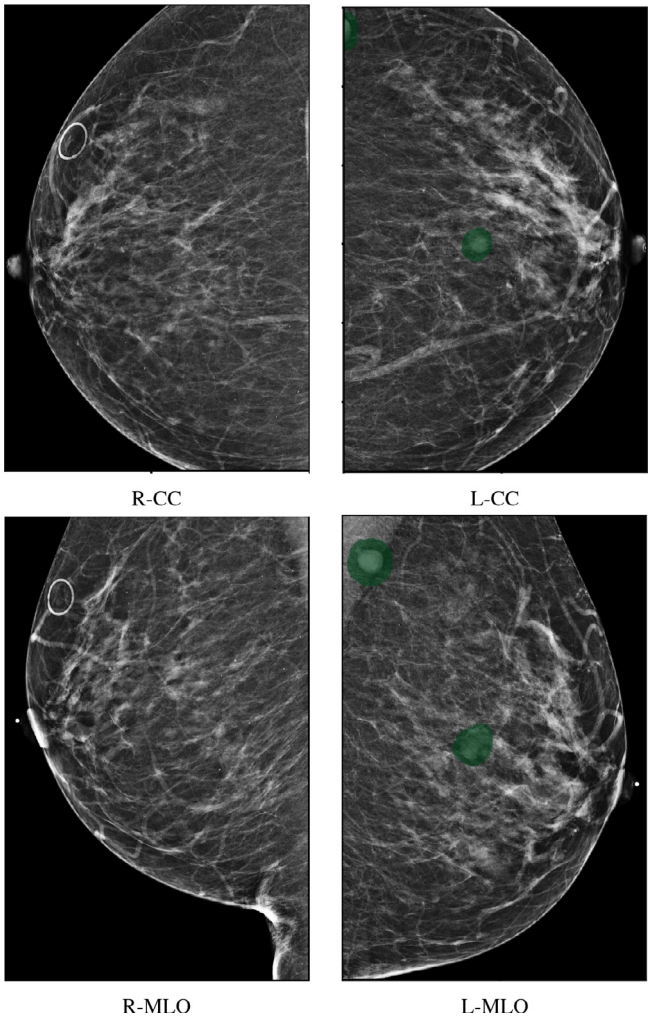

Fig. 3. Example screening mammography exam. Each exam is associated with four images that correspond to the CC and MLO view of both left and right breast. The left breast is diagnosed with benign findings which are highlighted in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CBIS-DDSM The Curated Breast Imaging Subset of DDSM dataset includes mammography images, lesion segmentation, lesion ROI patches, and pathologic diagnosis for 753 breast screening exams with calcification findings and 891 breast screening exams with mass findings. Each mammography image is associated with two binary labels indicating the presence of any benign and malignant lesions. For a fair comparison, in all experiments, we only utilize the entire mammography images with the image-level cancer labels for training our model. We refer the readers to Lee et al. (2017) for more details about this public dataset.

For the purpose of comparison to Zhu et al. (2017) and Shu et al. (2020), we applied the standardized splitting provided in Lee et al. (2017) (85% for training/validation and 15% for testing). We further randomly split training-validation subset into a training subset (80%) and a validation subset (20%). All models

[3] Our retrospective study was approved by our institutional review board and was compliant with the Health Insurance Portability and Accountability Act. Informed consent was waived.

were trained using only the training subset. The validation subset was used for hyperprameter tuning and model selection.

To preprocess mammography images in CBIS-DDSM, we first found the largest connected component containing only non-zero pixels to locate the breast. We then applied erosion and dilation to refine the breast margin. Lastly, we re-oriented all mammography images so that the breasts are always on the left side of the image. All images are resized to 2944×1920 pixels and pixels values were normalized to the range [0,1]. During training, we use image augmentations including random horizontal flipping (p=0.5), random rotation (-15° to 15°), random translation (up to 10% of image size), scaling by a random factor between 0.8 and 1.6, random shearing (-25° to 25°), and pixel-wise Gaussian noise ($\mu = 0$, $\sigma = 0.005$).

3.2. Evaluation metrics

To measure classification performance, we report area under the ROC curve (AUC), on the breast-level for NYUBCS and, for consistency with prior work, on the image-level for CBIS-DDSM. As each breast is associated with two images (CC and MLO views) and our model generates a prediction for each image, we define breast-level predictions as the average of the two image-level predictions. In the reader study, we also used area under the precision-recall curve (PRAUC) to compare radiologists and the proposed model. We computed the radiologists sensitivity which served as a threshold to derive the specificity of GMIC. To assess statistical significance, we computed binomial proportion confidence intervals for specificity. To quantitatively evaluate our model's localization ability, we calculate the Dice similarity coefficient (DSC) and pixel average precision (PxAP) proposed by Choe et al. (2020). Both the DSC and PxAP values we report are computed as an average over images for which segmentation labels are available (i.e. images from breasts which have biopsied findings which were not mammographically occult).

In addition to accuracy, computation time and memory efficiency are also important for medical image analysis. To measure memory efficiency, we report the peak GPU memory usage during training as in Canziani et al. (2016). Similar to Schlemper et al. (2019), we also report the run-time performance by recording the total number of floating-point operations (FLOPs) during inference and elapsed time for forward and backward propagation. Both memory and run-time statistics were measured by benchmarking each model on a single exam (4 images), averaged across 100 exams. All experiments are conducted on an NVIDIA Tesla V100 GPU.

3.3. Classification performance

3.3.1. NYU Breast Cancer Screening Dataset

Implementation details

We parameterize f_g as a ResNet-22 whose architecture is shown in Fig. 2. We pretrain f_g on BI-RADS labels as described in Geras et al. (2017) and Wu et al. (2019b). For f_l, we experiment with three different architectures with varying levels of complexity (ResNet-18, ResNet-34, ResNet-50). We extract $K = 6$ ROI patches from each image. In all experiments (except the ablation study in Section 3.6), we only used image-level labels to train GMIC. In all experiments, the training loss is optimized using Adam (Kingma and Ba, 2014) with learning rate fine-tuned as described in Section 3.3. Our PyTorch (Paszke et al., 2017) implementation (the code and the trained weights of the model) is available at https://github.com/nyukat/GMIC.

Baselines The proposed model is compared against three baselines. We first trained ResNet-34 (He et al., 2016a). ResNet-34 is the highest capacity model among the ResNet architectures that can process a mammography image in its original resolution while fitting in the memory of an NVIDIA Tesla V100 GPU. We also experimented with a variant of ResNet-34 (ResNet-34-1×1 conv) by replacing the fully connected classification layer with a 1×1 convolutional layer and *top t% pooling* as the aggregation function. In addition, we compared our model with Deep Multi-view CNN (DMV-CNN) proposed by Wu et al. (2019b) which has two versions. In the vanilla version, DMV-CNN applies a ResNet-based model on four standard views to generate two breast-level predictions for each exam. DMV-CNN can also be enhanced with pixel-level heatmaps generated by a patch-level classifier, which requires hand-annotated segmentation labels during training. Lastly, we also compared GMIC with the work of Févry et al. (2019) which used a model based on Faster R-CNN (Ren et al., 2015) that utilizes segmentation labels to localize anchor boxes that correspond to malignant or benign lesions. Unlike DMV-CNN and Faster R-CNN which rely on segmentation labels, GMIC can be trained with only image-level labels.

Hyperparameter tuning To make a fair comparison between model architectures, we optimize the hyperparameters with random search (Bergstra and Bengio, 2012) for both ResNet-34 baselines and GMIC. Specifically, for all models, we search for the learning rate $\eta \in 10^{[-5.5,-4]}$ on a logarithmic scale. Additionally, for GMIC and ResNet-34 with 1×1 filters in the last convolutional layer, we also search for the regularization weight $\beta \in 10^{[-5.5,-3.5]}$ (on a logarithmic scale) and for the pooling threshold $t \in \{1\%, 3\%, 5\%, 10\%, 20\%\}$. For all models, we train 30 separate models using hyperparameters randomly sampled from ranges described above. Each model is trained for 50 epochs, and we report the test performance using the weights from the training epoch that achieves highest validation performance.

Performance For each network architecture, we selected the top five models (referred to as *top-5*) from the hyperparameter tuning phase that achieved the highest validation AUC in identifying breasts with malignant findings and evaluated their performance on the held-out test set. In Table 1, we report the mean and the standard deviation of AUC for the *top-5* models in each network architecture. In general, the GMIC model outperformed all baselines. In particular, GMIC achieved higher AUC than Faster R-CNN and DMV-CNN (with heatmaps), despite GMIC not learning with pixel-level labels. We hypothesize that GMIC's superior performance is related to its ability to efficiently integrate both global features and local details. In Section 3.6, we empirically investigate this hypothesis with multiple ablation studies. Separately, we also observe that increasing the complexity of f_l brings a small improvement in AUC.

To further improve our results, we employed the technique of model ensembling (Dietterich, 2000). Specifically, we averaged the predictions of the *top-5* models for GMIC-ResNet-18, GMIC-ResNet-34, and GMIC-ResNet-50 to produce the overall prediction of the ensemble. Our best ensemble model achieved an AUC of 0.930 in identifying breasts with malignant findings.

In addition, GMIC is efficient in both run-time complexity and memory usage. Compared to ResNet-34, GMIC-ResNet-18 has 28.8% fewer parameters, uses 78.43% less GPU memory, is 4.1x faster during inference and 5.6x faster during training. GMIC achieved even more prominent superiority in both run-time and GPU memory usage compared to Faster R-CNN. This improvement is brought forth by its design that avoids excessive computation on the whole image while selectively focusing on informative regions.

3.3.2. CBIS-DDSM dataset

Implementation details We parameterize f_g as ResNet-18 with initial weights pretraiend on ImageNet (Deng et al., 2009). We experimented ResNet-18, ResNet-34, and ResNet-50 for f_l. Unlike NYUBCS, CBIS-DDSM does not include any exams without benign

Table 1
Comparison of performance of GMIC and the baselines on NYUBCS. For both GMIC and ResNet-34, we reported test AUC (mean and standard deviation) of *top-5* models that achieved highest validation AUC in identifying breasts with malignant findings. We also measure the total number of learnable parameters in millions, peak GPU memory usage (Mem) for training a single exam (4 images), time taken for forward (Fwd) and backward (Bwd) propagation in milliseconds, and number of floating-point operations (FLOPs) in billions.

