Clinically Explainable Disease Diagnosis based on Biomarker Activation Map

Pengxiao Zang, Carol Wang, Tristan T. Hormel, Steven T. Bailey, Thomas S. Hwang, and Yali Jia*

Abstract*— Objective: Artificial intelligence (AI)-based disease classifiers have achieved specialist-level performances in several diagnostic tasks. However, real-world adoption of these classifiers remains challenging due to the black box issue. Here, we report a novel biomarker activation map (BAM) generation framework that can provide clinically meaningful explainability to current AI-based disease classifiers. Methods: We designed the framework based on the concept of residual counterfactual explanation by generating counterfactual outputs that could reverse the decision-making of the disease classifier. The BAM was generated as the difference map between the counterfactual output and original input with postprocessing. We evaluated the BAM on four different disease classifiers, including an age-related macular degeneration classier based on fundus photography, a diabetic retinopathy classifier based on optical coherence tomography angiography, a brain tumor classifier based on magnetic resonance imaging (MRI), and a breast cancer classifier based on computerized tomography (CT) scans. Results: The highlighted regions in the BAM correlated highly with manually demarcated biomarkers of each disease. Conclusion/Significance: The BAM can improve the clinical applicability of an AI-based disease classifier by providing intuitive output clinicians can use to understand and verify the diagnostic decision.

***Index Terms*—Deep learning, Medical imaging, Explainable AI**

I. Introduction

END-to-end artificial intelligence (AI)-based classifiers have achieved state-of-the-art performances in a broad range of diagnostic tasks in medical imaging [1-5]. These disease classifiers can lower clinical burden and increase diagnostic efficiency [5]. However, the real-world adoption of such classifiers faces barriers due to the black-box nature of AI (Fig. 1a). For the output of a classifier to be acceptable to clinicians and patients, it must be explainable. In addition, when a classifier gives differential output for patients of different ethnic or racial backgrounds, it is impossible to determine whether the output is due to real differences in these groups or an unacceptable bias and disparate performance [6-8]. To address this, regulations in the United States and Europe [6] require that AI outputs must be explainable [6-8]. For example, clinicians should be able to understand salient features of AI decision-making through heuristic tools. The dearth of tools that explain AI decision-making is a major hurdle for the adoption of AI-based classifiers in the clinical setting [6-8].

Contemporary AI-based classifier explainability methods can be summarized into two categories [9-11]. The first is numerical methods, which generate activation maps directly from the parameters and gradients of the classifiers (Fig. 1b) [12-14]. These methods are usually easy to develop and deploy since they do not need extra parameters and training. Due to their ease of use, such methods, like class activation maps (CAM), have been widely used in AI-based disease diagnosis research projects [13]. However, the activation maps generated by these methods are not always appropriate for clinical practice since they focus more on algorithm explanation rather than user explanation [15], failing to connect algorithmic output to clinical decision-making. They often do not provide meaningful explainability for the clinician or the regulators.

The second category is activation maps generated based on direct counterfactual explanations [15-23]. These methods usually involve the development of a deep learning generator, which is trained to directly generate a counterfactual output that can switch the decision-making or lower the classifier's confidence in the original input. The activation map is generated as the difference maps between the output and input of the generator (Fig. 1c). The counterfactual explanation concept fits well for disease diagnostic tasks since such tasks usually can be transferred to binary classification and the clinician's interest is whether each input reveals unique biomarkers belonging to a specific disease or is a sign of greater disease severity. Several methods have been proposed based on counterfactual explanations for deep-learning disease classifiers [15-23]. The biggest challenge for the counterfactual explanation methods is restricting the difference between counterfactual output and original input to just the classifier-utilized biomarkers. The reconstruction error in current direct generation methods is one of the major causes of erroneously highlighting areas not utilized by the classifier.

This work was supported by grants from National Institutes of Health (R01 EY 036429, R01 EY035410, R01 EY024544, R01 EY027833, R01 EY031394, R43EY036781, P30 EY010572, T32 EY023211, UL1TR002369); the Jennie P. Weeks Endowed Fund; the Malcolm M. Marquis, MD Endowed Fund for Innovation; Unrestricted Departmental Funding Grant and Dr. H. James and Carole Free Catalyst Award from Research to Prevent Blindness (New York, NY); Edward N. & Della L. Thome Memorial Foundation Award, and the Bright Focus Foundation (G2020168, M20230081).

*Y. Jia is with Casey Eye Institute, Oregon Health & Science University, Portland, OR 97239 USA, and also with Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR 97239 USA (correspondence e-mail: jiaya@ohsu.edu).

P. Zang, T. T. Hormel, S. T. Bailey, and T. S. Hwang are with Casey Eye Institute, Oregon Health & Science University, Portland, OR 97239 USA.

C. Wang is with Jesuit High School, Portland, OR 97225, USA.

Published methods of direct counterfactual explanations were designed to explain all the differences instead of developing methods that restrict differences to the areas with utilized biomarkers [16-23]. An additional limitation is these methods were developed exclusively on a single imaging modality for a single disease. It is unknown if these methods can be applied to other imaging modalities and diseases.

Here, we propose a novel biomarker activation map (BAM) generation framework based on the concept of residual counterfactual explanation. We designed the framework and study focused on generating activation maps that only highlight the classifier-utilized biomarkers (Fig. 1d). Expanding on our previous study [15], we optimized the framework and evaluated the BAM on four different disease classifiers, including an age-related macular degeneration (AMD) classifier based on color fundus photography (CFP), a diabetic retinopathy (DR) classifier based on optical coherence tomography angiography (OCTA), a brain tumor classifier based on magnetic resonance imaging (MRI), and a breast cancer classifier based on computerized tomography (CT).

II. Related Work

Several methods have been proposed and try to solve the black-box issue of AI classifiers. Among these methods, there are three methods are commonly used in different studies. First, local interpretable model agnostic explanations (LIME) methods can generate local explanations by approximating the black-box model's behavior around specific instances through interpretable surrogate models, such as sparse linear regressions. Second, SHapley Additive exPlanations (SHAP) leverages cooperative game theory, specifically Shapley values, to quantify the contribution of each feature towards a particular prediction, offering both local and global insights. At last, in contrast to these feature attribution methods, CAM (Class Activation Mapping) provides visual explanations specifically for convolutional neural networks (CNNs) by highlighting regions in input images that significantly influence the network's predictions, thus aiding intuitive interpretation in image classification tasks. Despite their broad adoption and utility, these approaches generally explain predictions by attributing importance scores to existing features or image regions, but do not provide insights into actionable interventions required to alter model outcomes, which is the central motivation behind counterfactual explanation methods.

The counterfactual explanation method has been used to explain different AI-based disease classifiers [15-23]. A. Katzmann et al. proposed a method based on cycle-consistent activation maximization to highlight the biomarker changes correlated to the decision-making of a classifier [19]. A. Dravid et al. proposed generating synthesized reconstructed and negative images based on latent space manipulation [18]. The difference map between the two images was used to highlight the classifier-utilized biomarkers. S. Singla et al. proposed a method to generate counterfactual outputs by combining a conditional generative adversarial network (GAN) with a pre-trained semantic segmentation network [20]. A. J. DeGrave et al. proposed a method to generate outputs with different diagnostic results based on a previously proposed study called "Explanation by Progressive

