MedLIME: A Distribution-Aligned and Evidence-Supported Framework for Medical Saliency Explanations

Supplementary Material

8.1. Dataset Details

We perform evaluations on several chest X-ray datasets due to availability of rich open source benchmarks. These include RSNA Pneumonia Detection Challenge dataset [2], SIIM-ACR Pneumothorax Segmentation dataset [42], ChestX-Det10 dataset[20], CheXlocalize dataset [29] and VinDr-CXR[23]. Also, we show results on Breast Ultrasound Images Dataset (BUID) [1] and Kvasir-Capsule [31]. Also we show results on DeepPCB dataset [36] to evaluate our method on non-medical datasets. Here, we provide a detailed overview of each of the datasets.

- **Breast Ultrasound Images Dataset**: This dataset has ultrasound images with and without breast cancer. The images are categorized into three classes, which are normal, benign, and malignant. We used only the normal and malignant images in our experiments. There are total of 194 train images with 110 abnormal images and test set has 150 images with 100 abnormal images.
- **CheXlocalize**: This dataset consists of pixel-level segmentation maps and most representative points for chest X-rays with 10 pathologies which includes Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion,Edema etc. These annotations are on images from CheXpert [13] validation and test sets. We sample 6K images from the training data, of which 3754 are abnormal, which have over 200K images. In the test set, there are 518 images with 450 abnormal images. The ground truth localization maps are precise segmentation mask with multiple abnormalities.
- **ChestX-Det10**: This dataset consists of 3543 chest X-ray images covering 10 types of abnormalities namely Atelectasis, Calcification, Consolidation, Effusion, Emphysema, Fibrosis, Fracture, Mass, Nodule, and Pneumothorax. In train set we take 3001 images with 2320 abnormal images and we have 542 test images in which 459 are abnormal. The ground truth localization maps are in the form of segmentation maps, but on average are large than those in SIIM-ACR dataset
- **RSNA Pneumonia Detection Challenge**: This dataset is focused on the presence and absence of Pneumonia in chest X-ray images. It consists of a total of 14863 images, which are split into train, validation and test. The training set comprises 12039 images with 4870 abnormal images, while the validation set contains 1338 images with 541 abnormal images and the test set contains 1486 images with 601 abnormal images. All abnormal images in the test set have respective abnormality localization maps. The maps are in the form of 2D bounding boxes.
- **SIIM-ACR Pneumothorax Segmentation Dataset**. This dataset labels images based on the absence or presence of Pneumothorax. It has 10675 images with training, validation and test sets having 8646, 961 and 1068 images in total and 1931, 202 and 246 abnormal images respectively. Here, the test set is provided with precise segmentation maps for the abnormal regions
- **VinDr-CXR**: This dataset consists of chest X-rays covering over 28 different radiographic findings and diagnoses. The dataset has over 18K images in total. We sample 5000 train images with 2500 abnormal images and in the test set we have 800 images with 400 abnormal ones. The ground truth saliency maps are in bounding boxes, but with multiple of them in a single image.
- **Kvasir-Capsule**: This is a video capsule endoscopy dataset with 118 endoscopy videos from which 4,820,739 frames have been derived. 44,228 of these frames are labeled with bounding boxes around the abnormality. We sample 6000 images for our training set, of which 3000 have abnormality. For the test set we have 800 images, with 400 abnormal images. The dataset includes 13 abnormalities including Ampulla of Vater, Angiectasia, Erythema, Ileocecal valve, Polyp, Pylorus, Ulcer, Lymphangiectasia etc. The test set has bounding boxes around the abnormal region.
- **DeepPCB**: This is an anomaly detection dataset consisting of 1500 pairs of PCB (Printed Circuit Board) images. One image in the pair is a defect free image and the other has one of the 6 most common PCB defects : open, short, mouse-bite, spur, pin hole and spurious copper. The defective images are accompanied with the respective segmentation maps of the defect. We sample 1000 image pairs for training and keep the remaining 500 pairs for evaluation.