Model	AUC(M)	AUC(B)	#Param	Mem(GB)	Fwd/Bwd (ms)	FLOPs
ResNet-34	0.736 ± 0.026	0.684 ± 0.015	21.30M	13.95	189/459	1622B
ResNet-34-1 $\times$ 1 conv	0.889 ± 0.015	0.772 ± 0.008	21.30M	12.58	201/450	1625B
DMV-CNN (w/o heatmaps)	0.827 ± 0.008	0.731 ± 0.004	6.13M	2.4	38/86	65B
DMV-CNN (w/ heatmaps)	0.886 ± 0.003	0.747 ± 0.002	6.13M	2.4	38/86	65B
Faster R-CNN	0.908 ± 0.014	0.761 ± 0.008	104.8M	25.75	920/2019	–
GMIC-ResNet-18	0.913 ± 0.007	0.791 ± 0.005	15.17M	3.01	46/82	122B
GMIC-ResNet-34	0.909 ± 0.005	0.790 ± 0.006	25.29M	3.45	58/94	180B
GMIC-ResNet-50	$\mathbf{0.915} \pm 0.005$	$\mathbf{0.797} \pm 0.003$	27.95M	5.05	66/131	194B
GMIC-ResNet-18-ensemble	**0.930**	0.800	–	–	–	–
GMIC-ResNet-34-ensemble	0.920	0.795	–	–	–	–
GMIC-ResNet-50-ensemble	0.927	**0.805**	–	–	–	–

or malignant findings. All breasts in CBIS-DDSM contain either benign or malignant lesions. To be consistent with the baseline approaches, we adopted a binary classification framework and only computed the probability for the presence of malignant findings. We adopted the same setting for hyperparameter tuning as for the NYUBCS.[4]

Baselines We compared GMIC to ResNet-34, ResNet-34-1$\times$1 conv, Deep MIL (Zhu et al., 2017), and two other models based on Deep MIL with more elaborate pooling mechanism proposed by Shu et al. (2020).[5] Deep MIL consists of a CNN applied on downsampled mammography images, followed by a multiple instance learning (MIL) pooling layer to aggregate predictions from all spatial positions. Shu et al. (2020) further extended Deep MIL with two new pooling mechanisms: region-based group-max pooling (RGP) and global group-max pooling (GGP) to address the variability of lesion size. To make the comparison to GMIC fair, for ResNet-34 and ResNet-34-1$\times$1 conv, we used the training and hyperparameter search procedure described in Section 3.3.1. For Deep MIL, RGP, and GGP, we used the performance reported by Shu et al. (2020).

Performance We report the classification performance on the test set in Table 2. On average, the *top-5* GMIC-ResNet-18 achieved the AUC of 0.833 (std:0.004) in identifying breasts with malignant lesions. This result is on par with the two state-of-the-art approaches. Moreover, we observed that increasing the complexity of f_l does not improve the classification performance, which is consistent with our observation in Section 3.3.1. In addition, we similarly applied model ensembling as with NYUBCS which further improves GMIC's performance (AUC = 0.858). In summary, the classification performance on CBIS-DDSM further confirms the generalizability of GMIC.

3.4. Reader study

Organization To evaluate the potential clinical impact of our model, we compared the performance of GMIC to the performance of radiologists using data from the reader study conducted by

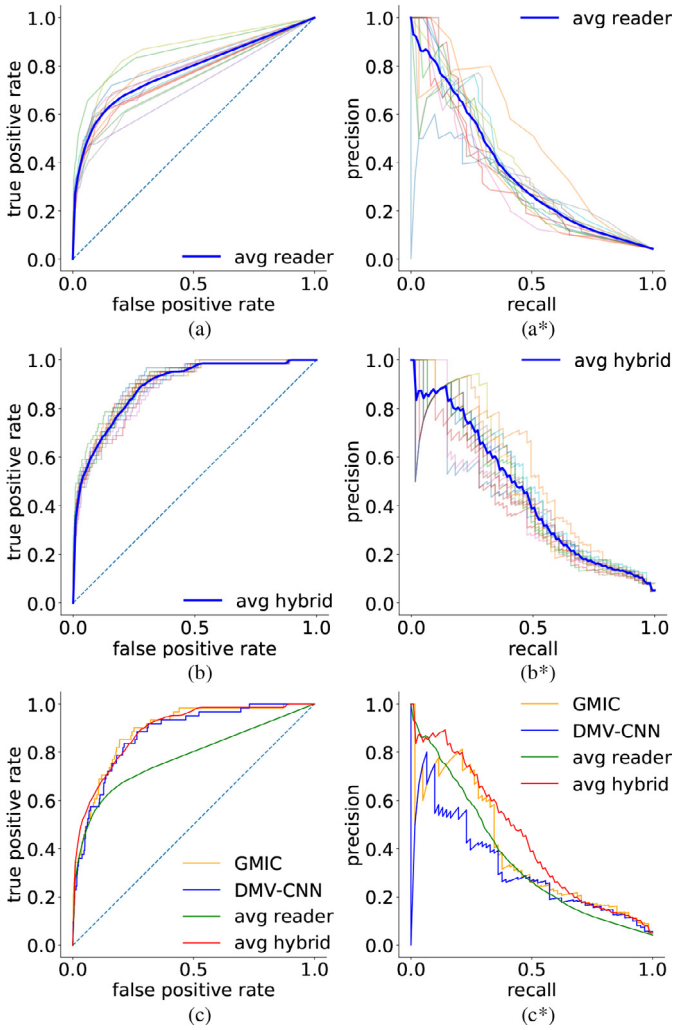

Fig. 4. The ROC curves ((a), (b), (c)) and the precision-recall curves ((a*), (b*), (c*)) computed on the reader study dataset. (a) & (a*): curves for all 14 readers. We derive the ROC/PRC for the average reader by computing the average true positive rate and precision across all readers for every false positive rate and recall. (b) & (b*): curves for hybrid models with each single reader. The curve highlighted in blue indicates the average performance of all hybrids. (c) & (c*): comparison among the GMIC, DMV-CNN, the average reader, and average hybrid. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[4] The implementation of Faster R-CNN by Févry et al. (2019) is not compatible with our framework of FLOPs calculation.

[5] In this work, when using the CBIS-DDSM dataset, we compared GMIC to models that are trained only with image-level labels as this is the scenario for which GMIC is primarily designed for. While existing works demonstrate that incorporating pixel-level segmentations can improve classification performance (Ribli et al., 2018; Shen et al., 2019; Li, Chen, Nailon, Davies, Laurenson), they are evaluated using different subsets of CBIS-DDSM as the test set. This makes direct numerical comparisons to our work, as well as comparisons between them, inappropriate. Therefore, we leave evaluating the utility of pixel-level segmentations for feature work.

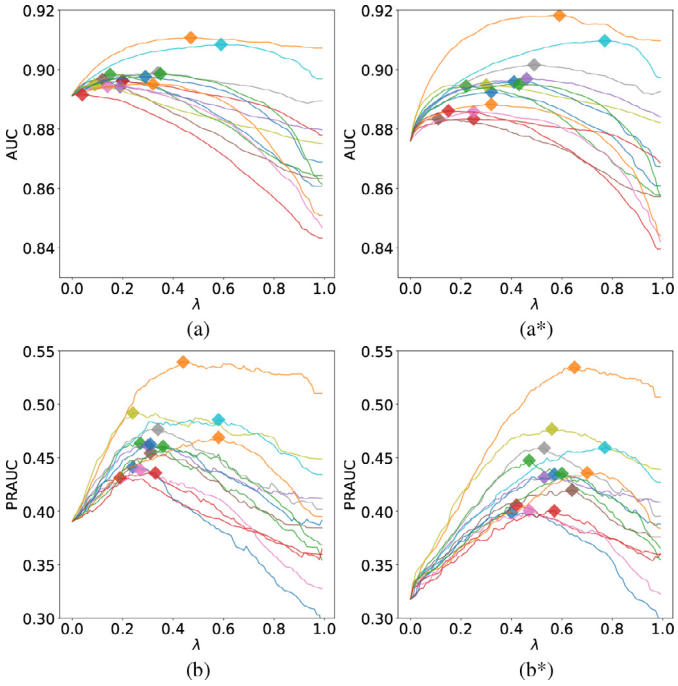

Fig. 5. AUC and PRAUC as a function of $\lambda \in [0, 1)$ for hybrids between each reader and GMIC (left)/DMV-CNN (right) ensemble. Each hybrid achieves the highest AUC/PRAUC for a different λ (marked with $\diamond$).

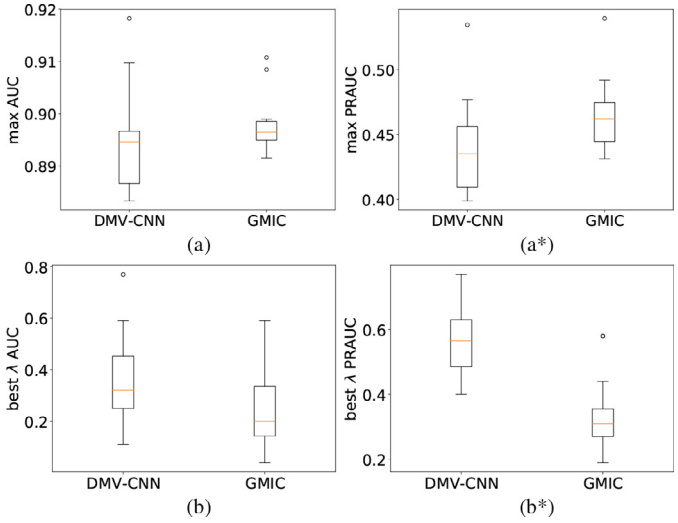

Fig. 6. (a) and (a*): the distribution of maximum AUC/PRAUC achieved for hybrids between each reader and GMIC/DMV-CNN ensemble. (b) and (b*): the distribution of the optimal λ^* that achieves the maximum AUC/PRAUC for both GMIC/DMV-CNN hybrids. GMIC hybrids achieve higher AUC and PRAUC than DMV-CNN hybrids. Moreover, GMIC plays a more important role than DMV-CNN in the hybrid models as indicated by the distribution of λ^*.

Wu et al. (2019b). This study includes 14 readers: 12 attending radiologists at various level of experience (between 2 and 30 years), a medical resident, and a medical student. Each reader was asked to provide probability estimates as well as binary predictions of malignancy for 720 screening exams (1440 breasts). Among the 1440 breasts, 62 breasts were associated with malignant findings and 356 breasts were associated with benign findings. Among the breasts in which there were malignant findings, there were 21 masses, 26 calcifications, 12 asymmetries and 4 architectural distortions. The radiologists were only shown images with no other data.