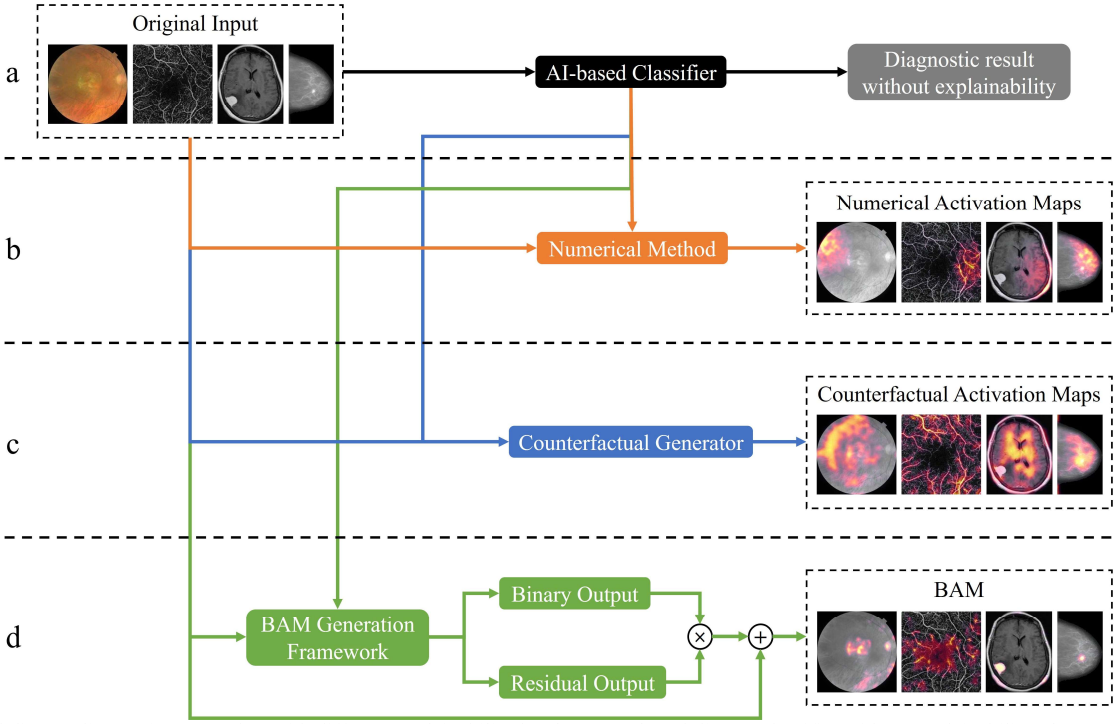

Fig. 1. AI-based disease diagnostic approaches with different explainability levels. (a) AI-based disease classifier without explainability. (b) Numerical methods provide explainability that may be clinically meaningless. In this example, the highlighted regions are mostly healthy tissues, which may be misleading in a medical imaging context. (c) Current counterfactual explanation methods provide insufficient clinical explainability. In this example, the highlighted regions include classified utilized biomarkers and healthy tissues, which means clinicians still need to identify actual biomarkers. (d) Our BAM generation framework provides clinically meaningful explainability. In this example, only the classifier-utilized biomarkers are highlighted.

Exaggeration" [17]. This method was developed on a large skin cancer data set. Physicians analyzed the difference between the original input and generated outputs. However, all these methods were evaluated on classifiers developed for a single disease on a single imaging modality. In the present work, we attempted to develop a general framework that could be used across multiple imaging modalities. A general framework can provide explainability to classifiers in various diagnostic tasks without the need for redevelopment, which is time-consuming and clinically impractical.

III. Materials

Four disease classifiers based on different imaging modalities were used in this study to evaluate the generalization performance of our BAM. The first AMD classifier was derived from a data set of 826 fundus photography obtained from two public data sets [24, 25]. This data set includes 326 AMD subjects and 500 healthy controls. After cropping the blank edges, all the images in this data set were resized to 224 × 224 × 3. The second DR classifier was derived from a data set of 456 en face OCTA images obtained from Casey Eye Institute, Oregon Health & Science University [15, 26]. This data set includes 257 referable and 199 non-referable DR subjects. All the images in this data set were resized to 224 × 224. The third brain tumor classifier was derived from a data set of 3000 MRI images obtained from a public data set [27]. This data set includes 1500 brain tumor subjects and 1500 healthy controls. After cropping the blank edges, all the images in this data set were resized to 192 × 160. The last breast cancer classifier was derived from a data set of 2400 CT images obtained from a public data set [28]. This data set includes 1200 breast cancer subjects and 1200 healthy controls. After cropping the blank edges, all the images in this data set were resized to 224 × 128. For unbiased development and evaluation, each data set was split as training (60%), validation (20%), and testing (20%), respectively.

IV. Methods

To provide a clinically meaningful explanation to a trained disease classifier, our BAM generation framework was designed to learn which unique biomarkers that only belong to the positive class were potentially utilized by the classifier. Based on the counterfactual explanation, the classifier-utilized biomarkers were defined as the necessary changes that would change the decision-making of the classifier from positive to negative. A generator was trained to generate counterfactual outputs based on the input images and reversed class labels to detect these necessary changes. The outputs should switch the decision-making of the classifier by only adding necessary changes to the input images. A cycle consistency learning strategy was used during training to reduce the unnecessary changes made by the generator [29]. The predicted positive images and negative labels would be used as inputs to generate counterfactual negative outputs during the inference. The BAM was then generated as a different map between the output and the generator's input.

A. Generator Architecture

The main and the assistant generators were established based on the same architecture. This architecture was designed by combining self-attention and UNet [30-32] (Fig. 2). The self-attention was performed using transformer blocks at the end of each residual block in both the encoder and decoder. A reconstructed block was designed and used after the last residual block of the decoder. The changes map and a binary mask of the necessary changes were generated. A Tanh activation was applied on the changes map to fulfill both plus and minus changes within the value range of the input. A SoftMax activation was applied on the binary mask to classify each position as unnecessary or necessary changes. The residual counterfactual output was calculated as the input plus the multiplication between the changes map and the binary mask.

B. Framework Training

The main G_0 and assistant G_1 generators were trained to generate outputs that can be diagnosed as negative and positive, respectively. Here, residual counterfactual outputs should have the opposite diagnosis of the inputs, while preserved outputs should have no changes compared to the inputs. The main generator was trained to generate counterfactuals $G_0(x_+)$ and preserve $G_0(x_-)$ negatives from positive and negative inputs. The assistant generator was trained to generate counterfactual $G_1(x_-)$ and $G_1(x_+)$ preserved positive from the negative and positive inputs. That is, the main generator always produces images classified as negative diagnoses, and the assistant generator always produces images classified as positive diagnoses. The counterfactual negative was also used as the input of the assistant generator to generate a cycled positive output $G_1(G_0(x_+))$, which should be the same as the positive input x_+ used to generate the counterfactual negative. And vice versa for the counterfactual positive. To improve the similarity between residual counterfactual output and the actual image in the corresponding class, two discriminators were D_- and D_+ were trained for negative and positive images, respectively. For the counterfactual output of the main generator, the negative cross entropy was utilized as the loss function to make sure this output could switch the decision making y_+ of the classifier. The adversarial loss was used to improve the similarity between the counterfactual negative and actual negative images.

$$L_{COU}(G_0(x_+), y_+) = -\begin{bmatrix}(1-y_+)\cdot\log\left(F\left(G_0(x_+)\right)\right)+\\ y_+\cdot\left(1-\log\left(F\left(G_0(x_+)\right)\right)\right)\end{bmatrix} \quad (1)$$

$$L_{GAN}(G_0(x_+), x_-) = E_{X_-}[\log D_-(x_-)] + E_{X_+}\left[\log\left(1 - D_-\left(G_0(x_+)\right)\right)\right] \quad (2)$$

For the preserved output of the main generation, both L1 and L2 norm were used to keep it same as the input.