Figure 11 shows images and ground truth abnormality localization map samples from the datasets.

8.2. Implementation Details

8.2.1. Models

The pretrained models for the three architectures, namely Inception-V3, ViT and SwinViT are taken from `torchvision.models` with the weights parameter set to `IMAGENET1K_V1`. For Inception-V3 we use `models.inception_v3`, for ViT we use `vit_base_patch16_224` and for SwinViT we use `models.swin_b`

Dataset	Model	XRAI	IG	Grad CAM	Layer CAM	Finer CAM	LIME	S-LICE	G-LIME	Bay LIME	Med LIME
SIIM-ACR	Incep-V3	0.047	0.018	0.024	0.021	0.027	0.021	0.032	0.023	0.020	**0.112**
	ViT	0.041	0.018	0.018	0.013	0.014	0.028	0.034	0.030	0.026	**0.094**
	SwinViT	0.117	0.042	0.043	0.054	0.053	0.036	0.029	0.035	0.032	**0.123**
	Average	0.068	0.026	0.028	0.029	0.031	0.028	0.032	0.029	0.026	**0.110**
VinDr-CXR	Incep-V3	0.264	0.188	**0.283**	0.260	0.260	0.149	0.193	0.154	0.110	0.241
	ViT	0.134	0.149	0.117	0.210	0.095	0.128	0.194	0.137	0.101	**0.250**
	SwinViT	0.259	0.181	**0.282**	0.239	0.126	0.158	0.174	0.178	0.103	0.272
	Average	0.219	0.173	0.227	0.236	0.160	0.145	0.187	0.156	0.105	**0.254**
Kvasir-Capsule	Incep-V3	0.244	0.175	0.150	0.253	0.220	0.169	0.186	0.156	0.145	**0.301**
	ViT	0.230	0.187	0.142	0.163	0.135	0.178	0.191	0.164	0.151	**0.332**
	SwinViT	0.339	0.215	0.252	0.378	0.335	0.230	0.267	0.221	0.179	**0.472**
	Average	0.271	0.192	0.181	0.265	0.230	0.192	0.215	0.180	0.158	**0.368**
DeepPCB	Incep-V3	0.131	0.052	0.071	0.048	0.048	0.066	0.038	0.039	0.041	**0.210**
	ViT	0.155	0.100	0.093	0.101	0.045	0.076	0.045	0.035	0.030	**0.330**
	SwinViT	0.182	0.091	0.137	0.064	0.039	0.096	0.039	0.027	0.035	**0.230**
	Average	0.156	0.081	0.100	0.071	0.044	0.079	0.041	0.034	0.035	**0.257**

Table 4. **Quantitative comparison against explainability methods.** We report the average AUPRC score on the test set for each dataset and three model architectures. Bold indicates the best score per row; underlined indicates the second-best.

8.2.2. Hyperparameters and Configurations

In this section, we provide details of hyperparameter values used in MedLIME , which have been discussed throughout the main paper.

In the Supervised Test Time Adaptation (STTA) stage, we apply

- `RandomRotation(degrees=10)`
- `RandomResizedCrop(size=224, scale=(0.9, 1.0))`

from `torchvision.transforms` package to get $S = 100$ training samples. We train the models for 20 epochs with a learning rate of 1e-4

Next in the Evidence-based Regularization phase we first resize the image to $(224, 224)$ and partition it into non-overlapping square patches of side $s = 16$. Hence, we get $P = 196$ patches. Next we take the nearest $N_T \in [300, 800]$ neighbors for each test sample and apply $N = 4000$ binary masks, with different masking ratios in the range $[0.3, 0.7]$, on each of the neighbors and pass them through the MAE for Generative Masking and then get their abnormality localized by vanilla LIME. The Gaussian Bandwidth for this step is in the range $h \in [1, 5]$