Table 2
Classification performance on CBIS-DDSM. For GMIC, we reported test AUC of *top-5* models that achieved highest validation AUC in identifying breasts with malignant findings. We compared GMIC with five baselines. The performance of Deep MIL, RGP, and GGP in this table was originally reported in Shu et al. (2020).

Model	AUC(M)
ResNet-34	0.792 ± 0.014
ResNet-34-1×1 conv	0.800 ± 0.011
Deep MIL (Zhu et al., 2017)	0.791 ± 0.0002
RGP (Shu et al., 2020)	0.838 ± 0.0001
GGP (Shu et al., 2020)	0.823 ± 0.0002
GMIC-ResNet-18	0.833 ± 0.004
GMIC-ResNet-18 (best)	0.840
GMIC-ResNet-34	0.830 ± 0.003
GMIC-ResNet-50	0.828 ± 0.001
GMIC-ResNet-18-ensemble	0.858
GMIC-ResNet-34-ensemble	0.849
GMIC-ResNet-50-ensemble	0.849

Table 3
Performance of readers, GMIC, and the hybrid model in the reader study. The specificity of GMIC and hybrid model is computed at readers' average sensitivity level (62.1%). In all metrics, GMIC outperforms the readers and the hybrid model outperforms both the readers and GMIC.

	AUC	PRAUC	specificity
readers	0.779 ± 0.044	0.364 ± 0.05	85.2%
GMIC	0.891	0.39	90%
hybrid	0.892 ± 0.009	0.449 ± 0.036	91.5%

Comparison to radiologists We calculate AUC and PRAUC on the reader study dataset to measure the performance of radiologists and GMIC. We obtain GMIC's predictions by ensembling the predictions of the *top-5* GMIC-ResNet-18 models. In Fig. 4 ((a) and (a*)), we visualize the receiver operating characteristic curve (ROC) and precision-recall curve (PRC) for each individual reader using their probability estimates of malignancy. We also compared GMIC with DMV-CNN and the radiologists ((c) and (c*)). GMIC achieves an AUC of 0.891 and PRAUC of 0.39 outperforming DMV-CNN (AUC: 0.876, PRAUC: 0.318). The AUCs associated with each individual reader ranges from 0.705 to 0.860 (mean: 0.778, std: 0.0435) and the PRAUCs for readers vary from 0.244 to 0.453 (mean: 0.364, std: 0.0496). GMIC achieves a higher AUC and PRAUC than the average reader. We note that there is a limitation associated with AUC and PRAUC. While AUC and PRAUC are calculated on continuous predictions, radiologists are trained to make diagnosis by choosing from a discrete set of BI-RADS scores (D'Orsi, 2013). Indeed, even though the readers were given a possibility to predict any number between 0% and 100%, they chose to stick to the probability threshold corresponding to BI-RADS scores.

To compare GMIC to radiologists, we also use sensitivity and specificity as additional evaluation metrics. We first compute the radiologists' sensitivity and specificity using the data from the reader study. We then use the average specificity and sensitivity among readers as the proxy for radiologists' performance under a single-reader setting and use the statistics of the consensus reading to approximate the performance under a multi-reader setting. The predictions for the consensus reading are derived using majority voting. We summarize the performance of both GMIC and radiologists in Table 3. The 14 radiologists achieved an average specificity of 85.2% (std:5.5%) and average sensitivity of 62.1% (std:9%). The consensus reading yields a specificity of 94.6% and a sensitivity of 76.8%. The performance of the radiologists in the reader study is lower than that for community practice radiologists performance (Lehman et al., 2016) which reported a sensi-

tivity of 86.9% and a specificity 88.9%. However, the overall sensitivity in our study falls within acceptable national performance standards (Lehman et al., 2016) and likely reflects the lack of prior imaging and other clinical data available during interpretation. At the average radiologists' sensitivity level (62.1%), GMIC achieves a specificity of 90% which is higher (P<0.001) than the average radiologists' specificity (85.2%). At the consensus reading sensitivity level (76.8%), GMIC's specificity is 83.6% which is lower than consensus reading specificity (94.6%). While the proposed model underperforms the consensus reading, the results demonstrate the potential value of GMIC as a second reader.

Human-machine hybrid To further demonstrate the clinical potential of GMIC, we create a hybrid model whose predictions are a linear combination of predictions from each reader and the model: $\hat{\mathbf{y}}_{\text{hybrid}} = \lambda\hat{\mathbf{y}}_{\text{reader}} + (1-\lambda)\hat{\mathbf{y}}_{\text{GMIC}}$. We compute the AUC and PRAUC of the hybrid models by setting $\lambda = 0.5$. We note that $\lambda = 0.5$ is not the optimal value for all hybrid models. On the other hand, the performance obtained by retroactively fine-tuning λ on the reader study is not transferable to realistic clinical settings. Therefore, we chose $\lambda = 0.5$ as the most natural way of aggregating two sets of predictions when not having prior knowledge of their quality. In Fig. 4((b) and (b*)), we visualize the ROC and PRC curves of the hybrid models ($\lambda = 0.5$) which on average achieve an AUC of 0.892 (std: 0.009) and an PRAUC of 0.449 (std: 0.036), improving radiologists' mean AUC by 0.114 and mean PRAUC by 0.085. For each of the hybrid models, we also calculate its specificity at the average radiologists' sensitivity (62.1%). The 14 hybrid models achieve an average specificity of 91.5% (std: 1.8%) which is higher than ($P < 0.001$) the average radiologists' specificity (85.2%). These results indicate that our model captures different aspects of the task compared to radiologists and can be used as a tool to assist in interpreting breast cancer screening exams.

In addition, in Fig. 5, we visualize the AUC and PRAUC achieved by combining predictions from each of these 14 readers with GMIC ((a) and (b)) and DMV-CNN ((a*) and (b*)) with varying λ. The diamond mark on each curve indicates the λ^* that achieves the highest AUC/PRAUC. As shown in the plot, the predictions from all radiologists could be improved ($\lambda^* < 1.0$) by incorporating predictions from GMIC. More specifically, as shown in Fig. 6 ((a) and (a*)), with the optimal λ^*, GMIC hybrids achieves a mean AUC of 0.898 ± 0.005 and mean PRAUC of 0.465 ± 0.03 both of which are higher than the counterparts of DMV-CNN hybrids (AUC:0.895 ± 0.01, PRAUC:0.439 ± 0.035). In addition, we compare the distribution of λ^* for GMIC and DMV-CNN. The average value of λ^* associated with GMIC hybrid models to achieve maximum AUC/PRAUC is $0.25 \pm 0.15/0.34 \pm 0.11$ which is lower than DMV-CNN ($0.34 \pm 0.15/0.59 \pm 0.12$). This result shows that, the more accurate the model used in the human-machine hybrid is, the more weight is attached to its predictions.

3.5. Localization performance

To evaluate the localization performance of GMIC, we select the model with the highest DSC for malignancy localization using the validation set. During inference, we upsample saliency maps using nearest neighbour interpolation to match the resolution of the input image. Our best localization model achieved a mean test DSC of 0.325 (std:0.231) for localization of malignant lesions and 0.240 (std:0.175) for localization of benign lesions. The extracted ROI patches correctly indicate the biopsy-confirmed lesions in 78.1% of all annotated images in the test set (a lesion is considered to be correctly indicated by the ROI patches if they cover at least 70% of its pixels). The best localization model achieves an AUC of 0.886 and 0.78 on classifying malignant and benign lesions, respectively. We observe that localization and classification performance are not perfectly correlated. The trade-off between classification and localization has been discussed in the weakly supervised object detection literature (Feng et al., 2017; Sedai et al., 2018; Yao et al., 2018). To estimate an upper bound of localization performance for this dataset, we use the pixel-level segmentations to train a U-Net (Ronneberger et al., 2015). To estimate performance of a model which did not learn anything, we generate random saliency maps whose pixel values were obtained by uniformly sampling values in the range of [0,1]. We summarize the localization performance of GMIC in Table 4. While GMIC underperformed U-Net, it achieved higher DSC and PxAP than random baseline indicating that GMIC provides non-trivial localization on the lesions of interest.

Table 4
Localization performance of GMIC on malignant (M) and benign (B) lesions. We adopt Dice similairty score (DSC) and pixel average precision (PxAP) as evaluation metrics. We compared GMIC with U-Net and a random baseline whose pixel-level predictions are randomly drawn from the standard uniform distribution.

Model	DSC(M)	DSC(B)	PxAP(M)	PxAP(B)
GMIC	0.325 ± 0.231	0.240 ± 0.175	0.396 ± 0.275	0.283 ± 0.244
U-Net	0.504 ± 0.283	0.412 ± 0.316	0.589 ± 0.329	0.498 ± 0.357
random	0.039 ± 0.044	0.021 ± 0.030	0.012 ± 0.015	0.006 ± 0.010

In Fig. 7, we visualize saliency maps for four samples selected from the test set. In the first two examples, the saliency maps are highly activated on the annotated lesions, suggesting that our model is able to detect suspicious lesions without pixel-level supervision. Moreover, the attention α_k is highly concentrated on ROI patches that overlap with the annotated lesions. In the third example, the saliency map for benign findings identifies three abnormalities. Although only the top abnormality was escalated for biopsy and hence annotated by radiologists, the radiologist's report confirms that the two non-biopsied findings have a high probability of benignity and a low probability of malignancy. In the fourth example, we illustrate a case when there is some level of disagreement between our model and the annotation in the dataset. The malignancy saliency map only highlights part of a large malignant lesion with segmental coarse heterogeneous calcifications. This behavior is related to the design of f_{agg}: a fixed pooling threshold t cannot be optimal for all sizes of ROI. The impact of f_{agg} is further studied in 3.6. This example also illustrates that while human experts are asked to annotate the entire lesion, CNNs tend to emphasize only the most informative regions. While no benign lesion is present, the saliency map of benign findings still highlights regions similar to that in the malignancy saliency map, but with a lower probability than the malignancy saliency map. In fact, calcifications with this morphology and distribution can also result from benign pathophysiology (Liberman and Menell, 2002). We provide additional visualization of both successful and failed localization of benign and malignant lesions in Supplementary Figs. 14–16.