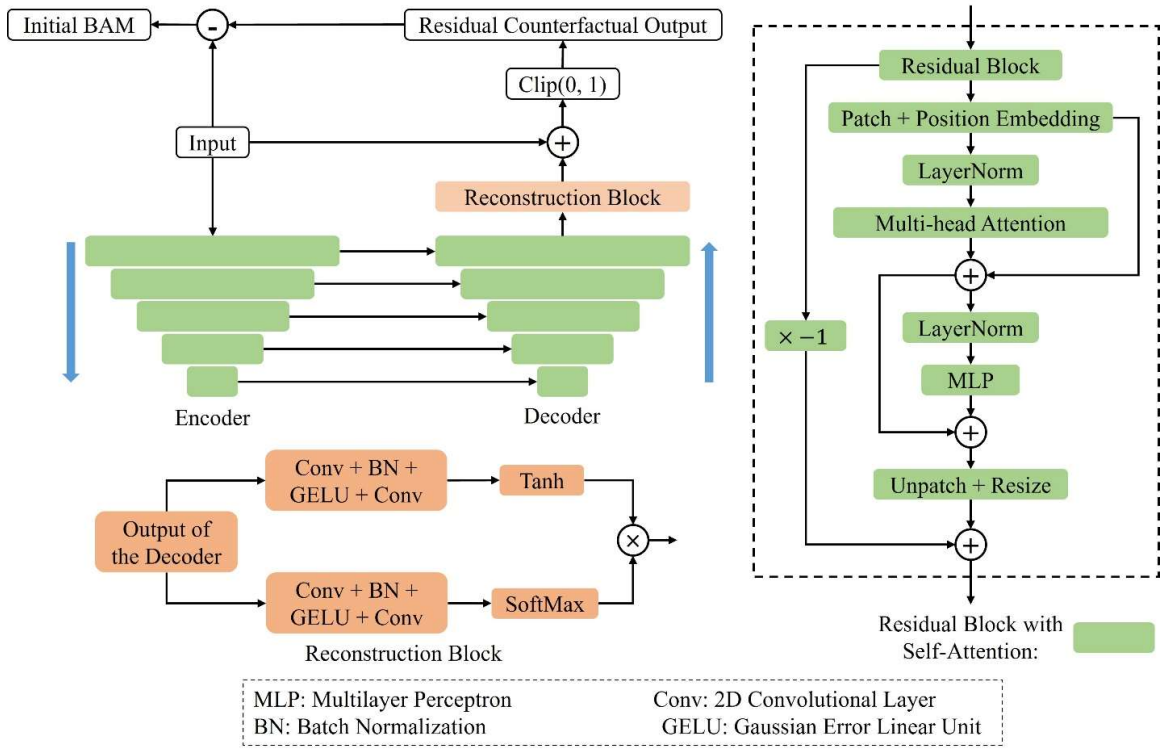

Fig. 2. Detailed architecture of the generator. The green patches represent the innovated self-attention residual block with stride size 2. A deconvolutional block is utilized between every two adjacent green patches in the decoder. The blue arrows represent the forward direction in encoder and decoder. The black arrows represent skip connections between encoder and decoder. The innovative reconstruction block is represented as orange patch.

$$L_{PRE}(G_0(x_-), x_-) = \|G_0(x_-) - x_-\|_1 + \|G_0(x_-) - x_-\|_2 \quad (3)$$

Similar equations above were used for the assistant generator by replacing G_0 to G_1, D_- to D_+, x_+ to x_-, and y_+ to y_-. In addition, the cycled loss was used for both main and assistant generator.

$$L_{CYC} = \|G_1(G_0(x_+)) - x_+\|_1 + \|G_1(G_0(x_+)) - x_+\|_2 + \|G_0(G_1(x_-)) - x_-\|_1 + \|G_0(G_1(x_-)) - x_-\|_2 \quad (4)$$

Because our BAM was designed to focus on the unique biomarkers that only belong to the positive class. The binary mask $M_{G_0(x_-)}$ generated with preserved negative $G_0(x_-)$ was trained to be blank image to avoid the changes of biomarkers not utilized by the classifier. The F1-score between this mask and a white image M_1 was minimized during the training.

$$L_M = \frac{2 \cdot \mathrm{sum}\left(M_{G_0(x_-)} \cdot M_1\right)}{\mathrm{sum}\left(M_{G_0(x_-)}\right) + \mathrm{su}\ \ (M_1)} \quad (5)$$

By combining all the above loss functions, the whole object during training was:

$$\begin{aligned} Loss = {} & L_{COU}(G_0(x_+), y_+) + L_{GAN}(G_0(x_+), x_-) \\ & + L_{PRE}(G_0(x_-), x_-) + L_M \\ & + L_{COU}(G_1(x_-), y_-) + L_{GAN}(G_1(x_-), x_+) + \\ & L_{PRE}(G_1(x_+), x_+) + L_{CYC} \end{aligned} \quad (6)$$

By training our framework to fulfill all the requirements above, the main generator could learn which biomarkers only belonged to the positive class and were used by the classifier to diagnose. All the positive and negative labels just described were based on the classifier's diagnostic results, not the ground truth labels.

The framework was trained and validated on the classifier's data set. The main generator with the lowest validation loss was finally selected for the evaluation described in the results section. The initial BAM was generated as the absolute difference between the output and input of the selected main generator. A Gaussian filter with kernel size 3 was used to generate the final BAM.

C. *Implementation Details*

To evaluate our residual BAM generation framework, four different classifiers were trained on CFP for AMD diagnosis, OCTA for rDR diagnosis, MRI for brain tumor diagnosis and CT for breast cancer diagnosis. The four classifiers were constructed with the same architecture based on VGG19 [33] with batch normalization and only one fully connected layer. Each classifier was trained and validated on the training and validation data set, respectively. For each disease, the trained classifier with highest validation accuracy was selected to evaluate our BAM generation framework.

Four separate models were independently developed for each of the classifiers described above. In the framework's training of each model, two stochastic gradient descent optimizers with Nesterov momentum (momentum = 0.9) were used simultaneously for the main and assistant generators. Hyperparameters during training included a batch size of 3 and 500 training epochs, and a learning rate of 0.0001 was used for all the training steps. To reduce visual distortions in the shown BAMs, perceptually uniform color maps were used on each BAM [34].

This study was implemented in PyTorch version 1.12 on

TABLE I
DATA DISTRIBUTION AND DIAGNOSTIC PERFORMANCE

Modality	CFP	OCTA	MRI	CT
Disease	AMD	Referable DR	Brain Tumor	Breast Cancer
Negative	2802	199	1500	1200
Positive	448	257	1500	1200
AROC	0.91	0.97	0.99	0.82

Ubuntu 20.04 server. The server has an Intel(R) Xeon(R) Gold 6254 CPU @ 3.10GHz ×2, 512.0 GB RAM and four NVIDIA RTX 3090 GPUs. But only two GPUs were used in the development of the framework. The inference time is about 2.8 seconds for each input.

V. Results

A. Generalizable Explainability

To test the generalizability of our framework, we evaluated the BAM on four disease classifiers (Table I). The first classifier was trained for AMD diagnosis based on CFP data extracted from two public datasets [24, 25]. The second classifier was trained for referable DR diagnosis using an OCTA dataset from Oregon Health & Science University [15, 26]. The third classifier was trained for brain tumor diagnosis based on a public MRI dataset [27]. The last classifier was trained for breast cancer diagnosis based on a public CT dataset [28]. The trained classifiers were evaluated on the testing set and achieved 0.91, 0.97, 0.99, and 0.82 area under the receiver operating characteristic curve (AROC) for AMD, referable DR, brain tumor, and breast cancer diagnosis, respectively.

B. Residual Mechanism for Attention Control

The activation maps generated by current counterfactual explanation methods have insufficient clinical explainability because they highlight healthy tissues or biomarkers which are not utilized by the classifier (Fig. 1c). This limitation is due in part to the direct generation strategy which reconstructs errors on the entire counterfactual output rather than just the relevant biomarkers. These errors can arise even when the exact original input is reconstructed. In the context of counterfactual generation, these reconstruction errors become more pronounced, as modifications are necessary to alter the classifier's decision-making, further complicating the accurate highlighting of utilized areas. As a result, the reconstruction errors in the areas not utilized by the classifier will lower the explainability of the outputs. The other challenge is that, to restrict the highlighting, the generator needs to learn how to differentiate the classifier-utilized biomarkers from other tissues. Compared to current methods, our BAM differs in that the counterfactual output is generated by adding a residual output to the original input (Fig. 1d). To restrict the residual output to the classifier-utilized biomarkers, a binary output is generated to filter the residual output before adding to the original input (Fig. 1d). By indirectly generating the counterfactual output, our BAM can significantly lower the influence of reconstruction errors.