Finally in the MedLIME Estimation stage, Generative masking is performed as in the previous stage. The reconstructed images are passed through the optimized module ϕ^* and then the backbone $g()$ to finally get the MedLIME saliency map. Here, the coefficients of the regularization term λ_1, λ_2 are dataset specific. Typically $\lambda_1 \in [1, 3]$ and $\lambda_2 \in [5, 15]$

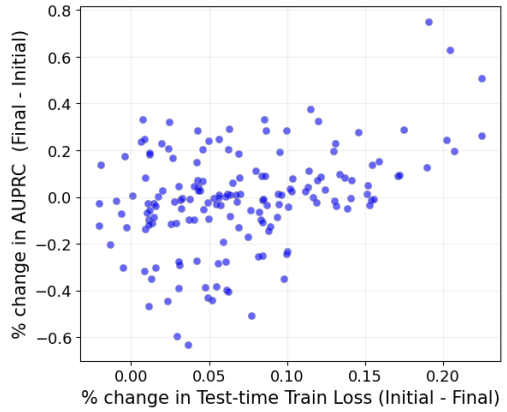

Figure 9. **Scatter plot between the percentage reduction in test-time training loss and the percentage increase in AUPRC**. The plot shows that test samples with larger decrease in test-time train loss have greater improvements in abnormality localization accuracy. The Pearson correlation across the RSNA test set is 0.39.

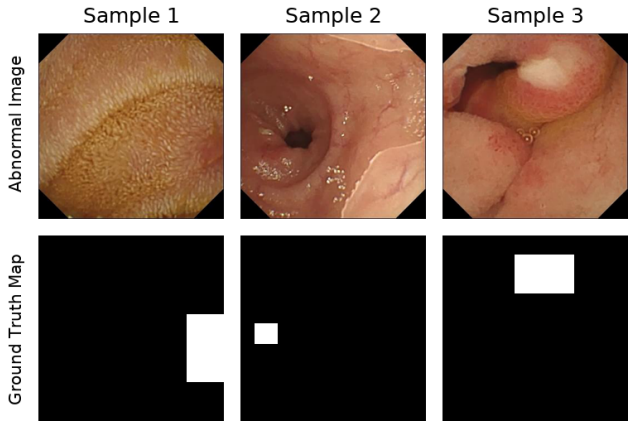

Figure 10. **Kvasir-Capsule Dataset.** The figure presents three abnormal image samples with their ground truth masks

8.3. Evaluation on More Datasets

Table 4 presents the evaluation result on SIIM-ACR, VinDr-CXR, Kvasir-capsule and DeepPCB dataset. SIIM-ACR is a challenging dataset due to its very small ground truth localization maps, as can be seen in Figure 11. Our method outperforms the baselines, with average gains in the range of 5%-7%. In the case of VinDr-CXR MedLIME outperforms the baselines by 2%-15% on average across architectures. In case of the Kvasir-Capsule dataset MedLIME achieves a gain of around 10% over the second best baseline. This shows that MedLIME generalizes not only across model architectures and datasets from different medical domains like X-ray, ultrasound images and endoscopy images.

8.4. Comparison with Perturbation-based Methods

In this section we present a comparison between MedLIME and RISE [25]. Table 5 shows that MedLIME outperforms RISE across all models and for all benchmarks

8.5. Computation Time

Since MedLIME follows the LIME pipeline, its computational cost is also dominated by perturbation sampling and repeated model queries as LIME. Our total per-image runtime is in a reasonable range and shown in Table 6

8.6. Visualization of Evidence-Based Regularizer

. Motivated by the fact that clinicians use historical evidence to diagnose diseases in patients, we introduced Evidence-Based Regularization (EBR) into our pipeline. Figure 12 provides a visualization of the EBR map along with the final MedLIME map obtained for an image from the ChestX-Det10 dataset. The image shows how the EBR map serves as a good starting point on top of which the final map is estimated.