In addition, we observe that GMIC is able to provide meaningful localization when the lesions are hardly visible to radiologists in the image. In Fig. 8, we illustrate a mammographically occult mammogram of a 59-year old patient with no family history of breast cancer and dense breasts. There is an asymmetry in the left lateral breast posterior depth which appears stable compared to prior mammograms and was determined to be benign by the reading radiologist. However, the saliency map of malignant findings successfully identifies the malignant lesion on the screening mammogram. Same day screening ultrasound (sagittal image) demonstrated a 1.2 cm irregular mass; ultrasound biopsy yielded moderate grade invasive ductal carcinoma.

3.6. Ablation study

We performed ablation studies to explore the effectiveness of global module, local module, fusion module, patch-level attention,

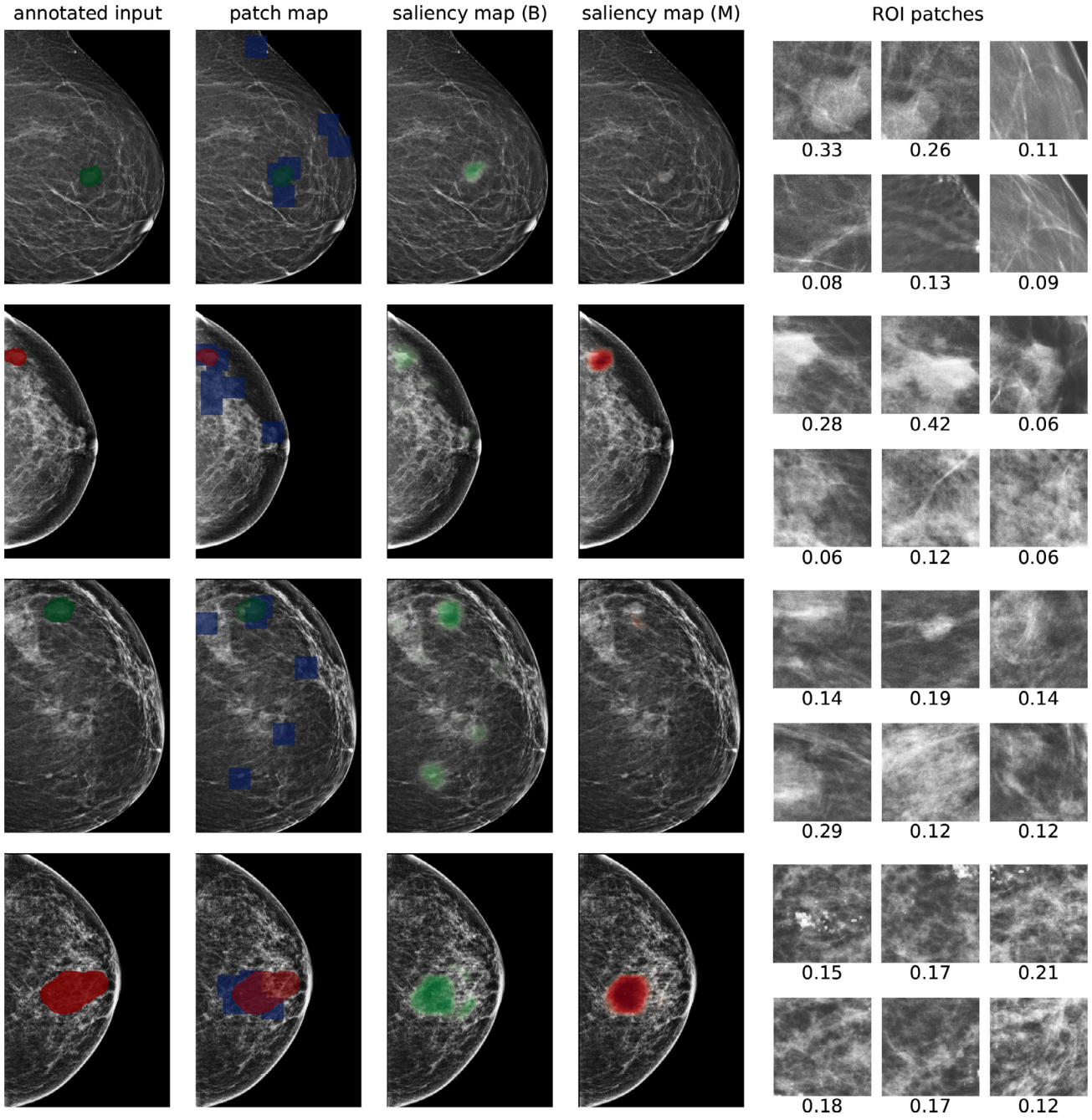

Fig. 7. Visualization of results for four examples. From left to right: input images annotated with segmentation labels (green = benign, red = malignant), locations of ROI patches (blue squares), saliency map for benign class, saliency map for malignant class, and ROI patches with their attention scores. The top example contains a circumscribed oval mass in the left upper breast middle depth which was diagnosed as a benign fibroadenoma by ultrasound biopsy. The second example contains an irregular mass in the right lateral breast posterior depth which was diagnosed as an invasive ductal carinoma by ultrasound biopsy. In the third example, the saliency map of benign findings identifies (from up to bottom) (a) a circumscribed oval mass in the lateral breast middle depth, (b) a smaller circumscribed oval mass in the media breast, and (c) an asymmetry in the left central breast middle depth. Ultrasound-guided biopsy of the finding shown in (a) yielded benign fibroadenoma. The medial breast mass (b) was recommended for short-term follow-up by the breast radiologist. The central breast asymmetry (c) was imaging-proven stable on multiple prior mammograms and benign. The bottom example contains segmental coarse heterogeneous calcifications in the right central breast middle depth. Stereotactic biopsy yielded high grade ductal carcinoma in situ. We provide additional visualizations of exams with benign and malignant findings in Supplementary Figs. 14–16. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the proposed *top t% pooling*. In addition, we also assess how much performance of GMIC could be improved by utilizing the pixel-level labels and ensembling GMIC with DMV-CNN and Faster R-CNN. All ablation experiments are based on the GMIC-ResNet-18 model.

Synergy of global and local information In the preliminary version of GMIC (Shen et al., 2019), the final prediction is defined as $\frac{1}{2}(\hat{\mathbf{y}}_{\text{global}} + \hat{\mathbf{y}}_{\text{local}})$. In this work, we enhance GMIC with a fusion module that combines signals from both global features and local details. To empirically evaluate the effectiveness of the fusion module, we compared the performance achieved using only global features ($\hat{\mathbf{y}}_{\text{global}}$), only local patches ($\hat{\mathbf{y}}_{\text{local}}$), the average prediction of two modules ($\frac{1}{2}(\hat{\mathbf{y}}_{\text{global}} + \hat{\mathbf{y}}_{\text{local}})$), and the fusion of the two ($\hat{\mathbf{y}}_{\text{fusion}}$). As shown in Table 5, $\hat{\mathbf{y}}_{\text{fusion}}$ achieved a higher AUC consistently for

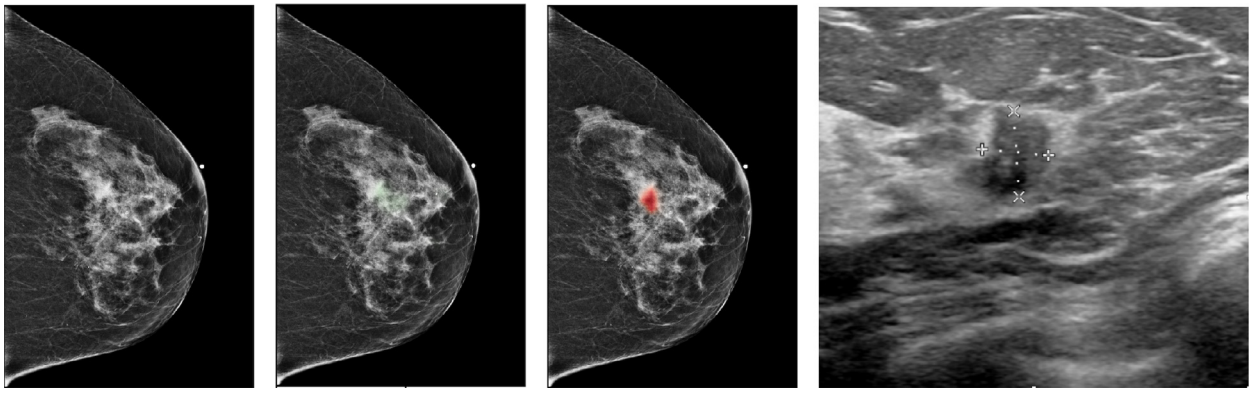

Fig. 8. A mammographically occult example with a biopsy-proven malignant finding. From left to right: the original image, the saliency map for benign findings, the saliency map for malignant findings, and the sagittal ultrasound image of this patient. While the asymmetry in the left lateral breast posterior depth was intepreted as benign by the radiologist, a subsequent screening ultrasound and ultrasound-guided biopsy yielded mammographically-occult moderate grade invasive ductal carcinoma. On saliency maps, this area shows a weak probability of benignity and a high probability of malignancy.

Table 5
Ablation study: effectiveness of incorporating both global and local features. We report the mean and standard deviation of the test AUC for *top-5* GMIC-ResNet-18. We experimented with 4 GMIC variants that use $\hat{\mathbf{y}}_{\text{global}}$, $\hat{\mathbf{y}}_{\text{local}}$, the average of $\hat{\mathbf{y}}_{\text{global}}$ and $\hat{\mathbf{y}}_{\text{local}}$, and $\hat{\mathbf{y}}_{\text{fusion}}$ as predictions. The proposed design that uses $\hat{\mathbf{y}}_{\text{fusion}}$ as predictions outperforms all variants.

Prediction	AUC(M)	AUC(B)
$\hat{\mathbf{y}}_{\text{global}}$	0.892 ± 0.009	0.776 ± 0.004
$\hat{\mathbf{y}}_{\text{local}}$	0.897 ± 0.004	0.778 ± 0.005
$\frac{1}{2}(\hat{\mathbf{y}}_{\text{local}} + \hat{\mathbf{y}}_{\text{global}})$	0.905 ± 0.006	0.785 ± 0.004
$\hat{\mathbf{y}}_{\text{fusion}}$	$\mathbf{0.913} \pm 0.007$	$\mathbf{0.791} \pm 0.005$

Table 6
To evaluate the effectiveness of the patch-wise attention, we compare the proposed model with the variant (uniform) that always assigns equal attention to all patches. To investigate the importance of the localization information in the saliency maps, we trained another variant (random) that randomly selects patches from the input image. We use GMIC-ResNet-18 model with *top 3% pooling* as the base model. The performance of the local module ($\hat{\mathbf{y}}_{\text{local}}$) is reported.