Just as clinicians focus more on pathologies than on healthy tissues, our BAM was trained with a similar approach and focused on the classifier-utilized biomarkers on the positive class. Under this strategy, if the BAM performs well, the highlighted areas in a negative input should be significantly smaller than those in a positive input. However, small areas of highlighting can still be observed in part of the negative inputs for two main reasons. First, the classifier is not flawless and may mistakenly identify some irrelevant tissues as pathologic. However, these tissues are typically too small and insignificant to cause a negative input to be classified as positive. Second, in classification tasks distinguishing between lower and higher severities of a disease, some pathologies may still be present in lower severity inputs (negative). As mentioned earlier, these pathologies are not substantial enough to alter the classification of the negative input to positive. To quantitatively evaluate the performance of our BAM to differentiate between classifier-utilized biomarkers and surrounding tissues, we compared the highlighted area percentages between BAMs generated for positive and negative predicted inputs (Table II, Fig. 3). For the BAM of each input, the highlighted areas were identified as values exceeding the sum of 1.2 times the mean value and 0.8 times the standard deviation across the entire map.

TABLE II
COMPARISON OF THE PERCENTAGE OF HIGHLIGHTED AREAS OF BAMS BETWEEN POSITIVE AND NEGATIVE PREDICTIONS

	CFP	OCTA	MRI	CT
Negative, %	1.58 ± 1.32	9.34 ± 0.89	3.96 ± 1.20	0.41 ± 0.42
Positive, %	2.26 ±1.02	10.08 ± 1.17	5.15 ± 1.73	1.14 ± 0.96
P-value	< 0.001			

C. BAM for AMD Classifier on CFP

AMD is a common eye condition that primarily affects older adults, leading to the deterioration of the macula, the central part of the retina responsible for sharp vision [35]. In CFP

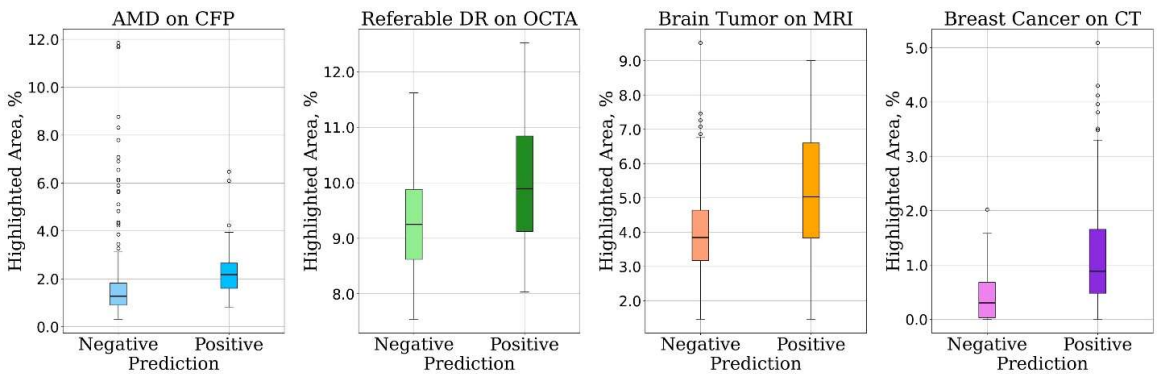

Fig. 3. Differentiation between classifier-utilized biomarkers and surrounding tissues. The highlighted areas of the BAMs of positively predicted inputs were significantly smaller than areas of negatively predicted inputs across all four classifiers.

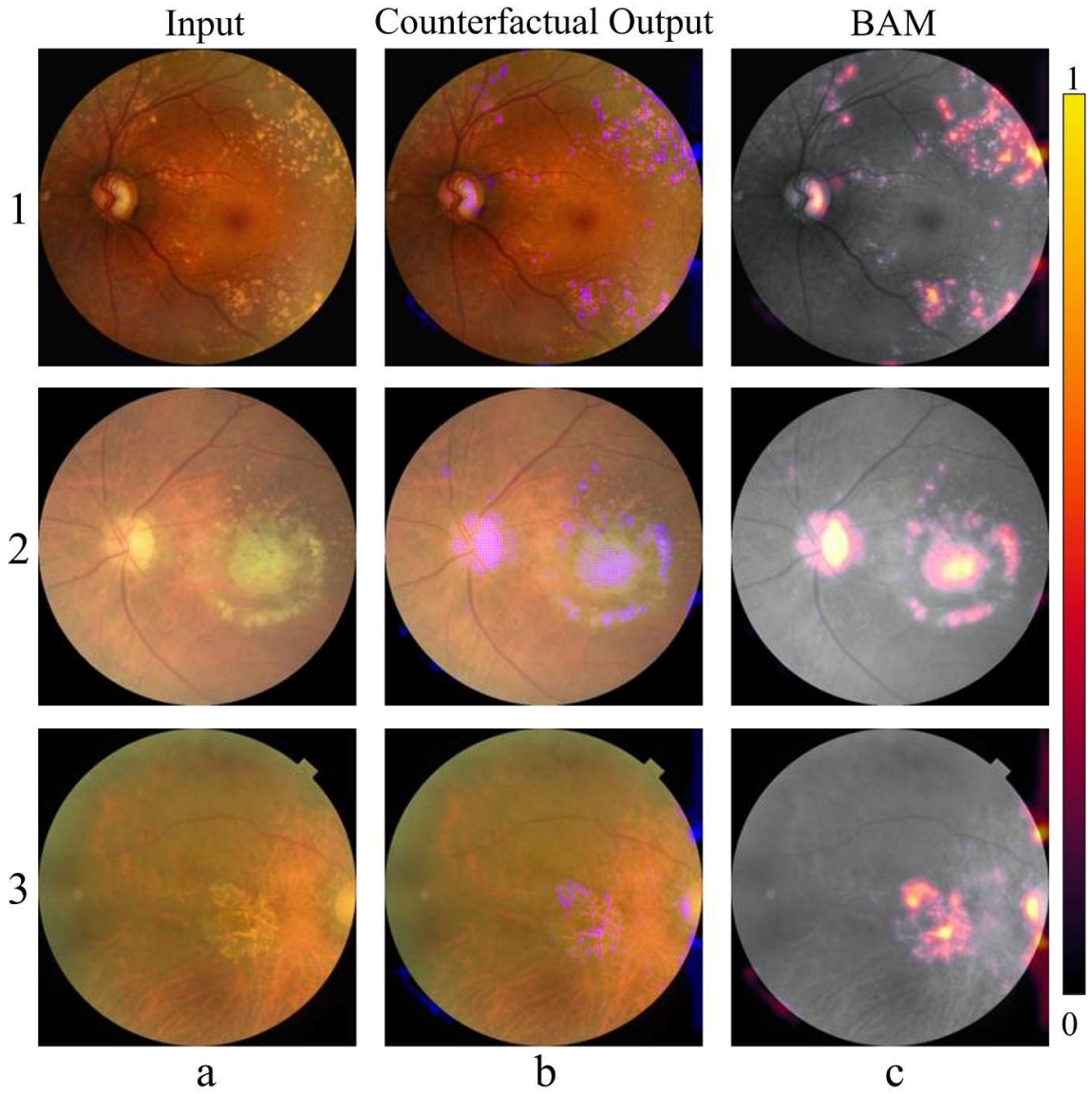

Fig. 4. BAMs for the AMD classifier on CFP. (a) Three representative AMD inputs. (b) The residual counterfactual outputs of the BAM generator, which are modified to produce images that would be diagnosed as non-AMD by the same classifier used on the inputs. To make the changes in the counterfactual output more easily identifiable, the change values have been amplified by a factor of five. (c) The BAM highlights the classifier-utilized biomarkers, drusen (1), macular neovascularization (2), and geographic atrophy (3). The optic cup is highlighted in all three inputs.

images, early-stage AMD is primarily characterized by the presence of drusen—visible yellow deposits of lipid and protein on CFP that have a variety of morphology (hard vs soft), sizes, and location. Our BAM successfully highlighted the hard drusen with different sizes and locations on a representative early AMD case (Fig. 4 Row 1). This indicates the classifier successfully learned that drusen is a major biomarker for AMD diagnosis.