Attention	ROI patches	AUC(M)	AUC(B)
uniform	`retrieve_roi`	0.874 ± 0.008	0.776 ± 0.007
gated	random	0.629 ± 0.042	0.658 ± 0.011
gated	`retrieve_roi`	$\mathbf{0.898} \pm 0.01$	$\mathbf{0.78} \pm 0.008$

classifying both benign and malignant lesions than either $\hat{\mathbf{y}}_{\text{global}}$ or $\hat{\mathbf{y}}_{\text{local}}$. This result suggests that the fusion module helps GMIC to aggregate signals from both global and local module. Moreover, $\hat{\mathbf{y}}_{\text{fusion}}$ also outperforms the ensemble prediction $\frac{1}{2}(\hat{\mathbf{y}}_{\text{local}} + \hat{\mathbf{y}}_{\text{global}})$, which further demonstrates that the fusion module promotes an effective synergy beyond an ensembling effect created from averaging predictions over two sets of parameters.

ROI proposals and patch-wise attention GMIC applies two mechanisms to control the quality of patches provided to the local module. First, the `retrieve_roi` algorithm utilizes localization information from the saliency maps and greedily selects informative patches of the input image. Those selected patches are then weighted using the Gated Attention network. To evaluate the effectiveness of both mechanisms, we trained two variants: one (uniform) that always assigns equal attention score to each patch and another (random) that randomly samples patches without using the saliency map. As shown in Table 6, if patch-wise attention is disabled, the AUC of classifying malignant lesions decreases from 0.898 to 0.874. If the `retrieve_roi` algorithm is replaced with random sampling, the local module suffers from a significant performance decrease. These results suggest that both the patch-wise attention and `retrieve_roi` procedure are essential for the local module to make accurate predictions.

Table 7
Ablation study: effect of different choice of aggregation function. We report the performance achieved by parameterizing f_{agg} as global average pooling (GAP), global maximum pooling (GMP), and *top t% pooling*. For each setting, we trained five GMIC-ResNet-18 models and report the mean and standard deviation of AUC and DSC.

f_{agg}	AUC(M)	AUC(B)	DSC(M)	DSC(B)
GMP	0.890 ± 0.02	0.785 ± 0.012	0.127 ± 0.052	0.103 ± 0.060
$t = 1\%$	0.906 ± 0.01	0.784 ± 0.007	0.190 ± 0.030	0.147 ± 0.053
$t = 2\%$	$\mathbf{0.916} \pm 0.009$	0.790 ± 0.007	0.203 ± 0.013	0.191 ± 0.042
$t = 3\%$	0.913 ± 0.007	$\mathbf{0.791} \pm 0.004$	$\mathbf{0.228} \pm 0.036$	0.178 ± 0.041
$t = 5\%$	0.912 ± 0.009	0.790 ± 0.002	0.172 ± 0.004	$\mathbf{0.194} \pm 0.027$
$t = 10\%$	0.914 ± 0.005	$\mathbf{0.791} \pm 0.008$	0.156 ± 0.050	0.182 ± 0.028
$t = 20\%$	0.907 ± 0.017	0.785 ± 0.008	0.126 ± 0.048	0.182 ± 0.040
GAP	0.903 ± 0.02	0.783 ± 0.012	0.065 ± 0.006	0.181 ± 0.011

Aggregation function In order to study the the impact of the aggregation function, we experimented with 8 parameterizations of f_{agg} including GAP, GMP, and *top t% pooling* with $t \in \{1, 2, 3, 5, 10, 20\}$. For each parameterization, we fixed other hyperparameters and trained five GMIC-ResNet-18 models with randomly initialized weights. In Table 7, we report the AUC and DSC achieved by each value of t. GMIC-ResNet-18 achieves the highest AUC on identifying malignant cases when using *top t% pooling* with $t = 2$. The performance of *top t% pooling* decreases as t moves away from 2 and converges to that of GAP/GMP when t is large/small. This observation is consistent with the intuition that GAP and GMP are two extremes of *top t% pooling*. We observe a similar but less pronounced trend on the AUC of identifying benign cases.

GMIC-ResNet-18 also obtains better localization performance with *top t% pooling* than with GAP or GMP. The highest DSC for localizing malignant and benign lesions is achieved when t is set to 3% and 5% respectively. To further study the effect of t, we visualize the saliency maps for four examples selected from the test set. As illustrated in Fig. 9, when t is small, the saliency maps tend to highlight a small area. When t is large, the highlighted region grows. Ideally, the choice of t should reflect the true size of lesions contained in the image and different images could use different t. In future research, we propose to learn t using information within the image.

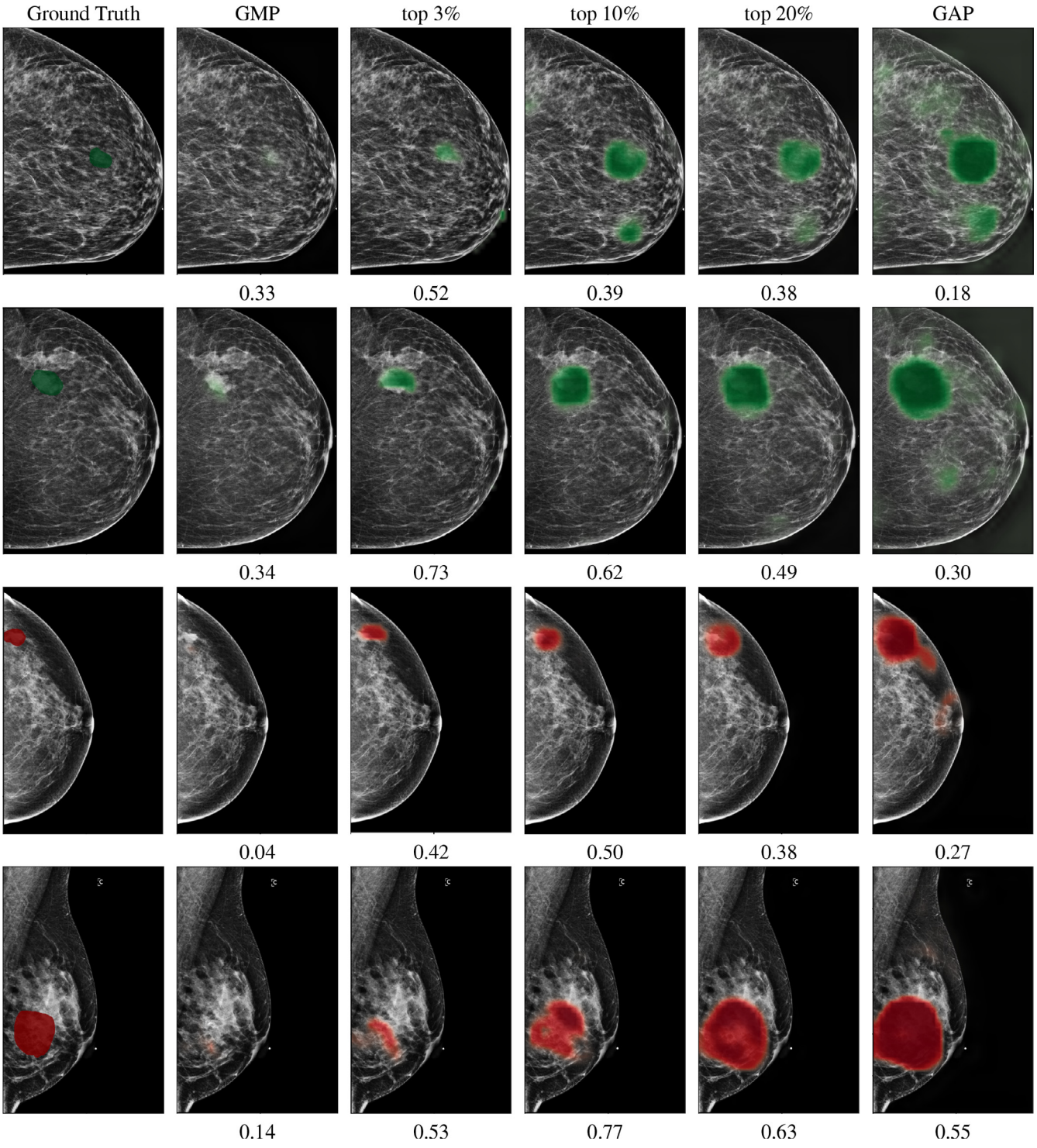

Fig. 9. In this figure we illustrate the effect of t in the pooling function on the saliency maps. From left to right: the mammogram with ground truth segmentation and the saliency map generated using GMP, top 3% pooling, top 10% pooling, top 20% pooling, and GAP. The corresponding DSC is specified below each saliency map. A benign lesion is found in the top two examples. A malignant lesion is found in the bottom two examples.

Number of ROI patches

We experimented with GMIC varying the number of patches $K \in \{1, 2, 3, 4, 6, 8, 10\}$. For each setting, we trained five GMIC-ResNet-18 models with *top t% pooling* ($t = 3\%$). In Fig. 10, we illustrate the mean and the standard deviation of AUC achieved by $\hat{\mathbf{y}}_{\text{fusion}}$ and $\hat{\mathbf{y}}_{\text{local}}$ on classifying benign and malignant lesions. Increasing K improves the classification performance when K is small. The improvement is more evident on $\hat{\mathbf{y}}_{\text{local}}$ than $\hat{\mathbf{y}}_{\text{fusion}}$, because $\hat{\mathbf{y}}_{\text{fusion}}$ also utilizes global features. However, for $K > 3$, the classification performance saturates. This observation demonstrates a trend of diminishing marginal return of incorporating additional ROI patches.

Utilizing segmentation labels

We also assessed how much performance of GMIC could be improved by utilizing pixel-level labels during training. Following Wu et al. (2019b), we used the pixel-level labels to train

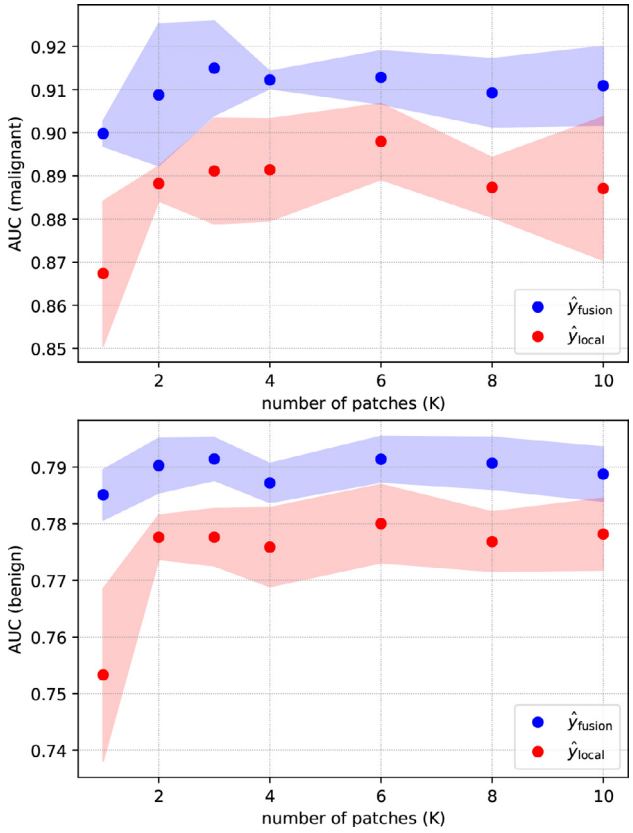

Fig. 10. The classification performance of GMIC-ResNet-18 with a varying number of patches $K \in \{1, 2, 3, 4, 6, 8, 10\}$. For each K, we trained five models and reported the mean and the standard deviation of test AUC on classifying malignant (top) and benign (bottom) lesions. We show the performance of both $\hat{\mathbf{y}}_{\text{fusion}}$ and $\hat{\mathbf{y}}_{\text{local}}$. The performance saturates for $K > 3$.

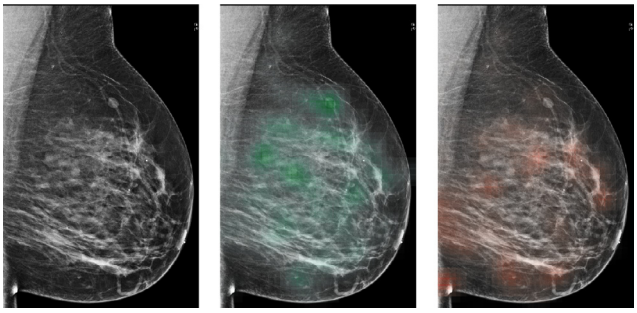

Fig. 11. Example heatmaps generated by the patch-level model proposed by Wu et al. (2019b). The original image (left), the "benign" heatmap over the image (middle), and the "malignant" heatmap over the image (right).

a patch-level model which classifies 256×256-pixel patches of mammograms, making two predictions: the presence or absence of malignant and benign findings in a given patch. We then apply the patch-level classifier to each full-resolution image in a sliding window fashion to create two heatmaps (illustrated in Fig. 11), one containing an estimated probability of a malignant finding for each pixel, and the other containing an estimated probability of a benign finding. In this comparison study, we concatenated the input images with these two heatmaps[6] to train 30 GMIC-ResNet-18 models (referred as GMIC-ResNet-18-heatmap models) using the hyperparameter optimization setting described in Section 3.3. We reported the test performance of the *top-5* GMIC-heatmap models that achieved the highest validation AUC on identifying breasts with malignant lesions. The *top-5* GMIC-ResNet-18-heatmap models achieved a mean AUC of 0.927 ± 0.04 / 0.792 ± 0.008 in identifying breasts with malignant/benign lesions, outperforming the vanilla GMIC models (0.913 ± 0.007 / 0.791 ± 0.005). The ensemble of the *top-5* GMIC-ResNet-18-heatmap models achieved an AUC of 0.931/0.80 in identifying breasts with malignant/benign lesions matching the performance of vanilla GMIC models (0.930/0.80). While augmenting GMIC with heatmaps improves its classification performance, the improvement is marginal especially when comparing to the ensemble of models. We conjecture that, for a sufficiently large dataset, image-level labels alone are powerful enough to capture most of the signal, and additional localization information from the pixel-level segmentation labels only slightly improves the performance of GMIC. In fact, sometimes it might even be biasing the model towards ignoring mammographically-occult findings.

Ensembling GMIC with other models In order to estimate a lower bound of what level of performance is possible to achieve on this task, we build a large "super-ensemble" of models by aggregating the predictions of: (a) an ensemble of *top-5* GMIC-ResNet-18, (b) an ensemble of 5 DMV-CNN model (with heatmaps) (Wu et al., 2019b), and (c) an ensemble of 3 Faster R-CNN models (Févry et al., 2019). Similar to the human-machine hybrid model, the predictions of the ensemble model are defined as $\hat{\mathbf{y}}_{\text{ensemble}} = \lambda_1 \hat{\mathbf{y}}_{\text{GMIC}} + \lambda_2 \hat{\mathbf{y}}_{\text{Faster R-CNN}} + \lambda_3 \hat{\mathbf{y}}_{\text{DMV-CNN}}$ where $\lambda_1 + \lambda_2 + \lambda_3 = 1$. On the test set, the ensemble model with equal weights associated with each of its components ($\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$) achieves an AUC of 0.936 in identifying breasts with malignant lesions. We note that the improvement against *top-5* GMIC-ResNet-18-ensemble (0.930) is small. We also note that utilizing this ensemble might be impractical, due to its complexity and computational cost.

We also checked what would be the AUC of this ensemble if we could tune the weighting coefficients of the ensemble on the test set. In Fig. 12, we visualize its classification performance on the reader study dataset and the full test set for different combinations of λ_1, λ_2 and λ_3. For the optimal combinations of λ_1, λ_2, and λ_3 that achieve the highest AUC on both datasets, the weight associated with GMIC (λ_1) is the largest, however, the two other weights are also non-negligible, suggesting that the three types of models are complementary, even though the improvement in terms of AUC is small.

4. Related work

4.1. High-resolution 2D medical image classification

The increased resolution level of medical images has posed new challenges for machine learning. Early works on applying deep neural networks to medical image classification typically utilize a CNN acting on the entire image to generate a prediction, resembling approaches developed for object classification in natural images. For instance, Roth et al. (2015) adopted a 5-layer CNN to perform anatomical classification of CT slices. A similar approach was adopted by Codella et al. (2015) to recognize melanoma on dermoscopy images. More recently, Rajpurkar et al. (2017) fine-tuned a 121-layer DenseNet (Huang et al., 2016) to classify thorax disease on chest X-ray images. However, this line of work suffers from two drawbacks. Unlike many natural images in which ROIs are sufficiently large, ROIs in medical images are typically small and sparsely distributed over the image. Applying a CNN indiscriminately over the entire image may include a considerable level of noise outside the ROI. Moreover, input images are commonly downsampled to fit in GPU memory. Aggressively downsampling

[6] The two heatmap channels are only used by the global network f_g. The local network f_l does not use them.

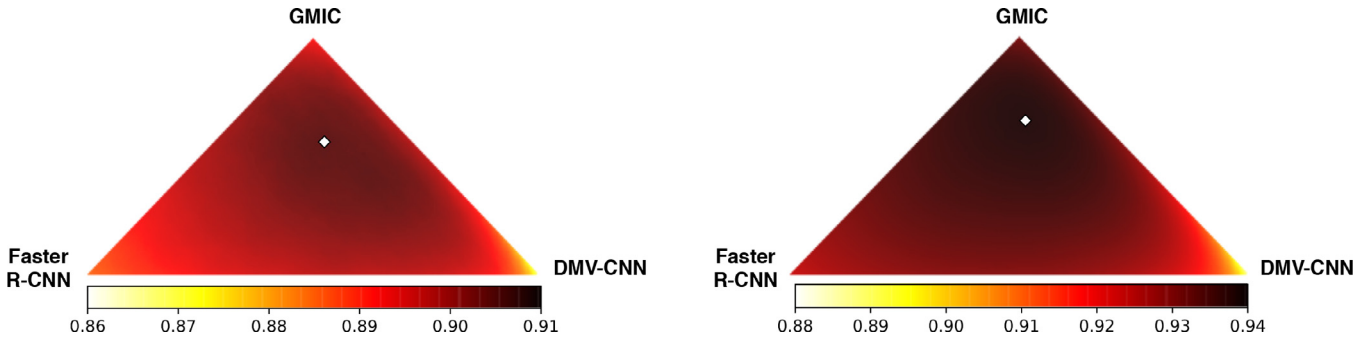

Fig. 12. We visualize the AUC of identifying breasts with malignant findings achieved by the ensemble model with varying λ_1, λ_2, and λ_3 on the reader study dataset (left) and the test set (right). The optimal combination of λ_1, λ_2, and λ_3 that achieves highest AUC is highlighted in white diamond. The weight associated with GMIC is the largest among the three models for both datasets. On the reader study dataset, the optimal combination ($\lambda_1 = 0.56$, $\lambda_2 = 0.2$, $\lambda_3 = 0.24$) achieves an AUC of 0.905. On the test set, the optimal combination ($\lambda_1 = 0.65$, $\lambda_2 = 0.16$, $\lambda_3 = 0.19$) achieves an AUC of 0.939.

medical images could distort important details making the correct diagnosis difficult (Geras et al., 2017).

In another line of research, input images are uniformly divided into small patches. A classifier is trained and applied to each patch, and patch-level predictions are aggregated to form an image-level prediction. This family of methods has been commonly applied to the segmentation and classification of pathology images (Campanella et al., 2019; Sun et al., 2019a,b). Coudray et al. (2018) used Inception V3 (Szegedy et al., 2016) on tiles of whole-slide histopathology images to detect adenocarcinoma and squamous cell carcinoma. Sun et al. (2019) proposed a multi-scale patch-level classifier using dilated convolutions to localize gastric cancer regions. For breast cancer screening, Wu et al. (2019b) utilized patch-level predictions as additional input channels to classify screening mammograms. A major limitation of these methods is that many of them require lesion locations to train the patch-level classifiers, which might be expensive to obtain. Moreover, global information such as the image structure could be lost by dividing input images into small patches.

Instead of applying patch-level model on all tiles, several methods have been proposed to select patches that are related to the classification task. Zhong et al. (2019) suggested selecting important patches based on a coarse attention map generated by applying an U-Net (Ronneberger et al., 2015) on downsampled input images. Guo et al. (2019) adopted a similar strategy to detect strut points on intravascular optical coherence tomography images. Guan et al. (2018) further developed this idea and proposed the attention guided convolution neural network (AG-CNN) that explicitly merges information from both the global image and a refined local patch to detect thorax disease on chest X-ray images. Our work is perhaps most similar to Guan et al. (2018). While AG-CNN only selects one patch for each class, our method is able to selectively aggregate information from a variable number of patches, which enables the model to learn from broader source of signal.