As AMD progresses, macular neovascularization may develop, signaling an advanced stage of the disease. In CFP, neovascularization is grayish-greenish areas of variable pigmentation and association with heme and lipid that have a largely distinct appearance relative to drusen. The BAM also accurately highlighted the lesion (Fig. 4 Row 2), which has indistinct borders—note the higher proportion of lower confidence (red) pixels. In this case, the drusen were also highlighted, indicating that the classifier still recognized them as significant biomarkers in the advanced stage of AMD. This observation is consistent with the fact that the classifier was trained solely for binary classification between non-AMD and AMD.

Another advanced form of AMD is geographic atrophy, which is visible in CFP as well (Fig. 4 Row 3). Geographic atrophy is marked by sharply demarcated areas of retinal thinning and depigmentation, making the underlying choroidal vasculature visible. Our BAM accurately identified the area of geographic atrophy. In addition to the relevant pathologies, the temporal sector of the optic disc is also highlighted in all three CFPs. By comparing all the highlighted areas between original inputs and residual counterfactual outputs, we found the BAM generator is generally transferring the yellow areas with a specific shape in the original inputs by adding pixel values mainly to the blue channel. This observation suggests that the classifier may produce false positives when analyzing CFPs with irregularly shaped yellow areas not caused by AMD.

D. BAM for Referable DR Classifier on OCTA

DR is a microvasculopathy of the eye and is a leading cause of preventable blindness globally [36]. In clinical practice, distinguishing between referable (rDR) and non-referable DR (nrDR) is essential since rDR indicates a more advanced stage with higher risk of vision loss that requires specialist referral. Traditionally, rDR is mostly defined using CFP or clinical examination [36]. Recently, OCTA and its biomarkers have been extensively studied for objective diagnosis of different levels of DR. [37] It provides a non-invasive method to visualize blood flow in the retinal and choroidal vasculature, offering detailed insights into the microvascular changes associated with rDR. One common OCTA feature of rDR is nonperfusion areas, where the retinal capillaries are absent or significantly reduced. The nonperfusion areas were prominently highlighted in our BAM (Fig. 5), indicating that the classifier relied on nonperfusion as a primary biomarker for rDR classification. These highlighted regions closely matched the manually segmented nonperfused areas (Fig. 5c), demonstrating that the classifier effectively learned this pathology. In addition to nonperfusion areas, we also observed highlighting on vessels with abnormal shapes, which corresponded to DR-related vasculopathies such as microaneurysms and dilated vessels, both of which can be readily identified on OCTA. In addition, similar highlighting was also shown in our previous study [15].

E. BAM for Brain Tumor Classifier on MRI

In MRI images, brain tumors are primarily characterized by the presence of abnormal lesions. These tumors can have various phenotypes, including gliomas, meningiomas, and pituitary adenomas, each with distinct characteristics. Brain tumors typically manifest as regions of altered signal intensity on MRI, with their appearance depending on the imaging sequence used. The size and shape of these tumors can vary widely, but the key feature is their distinct signal alteration compared to the surrounding healthy brain tissue (Fig. 6a) [38]. The BAMs generated for the brain tumor classifier based on MRI accurately and sharply highlighted the real tumor areas (Fig. 6). In residual counterfactual outputs, the pixels' intensity in the brain tumor areas of both cases was decreased, which means the classifier has learned that the pixel intensity of surrounding healthy tissues should be lower than the tumor area in these two cases.

F. BAM for Breast Cancer Classifier on CT

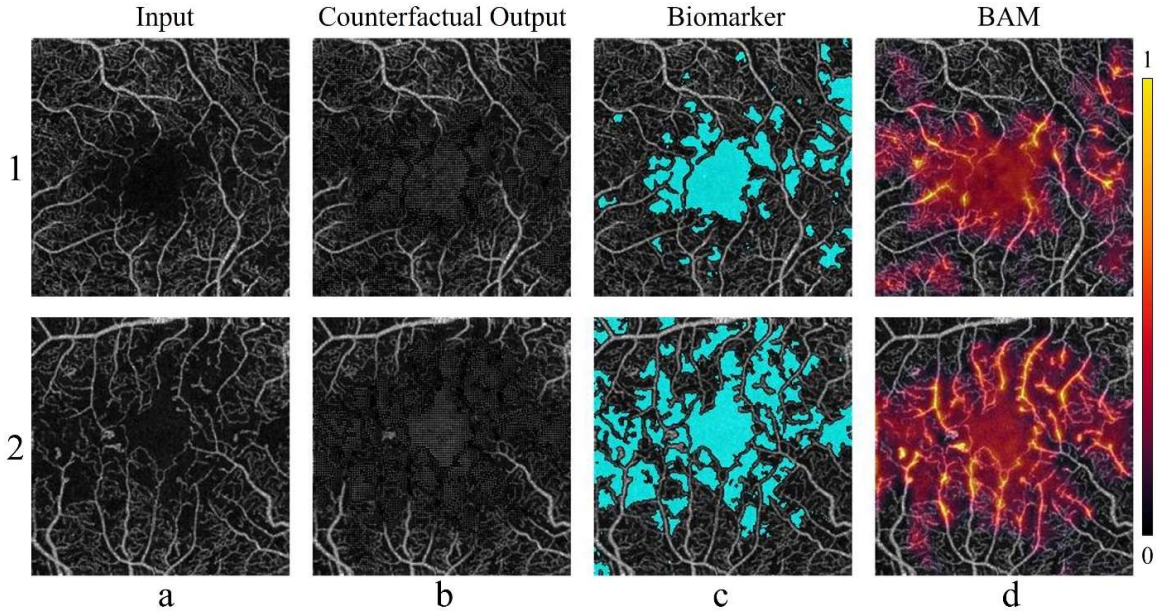

Fig. 5. BAMs for the referable DR classifier on OCTA. (a) Two representative referable DR inputs. (b) The residual counterfactual outputs of the BAM generator are modified to produce images that would be diagnosed as non-referable DR by the same classifier used on the inputs. To make the changes in the counterfactual output more easily identifiable, the change values have been amplified by a factor of five. (c) nonperfusion areas (NPAs) detected on OCTA, which are critical pathology for DR diagnosis on OCTA. (d) BAM, which highlights the classifier-utilized biomarkers. The highlighted areas are highly correlated with the NPA. In addition, some abnormal vessels are also highlighted.