4.2. Breast cancer classification in mammography

Early works on breast cancer screening exam classification were computer-aided detection (CAD) systems built with handcrafted features (Li et al., 2001; Wu et al., 2007; Masotti et al., 2009; Oliver et al., 2010). Despite their popularity, clinical study has suggested that CAD systems do not improve diagnostic accuracy (Lehman et al., 2015). With the advances in deep learning in the last decade (LeCun et al., 2015), neural networks have been extensively applied to assist radiologists in interpreting screening mammograms (McKinney et al., 2020; Rampun et al., 2019; Wu et al., 2018; Zhu et al., 2017). In particular, Geras et al. (2017) adopted a multi-view CNN that jointly utilizes information from four standard views to classify the BI-RADS category associated with mammograms. To accurately detect small lesions on mammograms, segmentation labels have been utilized to train patch-level classifiers (Lotter et al., 2017; Kooi and Karssemeijer, 2017; Shen, 2017; Teare et al., 2017; Wu et al., 2019b). Hagos et al. (2018) further designed a multi-input CNN that learns symmetrical difference among patches to detect breast masses. Another popular way of utilizing segmentation labels is to train anchor-based object detection models. For instance, Ribli et al. (2018) and Févry et al. (2019) fine-tuned a Faster RCNN (Ren et al., 2015) to localize lesions on mammograms. Xiao et al. (2019) integrated object detector in a Siamese structure with explicit loss terms to differentiate anchor proposals containing lesion from those with only normal tissues. We refer the readers to Hamidinekoo et al. (2018); Gao et al. (2019); Geras et al. (2019) for comprehensive reviews of prior works on machine learning for mammography.

4.3. Weakly supervised object detection

Recent progress demonstrates that CNN classifiers, trained with image-level labels, are able to perform semantic segmentation at the pixel level (Oquab et al., 2015; Pinheiro and Collobert, 2015; Bilen and Vedaldi, 2016; Zhou et al., 2016; Diba et al., 2017; Zeng et al., 2019). This is commonly achieved in two steps. First, a backbone CNN converts the input image to a saliency map which highlights the discriminative regions. A global pooling operator then collapses the saliency map into scalar predictions, which makes the entire model trainable end-to-end. Durand et al. (2017) devised a new pooling operator that performs feature pooling on both spatial space and class space. Wei et al. (2018) augmented the backbone network using convolution filters with varying dilation rates to address scale variation among object classes. Zhu et al. (2019) refined segmentation masks using pseudo-supervision from noisy segment proposals.

Weakly supervised object detection (WSOD) has become increasingly popular in the field of medical image analysis as it eliminates the reliance of models on segmentation labels which are often expensive to obtain. WSOD has been broadly utilized in medical applications including disease classification (Yao et al., 2018; Liu et al., 2019), cell segmentation (Li et al., 2019; Yoo et al., 2019), and lesion detection (Xu et al., 2014; Luo et al., 2019; Wu et al., 2019a). Schlemper et al. (2019) designed a novel attention gate unit that can be integrated with standard CNN classifiers to localize objects of interest in ultrasound images. Ouyang et al. (2019) proposed a spatial smoothing regular-

ization to model the uncertainty associated with the segmentation mask. Kervadec et al. (2019) demonstrated that regularization terms stemming from inequality constraints can significantly improve the localization performance of a weakly supervised model. While many works still rely on weak localization labels such as point annotations (Yoo et al., 2019) and scribbles (Ji et al., 2019) to produce saliency maps, our approach requires only image-level labels that indicate the presence of an object of a given class. In addition, to make an image-level prediction, most existing models only utilize global information from the saliency maps which often neglect fine-grained details. In contrast, our model also leverages local information from ROI patches using a dedicated network. In Section 3.6, we empirically demonstrate that the ability to focus on fine visual detail is important for classification.

5. Discussion and conclusion

Medical images differ from typical natural images in many ways such as much higher resolutions and smaller ROIs. Moreover, both the global structure and local details play essential roles in the classification of medical images. Because of these differences, deep neural network architectures that work well for natural images might not be applicable to many medical image classification tasks. In this work, we present a novel framework, GMIC, to classify high-resolution screening mammograms. GMIC first applies a low-capacity, yet memory-efficient, global module on the whole image to extract the global context and generate saliency maps that provide coarse localization of possible benign/malignant findings. It then identifies the most informative regions in the image and utilizes a local module with higher capacity to extract fine-grained visual details from the chosen regions. Finally, it employs a fusion module that aggregates information from both global context and local details to produce the final prediction.

Our approach is well-suited for the unique properties of medical images. GMIC is capable of processing input images in a memory-efficient manner, thus being able to handle medical images in their original resolutions while still using a high-capacity neural network to pick up on fine visual details. Moreover, despite being trained with only image-level labels, GMIC is able to generate pixel-level saliency maps that provide additional interpretability.

We applied GMIC to interpret screening mammograms: predicting the presence or absence of malignant and benign lesions in a breast. Evaluated on a large mammography dataset, the proposed model outperforms the ResNet-34 while being **4.3x** faster and using **76.1**% fewer memory of GPU. Moreover, we also demonstrated that our model can generate predictions that are as accurate as radiologists, given equivalent input information. Given its generic design, the proposed model could be widely applicable to various high-resolution image classification tasks. In future research, we would like to extend this framework to other imaging modalities such as ultrasound, tomosynthesis, and MRI.

In addition, we note that training GMIC is slightly more complex than training a standard ResNet model. As shown in Fig. 13, the learning speeds for the global and local module are different. As learning of the global module stabilizes, the saliency maps tend to highlight a fixed set of regions in each example, which decreases the diversity of patches provided to the local module. This causes the local module to overfit, causing its validation AUC to decrease. We speculate that GMIC could benefit from a curriculum that optimally coordinates the learning of both modules. A learnable strategy such as the one proposed in Katharopoulos and Fleuret (2019) could help to jointly train both global and local module.

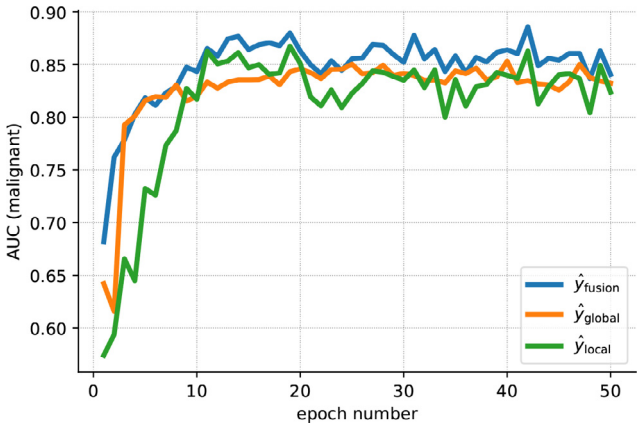

Fig. 13. Learning curves for a GMIC-ResNet-18 model. The AUC for malignancy prediction on the validation set is shown for $\hat{\mathbf{y}}_{\text{fusion}}$, $\hat{\mathbf{y}}_{\text{global}}$, and $\hat{\mathbf{y}}_{\text{local}}$.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yiqiu Shen: Conceptualization, Methodology, Software, Writing - original draft, Visualization. **Nan Wu:** Data curation, Visualization, Writing - review & editing. **Jason Phang:** Data curation, Writing - review & editing. **Jungkyu Park:** Data curation, Writing - review & editing. **Kangning Liu:** Investigation, Writing - review & editing. **Sudarshini Tyagi:** Investigation, Writing - review & editing. **Laura Heacock:** Resources, Writing - original draft, Writing - review & editing. **S. Gene Kim:** Resources, Writing - review & editing, Funding acquisition. **Linda Moy:** Resources, Writing - review & editing, Funding acquisition. **Kyunghyun Cho:** Resources, Conceptualization, Methodology, Writing - review & editing, Funding acquisition. **Krzysztof J. Geras:** Resources, Conceptualization, Methodology, Data curation, Writing - original draft, Writing - review & editing, Funding acquisition.

Acknowledgments

The authors would like to thank Joe Katsnelson, Mario Videna and Abdul Khaja for supporting our computing environment and Yizhuo Ma for providing graphical design consultation. We also gratefully acknowledge the support of Nvidia Corporation with the donation of some of the GPUs used in this research. This work was supported in part by grants from the National Institutes of Health (grants R21CA225175 and P41EB017183) and the National Science Foundation (grant 1922658).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101908.

References

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 (Feb), 281–305.

Bilen, H., Vedaldi, A., 2016. Weakly supervised deep detection networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2846–2854.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309.

Canziani, A., Paszke, A., Culurciello, E., 2016. An analysis of deep neural network models for practical applications. arXiv:1605.07678.

Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H., 2020. Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3133–3142.

Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R., 2015. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 118–126.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. 24 (10), 1559.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

DeSantis, C.E., Ma, J., Goding Sauer, A., Newman, L.A., Jemal, A., 2017. Breast cancer statistics, 2017, racial disparity in mortality by state. CA 67 (6), 439–448.

Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L., 2017. Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 914–922.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. Springer, pp. 1–15.

D'Orsi, C.J., 2013. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. American College of Radiology.

Duffy, S.W., Tabár, L., Chen, H.-H., Holmqvist, M., Yen, M.-F., Abdsalah, S., Epstein, B., Frodis, E., Ljungberg, E., Hedborg-Melander, C., et al., 2002. The impact of organized mammography service screening on breast carcinoma mortality in seven swedish counties: a collaborative evaluation. Cancer 95 (3), 458–469.

Durand, T., Mordan, T., Thome, N., Cord, M., 2017. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 642–651.

Feng, X., Yang, J., Laine, A.F., Angelini, E.D., 2017. Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. In: International Conference on Medical image Computing and Computer-Assisted Intervention. Springer, pp. 568–576.

Févry, T., Phang, J., Wu, N., Kim, S., Moy, L., Cho, K., Geras, K. J., 2019. Improving localization-based approaches for breast cancer screening exam classification. arXiv:1908.00615.

Gao, Y., Geras, K.J., Lewin, A.A., Moy, L., 2019. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. Am. J. Roentgenol. 212 (2), 300–307.

Geras, K.J., Mann, R.M., Moy, L., 2019. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. Radiology 293 (2), 246–259.

Geras, K. J., Shen, Y., Wolfson, S., Kim, S. G., Moy, L., Cho, K., 2017. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv:1703.07047v2.

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y., 2018. Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification. arXiv:1801.09927.

Guo, Y., Bi, L., Kumar, A., Gao, Y., Zhang, R., Feng, D., Wang, Q., Kim, J., 2019. Deep local-global refinement network for stent analysis in IVOCT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 539–546.

Hagos, Y.B., Mérida, A.G., Teuwen, J., 2018. Improving breast cancer detection using symmetry information with deep learning. In: Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 90–97.

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R., 2018. Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. 47, 45–67.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, pp. 630–645.

Huang, G., Liu, Z., Weinberger, K. Q., van der Maaten, L., 2016. Densely connected convolutional networks. arXiv:1608.06993.

Ilse, M., Tomczak, J. M., Welling, M., 2018. Attention-based deep multiple instance learning. arXiv:1802.04712.

Ji, Z., Shen, Y., Ma, C., Gao, M., 2019. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In: International Conference on Medical image Computing and Computer-Assisted Intervention. Springer, pp. 175–183.

Katharopoulos, A., Fleuret, F., 2019. Processing megapixel images with deep attention-sampling models. arXiv:1905.03711.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B., 2019. Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. 54, 88–99.

Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Kim, E.-K., Kim, H.-E., Han, K., Kang, B.J., Sohn, Y.-M., Woo, O.H., Lee, C.W., 2018. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Sci. Rep. 8 (1), 1–8.

Kooi, T., Karssemeijer, N., 2017. Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks. J. Med. Imaging 4 (4), 044501.

Kopans, D.B., 2002. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. Cancer 94 (2).

Kopans, D.B., 2015. An open letter to panels that are deciding guidelines for breast cancer screening. Breast Cancer Res. Treat. 151 (1), 19–25.

Kyono, T., Gilbert, F. J., van der Schaar, M., 2018. MAMMO: a deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. arXiv:1811.02661.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data 4, 170177.

Lehman, C.D., Arao, R.F., Sprague, B.L., Lee, J.M., Buist, D.S., Kerlikowske, K., Henderson, L.M., Onega, T., Tosteson, A.N., Rauscher, G.H., et al., 2016. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology 283 (1), 49–58.

Lehman, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N., Miglioretti, D.L., 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Med. 175 (11), 1828–1837.

Li, C., Wang, X., Liu, W., Latecki, L.J., Wang, B., Huang, J., 2019. Weakly supervised mitosis detection in breast histopathology images using concentric loss. Med. Image Anal. 53, 165–178.

Li, H., Chen, D., Nailon, W.H., Davies, M.E., Laurenson, D., 2020. Dual convolutional neural networks for breastmass segmentation and diagnosis inmammography. arXiv:2008.02957.

Li, L., Zheng, Y., Zhang, L., Clark, R.A., 2001. False-positive reduction in cad mass detection using a competitive classification strategy. Med. Phys. 28 (2), 250–258.

Liberman, L., Menell, J.H., 2002. Breast imaging reporting and data system (bi-rads). Radiol. Clin. 40 (3), 409–430.

Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., Yu, Y., 2019. Align, attend and locate: chest X-ray diagnosis via contrast induced attention network with limited supervision. In: International Conference on Computer Vision, pp. 10632–10641.

Lotter, W., Sorensen, G., Cox, D., 2017. A multi-scale CNN and curriculum learning strategy for mammogram classification. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 169–177.

Luo, L., Chen, H., Wang, X., Dou, Q., Lin, H., Zhou, J., Li, G., Heng, P.-A., 2019. Deep angular embedding and feature correlation attention for breast MRI cancer analysis. arXiv:1906.02999.

Luong, M.-T., Pham, H., Manning, C. D., 2015. Effective approaches to attention-based neural machine translation. arXiv:1508.04025.

Masotti, M., Lanconelli, N., Campanini, R., 2009. Computer-aided mass detection in mammography: False positive reduction via gray-scale invariant ranklet texture features. Med. Phys. 36 (2), 311–316.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A., et al., 2020. International evaluation of an ai system for breast cancer screening. Nature 577 (7788), 89–94.

Mongan, J., Moy, L., 2020. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. Radiology: Artificial Intelligence 2 (2), e200029. doi:10.1148/ryai.2020200029.

Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J., Denton, E.R., Zwiggelaar, R., 2010. A review of automatic mass detection and segmentation in mammographic images. Med. Image Anal. 14 (2), 87–110.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 685–694.

Ouyang, X., Xue, Z., Zhan, Y., Zhou, X.S., Wang, Q., Zhou, Y., Wang, Q., Cheng, J.-Z., 2019. Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest x-ray. In: International Conference on Medical image Computing and Computer-Assisted Intervention. Springer, pp. 613–621.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.

Pereira, S.M.P., McCormack, V.A., Moss, S.M., dos Santos Silva, I., 2009. The spatial distribution of radiodense breast tissue: a longitudinal study. Breast Cancer Res. 11 (3), R33.

Pinheiro, P.O., Collobert, R., 2015. From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1713–1721.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., 2017. Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225.

Rampun, A., López-Linares, K., Morrow, P.J., Scotney, B.W., Wang, H., Ocaña, I.G., Maclair, G., Zwiggelaar, R., Ballester, M.A.G., Macía, I., 2019. Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network. Med. Image Anal. 57, 1–17.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. 8 (1), 1–7.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Roth, H. R., Lee, C. T., Shin, H.-C., Seff, A., Kim, L., Yao, J., Lu, L., Summers, R. M., 2015. Anatomy-specific classification of medical images using deep convolutional nets. arXiv:1504.04003.
Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.
Sedai, S., Mahapatra, D., Ge, Z., Chakravorty, R., Garnavi, R., 2018. Deep multi-scale convolutional feature learning for weakly supervised localization of chest pathologies in x-ray images. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 267–275.
Shen, L., 2017. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv:1711.05775.
Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh,, W., 2019. Deep learning to improve breast cancer detection on screening mammography. Sci. Rep. 9, 1-12.
Shen, Y., Wu, N., Phang, J., Park, J., Kim, G., Moy, L., Cho, K., Geras, K.J., 2019. Globally-aware multiple instance classifier for breast cancer screening. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with International Conference on Medical image Computing and Computer-Assisted Intervention 2019, Shenzhen, China, October 13, 2019, Proceedings, 11861. Springer, p. 18.
Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z., 2020. Deep neural networks with region-based pooling structures for mammographic image classification. IEEE Trans. Med. Imaging 39 (6), 2246–2255.
Siegel, R.L., Miller, K.D., Jemal, A., 2020. Cancer statistics, 2020. CA 70 (1), 7–30. doi:10.3322/caac.21590.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
Sun, M., Zhang, G., Dang, H., Qi, X., Zhou, X., Chang, Q., 2019. Accurate gastric cancer segmentation in digital pathology images using deformable convolution and multi-scale embedding networks. IEEE Access 7, 75530–75541.
Sun, M., Zhou, W., Qi, X., Zhang, G., Girnita, L., Seregard, S., Grossniklaus, H.E., Yao, Z., Zhou, X., Stålhammar, G., 2019. Prediction of BAP1 expression in uveal melanoma using densely-connected deep classification networks. Cancers 11 (10), 1579.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.
Tan, M., Le, Q. V., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946.
Teare, P., Fishman, M., Benzaquen, O., Toledano, E., Elnekave, E., 2017. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J. Digit. Imaging 30 (4), 499–505.
Van Gils, C.H., Otten, J.D., Verbeek, A.L., Hendriks, J.H., 1998. Mammographic breast density and risk of breast cancer: masking bias or causality? Eur. J. Epidemiol. 14 (4), 315–320.
Wei, J., Chan, H.-P., Wu, Y.-T., Zhou, C., Helvie, M.A., Tsodikov, A., Hadjiiski, L.M., Sahiner, B., 2011. Association of computerized mammographic parenchymal pattern measure with breast cancer risk: a pilot case-control study. Radiology 260 (1), 42–49.
Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S., 2018. Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7268–7277.
Wu, K., Du, B., Luo, M., Wen, H., Shen, Y., Feng, J., 2019. Weakly supervised brain lesion segmentation via attentional representation learning. In: International Conference on Medical image Computing and Computer-Assisted Intervention. Springer, pp. 211–219.
Wu, N., Geras, K.J., Shen, Y., Su, J., Kim, S.G., Kim, E., Wolfson, S., Moy, L., Cho, K., 2018. Breast density classification with deep convolutional neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 6682–6686.
Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Févry, T., Katsnelson, J., Kim, E., et al., 2019b. Deep neural networks improve radiologists' performance in breast cancer screening. arXiv:1903.08297.
Wu, N., Phang, J., Park, J., Shen, Y., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K.J., 2019. The NYU Breast Cancer Screening Dataset v1.0. Technical Report. Available at https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf
Wu, Y.-T., Wei, J., Hadjiiski, L.M., Sahiner, B., Zhou, C., Ge, J., Shi, J., Zhang, Y., Chan, H.-P., 2007. Bilateral analysis based false positive reduction for computer-aided mass detection. Med. Phys. 34 (8), 3334–3344.
Xiao, L., Zhu, C., Liu, J., Luo, C., Liu, P., Zhao, Y., 2019. Learning from suspected target: Bootstrapping performance for breast cancer detection in mammography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 468–476.
Xu, Y., Zhu, J.-Y., Eric, I., Chang, C., Lai, M., Tu, Z., 2014. Weakly supervised histopathology cancer image segmentation and classification. Med. Image Anal. 18 (3), 591–604.
Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K., 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv:1803.07703.
Yoo, I., Yoo, D., Paeng, K., 2019. Pseudoedgenet: nuclei segmentation only with point annotations. arXiv:1906.02924.
Zeng, Y., Zhuge, Y., Lu, H., Zhang, L., 2019. Joint learning of saliency detection and weakly supervised semantic segmentation. In: International Conference on Computer Vision, pp. 7223–7233.
Zhong, Z., Li, J., Zhang, Z., Jiao, Z., Gao, X., 2019. An attention-guided deep regression model for landmark detection in cephalograms. arXiv:1906.07549.
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929.
Zhu, W., Lou, Q., Vang, Y.S., Xie, X., 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: International Conference on Medical image Computing and Computer-Assisted Intervention. Springer, pp. 603–611.
Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D., Jiao, J., 2019. Learning instance activation maps for weakly supervised instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3116–3125.