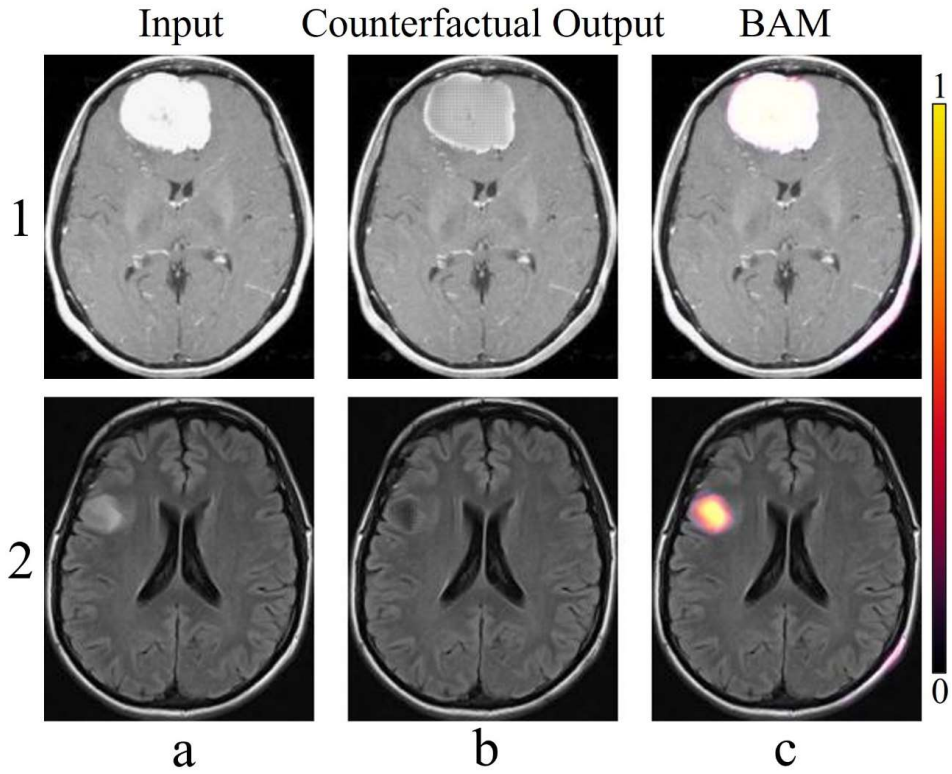

Fig. 6. BAMs for the brain tumor classifier on MRI. (a) Two representative brain tumor inputs. (b) The residual counterfactual outputs of the BAM generator, which are modified to produce images that would be diagnosed as non-tumor by the same classifier used on the inputs. (c) The BAM which highlights the classifier-utilized biomarkers. The large bright tumor in the first input and small dark tumor in the second input are both accurately and sharply highlighted.

In CT images, breast cancer is typically identified by the

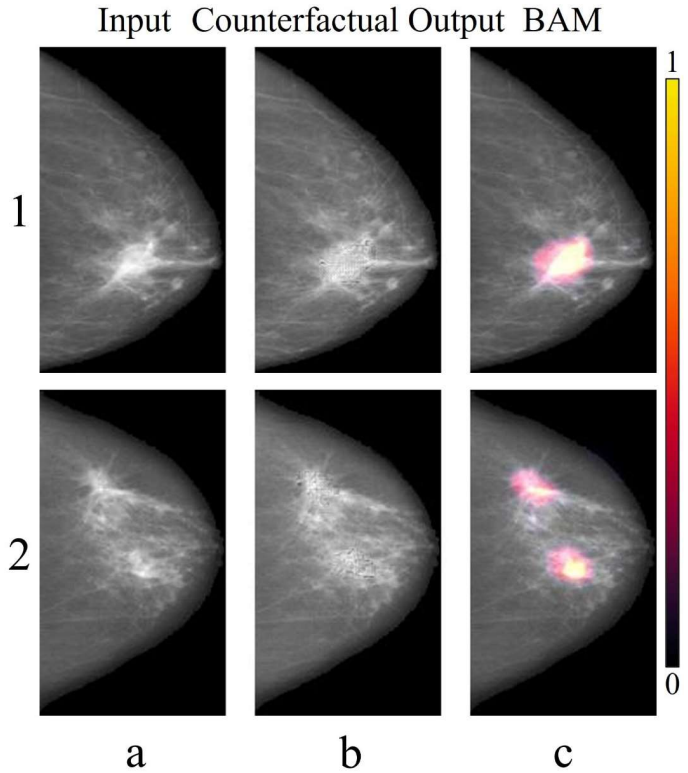

Fig. 7. BAMs for the breast cancer classifier on CT. (a) Two representative breast cancer inputs. (b) The residual counterfactual outputs of the BAM generator, which are modified to produce images that would be diagnosed as non-cancer by the same classifier used on the inputs. To make the changes in the counterfactual output more easily identifiable, the change values have been amplified by a factor of five. (c) The BAM which highlights the classifier-utilized biomarkers. The radiopaque areas with abnormal shapes are highlighted in both inputs.

presence of abnormal masses or lesions within the breast tissue. These masses can represent various types of breast cancer, including invasive ductal carcinoma and invasive lobular carcinoma. They commonly appear as regions of higher density compared to the surrounding tissue. The masses often exhibit irregular or spiculated borders, which suggest malignancy. Additionally, calcifications may be present within or around these masses, appearing as small, high-density spots on the CT scan [39]. The BAMs generated for the breast cancer classifier based on CT highlighted radiopaque areas with abnormal shapes (Fig. 7). But some small and darker radiopaque areas were not highlighted. This can be caused by an imperfect classifier or the highlighted biomarkers already sufficient for an accurate diagnosis.

G. *Clinical Applicable Analysis*

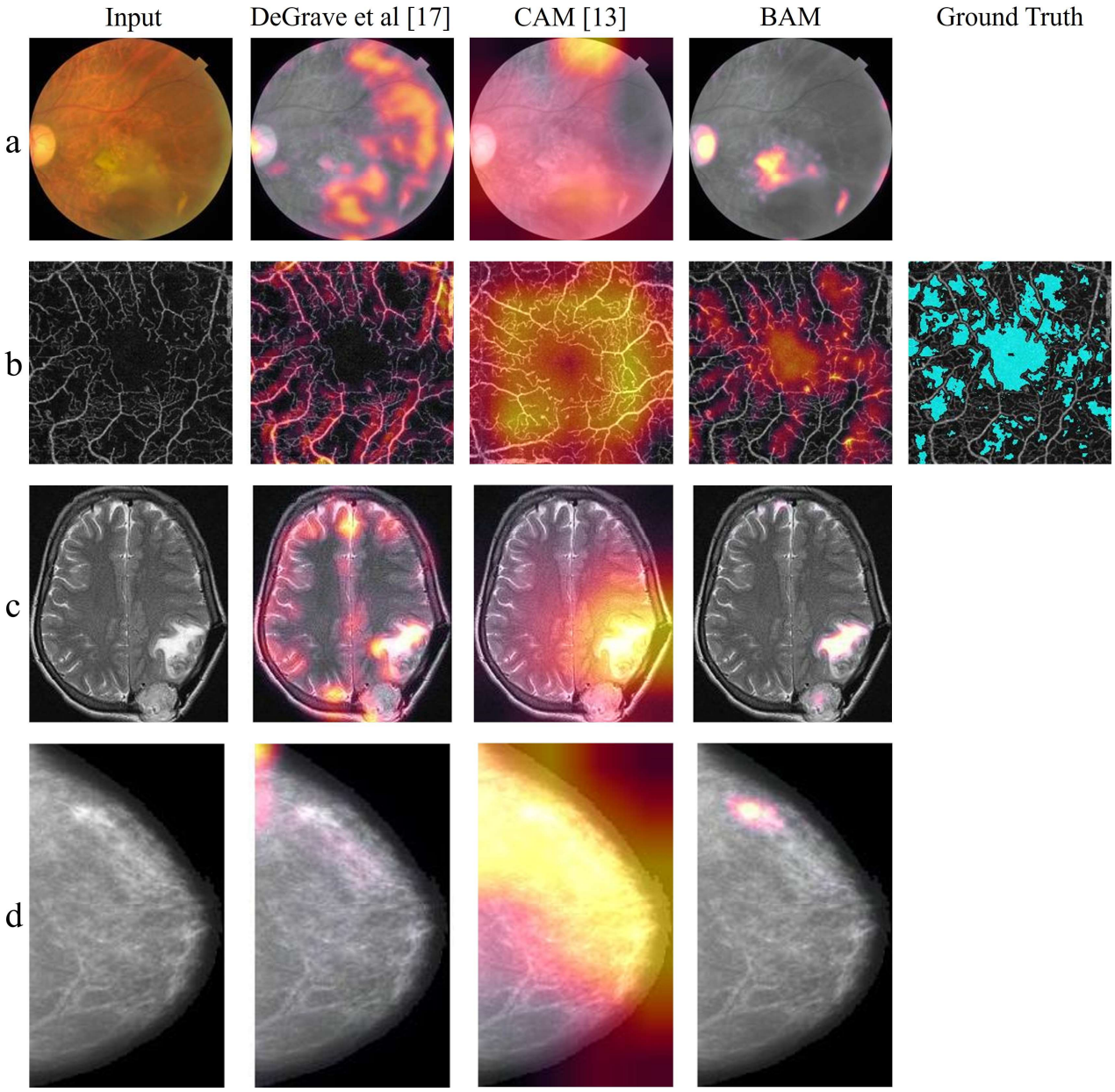

Fig. 8. Qualitative comparison of activation maps from DeGrave et al. (counterfactual audit) [17], class activation map (CAM) [13], and BAM across four correctly classified cases. (a) AMD, CFP: DeGrave et al. and CAM highlight broad nonpathological peripheral regions and do not clearly localize macular neovascularization, whereas BAM concentrates on the lesion while suppressing background. (b) DR, OCTA: DeGrave et al. emphasizes large nonpathological vessels and CAM produces diffuse, near-global activation; BAM focuses on nonperfusion areas and abnormal vessels—the hallmarks of DR—with higher correspondence to the ground-truth nonperfusion mask. (c) Brain tumor, MRI: DeGrave et al. activates both tumor and background and CAM overextends beyond tumor margins; BAM sharply delineates the tumor. (d) Breast cancer, CT: DeGrave et al. and CAM respond to masses and surrounding tissue, whereas BAM selectively highlights the malignant masses.

TABLE III
CLINICIAN PREFERENCE FOR THE BAM METHOD (%) OVER DEGRAVE ET AL. [17]

	CFP	OCTA	MRI	CT	All imaging modality
Clinician 1	100%	80%	90%	50%	80%
Clinician 2	90%	85%	95%	90%	90%
Average	95%	83%	93%	70%	85%

TABLE IV
QUANTITATIVE COMPARISON

	F1-score	IoU	Precision	Recall
BAM	**0.40 ± 0.09**	**0.25 ± 0.07**	**0.30 ± 0.10**	0.68 ± 0.12
DeGrave et al. [17] (counterfactual audit)	0.07 ± 0.05	0.04 ± 0.03	0.06 ± 0.05	0.09 ± 0.06
CAM [13]	0.30 ± 0.11	0.18 ± 0.07	0.18 ± 0.07	**0.99 ± 0.01**

To further analyze BAM in a real-world clinical setting, we compared it with a peer-reviewed, clinician-in-the-loop counterfactual framework (DeGrave et al., Nature Biomedical Engineering, 2023), selected for its rigorous clinical design and high-impact venue [17]. We asked two senior clinicians to evaluate the BAMs and the corresponding counterfactual activation maps generated from the same inputs (Fig. 8). For each classifier, 20 inputs were randomly selected for this blind experiment. For each input, our BAM and the activation map based on DeGrave et al. (counterfactual audit) were randomly inserted on the left or right side in a side-by-side image. The two clinicians were asked to pick the one with more clinically meaningful explainability. Our BAM achieved higher preference across all four classifiers (Table III).

To quantitatively compare BAM with DeGrave et al. (counterfactual audit) and the widely used numerical method, class activation map (CAM) [13, 40-41], we calculated the F1-score, IoU, precision, and recall between the segmented biomarkers and the binary masks derived from each activation map (Table IV). For each method, the binary mask was generated using the threshold that yielded the highest average F1-score. This quantitative comparison was conducted only on the OCTA dataset, which was acquired at our hospital and includes annotations of non-perfusion areas—clinically important biomarkers for rDR diagnosis. Our BAM demonstrated significantly superior performance across most of the evaluation metrics (Table IV). Recall was the only metric on which BAM did not surpass CAM. This precision–recall profile—higher precision with slightly lower recall—indicates that BAM concentrates on classifier-relevant DR biomarkers while suppressing activations in healthy tissue. In contrast, CAM attains higher recall at the cost of substantially lower precision because it highlights broad regions that include large amounts of non-pathologic tissue, yielding diffuse maps that are less clinically informative (Fig. 8).

H. Ablation Study

To evaluate the necessity of the proposed architecture and the contribution of each component, we compared BAMs generated from our full architecture with those from three ablated versions: (1) without the attention block, (2) without the assistant generator, and (3) without both. As shown in Supplementary Fig. 1, the BAMs produced by the ablated models often either highlighted irrelevant features not used by the classifier or failed to highlight critical features. For example, in BAMs for both CFP and CT classifiers without attention blocks, we observed spurious activations unrelated to any biomarkers. When the assistant generator was removed (CFP, OCTA, and CT), features correlated with disease but not utilized by the classifier—such as healthy vessels in OCTA images—were incorrectly highlighted. In the case of MRI, all three ablations resulted in partial omission of the relevant biomarkers.

VI. DISCUSSION

In this work, we developed an explainability method designed specifically for medical imaging and demonstrated its performance in four different imaging modalities. We generated the BAMs for disease classifiers focused on CFP-based AMD diagnosis, OCTA-based referable DR diagnosis, MRI-based brain tumor diagnosis, and CT-based breast cancer diagnosis. We first evaluated the BAM's ability to identify classifier-utilized biomarkers. The differences in highlighting between BAMs on predicted negative and positive inputs indicated that our BAM could successfully differentiate the unique biomarkers found only in positive inputs from other tissues shared among all the inputs. We then analyzed the correlations between highlighting in the BAMs and the corresponding pathologies of each diagnostic task. To evaluate the clinical applicability of our BAM, we asked clinicians to pick the most clinically meaningful maps between BAM and the maps generated from DeGrave et al. (counterfactual audit) [17]. Our BAM achieved higher preference across all diagnostic tasks. We believe these results are due to two major concepts we proposed in this study. First, the counterfactual output should be generated based on a residual mechanism, and second the model should focus solely on the unique biomarkers specific to the positive (disease or higher severity) input.

While existing counterfactual explanatory methods for machine decision-making can highlight local regions influencing classification, these regions may not map to the specific biomarkers the classifier utilizes. For medical imaging, a general counterfactual framework may not produce highlights that would help clinicians quickly and accurately assess the validity of classifier outputs. Instead, Our BAM framework restricts explainability maps to highlight only the most significant classifier-utilized biomarkers, which are usually clinical pathologies for classifiers with acceptable performance. As a result, our method allows for an interpretable output that demonstrates which biomarkers were or were not utilized by the classifier for decision-making. This

is a general strategy that can be employed across multiple imaging modalities, as we demonstrated in this work.

We designed the BAM generation framework to highlight the unique biomarkers that only belong to the positive class for a disease classifier. Consequently, our BAM provides more reasonable explanations for classifier decision-making than current counterfactual explanation methods proposed for AI-based disease classifiers. Providing clinically meaningful highlighting differentiating biomarkers utilized and ignored by the classifier was the top priority in our development. The result is that our BAM can highlight the classifier-utilized biomarkers accurately and sharply. Additionally, the highlighted biomarkers in BAMs can help clinicians quickly discern whether the classifier utilizes clinically meaningful biomarkers for each case.

There are many ways to generate a counterfactual output that can switch the decision-making of a disease classifier. However, there is only one clinically meaningful way: generating the output by changing the classifier-utilized biomarkers. We generated our output based on counterfactual and semantic segmentation losses to optimize the framework to better differentiate the classifier utilized or ignored biomarkers. The counterfactual loss was used to ensure the output could switch the decision-making of the classifier. At the same time, the segmentation ensured that the differences between the output and input were exclusively limited to the classifier-utilized biomarkers. Additionally, to make sufficient changes to the highlighted biomarkers, we improved the generator's ability to capture global correlations by adding an attention mechanism to the previous architecture [15].

To improve the generalizability of our BAM generation framework, we developed and evaluated four different disease classifiers, each utilizing a different imaging modality. In both development and evaluation, the architectures and most of the hyperparameters were the same across these four classifiers. Only some basic hyperparameters were different (e.g., learning rate). This consistency improved the reproducibility of the BAM on classifiers for other diseases and imaging modalities.

Beyond providing clinically meaningful explainability to different AI-based disease diagnostic systems, our BAM can also be used to explore new biomarker findings in research. For example, in a study exploring the diagnostic potential of a new imaging modality on a known disease with ground truth classes, we could identify relevant biomarkers by generating BAMs for this disease classifier trained on the new imaging modality. Furthermore, our BAM can also facilitate understanding new treatments by highlighting all the relevant differences after intervention.

One disadvantage of the BAM and other counterfactual explanation methods is that a new model needs to be trained for each new classifier. This issue makes deploying such methods more complicated than the numerical methods. In a future study, we plan to add the outputs of several hidden layers of the classifiers as inputs. The new BAM generation model could be trained using medical images and hidden layer outputs as inputs. The model can generate BAMs for an unseen classifier during inference without requiring retraining by training with multiple classifiers.

VII. Conclusion

We developed a method that can provide clinically applicable explainability to AI-based medical imaging classifiers. For each diagnostic task in this study, the BAM successfully differentiated classifier-utilized biomarkers from other unutilized biomarkers. Before the deployment of a classifier in a real-world clinical setting, the BAM could be used to verify whether clinically meaningful biomarkers are basically learned by the classifier in most cases in an existing data set. Clinicians then could use the verified BAM to quickly verify each AI diagnosis by reviewing the highlighted classifier-utilized biomarkers.

Acknowledgment

Oregon Health & Science University (OHSU), Yali Jia has significant financial interests in Visionix, Inc. and Optos, Inc. These potential conflicts of interest have been reviewed and managed by OHSU. Pengxiao Zang is currently employed by Topcon Healthcare Inc, this employment had no influence on the design, execution, or interpretation of the study.

References

[1] Y. LeCun et al., "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May. 2015.

[2] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60-88, Dec. 2017.

[3] D. Shen et al., "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.* vol. 19, no. 1, pp. 221–248, Jun. 2017.

[4] S. Suganyadevi et al., "A review on deep learning in medical image analysis," *Int. J. Multimed. Inf. Retr.* vol. 11, pp. 19–38, Sep. 2022.

[5] T. T. Hormel et al., "Artificial intelligence in OCT angiography," *Prog. Retin. Eye Res.*, vol. 85, pp. 100965, Nov. 2021.

[6] J. He et al., "The practical implementation of artificial intelligence technologies in medicine," *Nat. Med.*, vol. 25, no. 1, pp. 30–36, Jan. 2019.

[7] S. Gerke et al., "Ethical and legal challenges of artificial intelligence-driven healthcare," in *Artificial Intelligence in Healthcare*, 1st ed. Amsterdam, Netherlands: Elsevier Academic Press, 2020, pp. 295–336.

[8] Z. Salahuddin et al., "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, pp. 105111, Jan. 2022.

[9] Q. Zhang, and S. C. Zhu, "Visual interpretability for deep learning: a survey," *Front. Inf. Technol. Electron. Eng.*, vol. 19, pp. 27–39, Jan. 2018.

[10] D. V. Carvalho et al., "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, pp. 832, Aug. 2019.

[11] P. Linardatos et al., "Explainable AI: a review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, pp. 18, Dec. 2020.

[12] M. Sundararajan et al., "Axiomatic attribution for deep networks," in *Proc. ICML PMLR*, 2017, pp. 3319–3328.

[13] R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[14] B. K. Iwana et al., "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," in *Proc. IEEE/CVF ICCVW*, 2019, pp. 4176-4185.

[15] P. Zang et al., "Interpretable diabetic retinopathy diagnosis based on biomarker activation map," *IEEE Trans. Biomed. Eng.*, vol. 7, no. 1, pp. 14–25, Jul. 2023.

[16] A. Fontanella et al., "ACAT: Adversarial counterfactual attention for classification and detection in medical imaging," *arXiv preprint arXiv:2303.15421*, 2023.

[17] A. J. DeGrave et al., "Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians," *Nat. Biomed. Eng.*, vol. 7, no. 3, pp. 1–13, Mar. 2023.

[18] A. Dravid et al., "medXGAN: visual explanations for medical classifiers through a generative latent space," in *Proc. IEEE/CVF CVPR*, 2022, pp. 2936-2945.

[19] A. Katzmann et al., "Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization," *Neurocomputing*, vol. 458, pp. 141–156, Oct. 2021.

[20] S. Singla et al., "Explaining the Black-box Smoothly: A Counterfactual Approach," *Med. Image Anal.*, vol. 84, pp. 102721, Feb. 2023.

[21] J. P. Cohen et al., "Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays," in *Proc. Med. Imag. Deep Learn. PMLR*, 2021, pp. 74-104.

[22] J. J. Thiagarajan et al., "Training calibration-based counterfactual explainers for deep learning models in medical image analysis," *Sci. Rep.*, vol. 12, no. 1, pp. 597, Jan. 2022.

[23] V. Boreiko et al., "Visual explanations for the detection of diabetic retinopathy from retinal fundus images," in *Proc. MICCAI*, 2022, pp. 539-549.

[24] S. Pachade et al., "Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research," *Data*, vol. 6, no. 2, pp. 14, Feb. 2021.

[25] Kaggle. Ocular Disease Recognition [Data set]. Kaggle. https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k (2019).

[26] P. Zang et al., "A diabetic retinopathy classification framework based on deep-learning analysis of OCT angiography," *Transl. Vis. Sci. Technol.*, vol. 11, no. 7, pp. 10–10, Jul. 2022.

[27] A. Hamada, Br35H: Brain Tumor Detection 2020 [Data set]. Kaggle. https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection (2020).

[28] S. Malekzadeh, and G. Faramarzi, Breast Cancer CT (Fully Preprocessed) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/2026269 (2022).

[29] J. Y. Zhu, et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE ICCV*, 2017, pp. 2223–2232.

[30] O. Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234-241.

[31] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998-6008.

[32] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR*, 2022, pp. 10684-10695.

[33] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] P. Kovesi, "Good colour maps: How to design them," *arXiv preprint arXiv:1509.03700*, 2015.

[35] E. Agrón et al., "An updated simplified severity scale for age-related macular degeneration incorporating reticular pseudodrusen: age-related eye disease study report number 42," *Ophthalmology*, vol. 131, no. 10, pp. 1164-1174, Oct. 2024.

[36] T. Y. Wong et al., "Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings," *Ophthalmology*, vol. 125, no. 10, pp. 1608–1622, Oct. 2018.

[37] T. T. Hormel, and Y. Jia, "OCT angiography and its retinal biomarkers," *Biomed. Opt. Express*, vol. 14, no. 9, pp. 4542–4566, Aug. 2023.

[38] M. K. Abd-Ellah et al., "A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned," *Magn. Reson. Imaging*, vol. 61, pp. 300–318, Sep. 2019.

[39] S. J. Glick, "Breast CT," *Annu. Rev. Biomed. Eng.*, vol. 9, no. 1, pp. 501–526, Aug. 2007.

[40] S. Nazir et al., "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Comput. Biol. Med.*, vol. 156, pp. 106668, Apr. 2023.

[41] M. Ennab, and H. Mcheick, "Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 1, pp. 12, Feb. 2025.