Dataset	Model	RISE	MedLIME
RSNA	Incep-V3	0.155	0.332
	ViT	0.246	0.451
	SwinViT	0.452	0.471
	Average	0.281	0.418
CheXlocalize	Incep-V3	0.136	0.401
	ViT	0.373	0.467
	SwinViT	0.424	0.486
	Average	0.311	0.451
BUID	Incep-V3	0.203	0.430
	ViT	0.273	0.451
	SwinViT	0.231	0.504
	Average	0.236	0.462
DeepPCB	Incep-V3	0.110	0.210
	ViT	0.140	0.330
	SwinViT	0.210	0.230
	Average	0.153	0.257

Table 5. **Quantitative comparison against RISE** We report the average AUPRC score on the test set for each dataset and three model architectures.

Method	Time
LIME	24s
MedLIME	39s
SLICE	84s

Table 6. Execution time comparison of MedLIME and LIME-based methods for ViT model

8.7. Visualization of Saliency Maps for various λ_1 and λ_2 values

The final step in the estimation of MedLIME involves two regularization coefficients, namely λ_1 and λ_2. In Figure 13 we provide a visualization of how varying these effects the final saliency map. The first row has the abnormal image and ground truth abnormality localization map followed by a saliency map generated using LIME and Generative Masking, and finally the EBR map for the particular sample. λ_1 is the coefficient of the sparsity loss and λ_2 is the coefficient of the distance from EBR loss. We can see that for higher λ_1 the maps tend to be sparser, while for higher λ_2 the maps tend to be similar to the EBR map

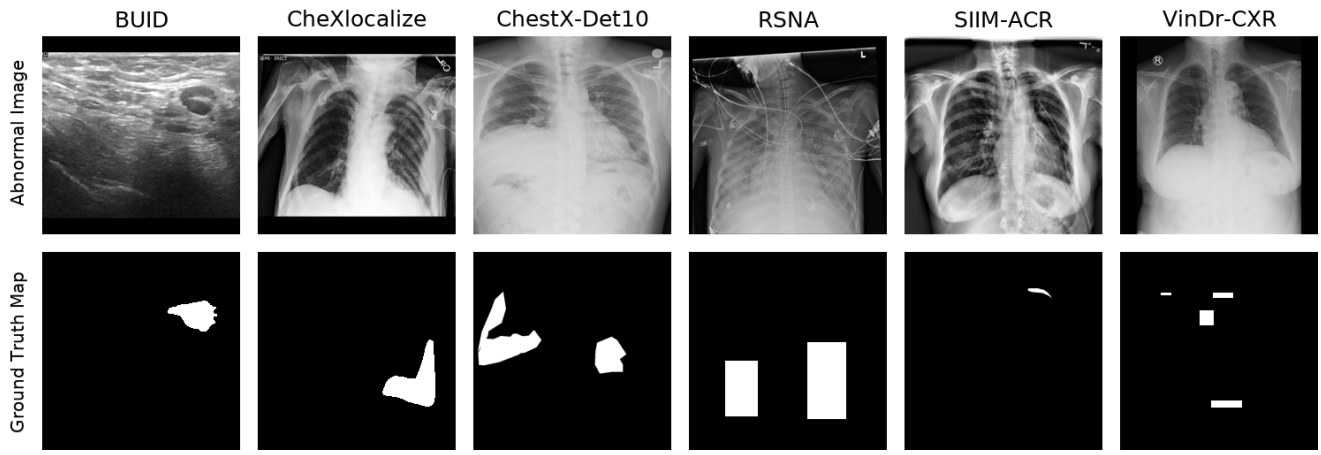

Figure 11. **Visual Comparison of datasets**. The figure shows abnormal images and the ground truth maps for various chest X-ray datasets that we shown results on. This highlights the varying complexity of each of the datasets as they have different sized and differently labeled abnormal regions

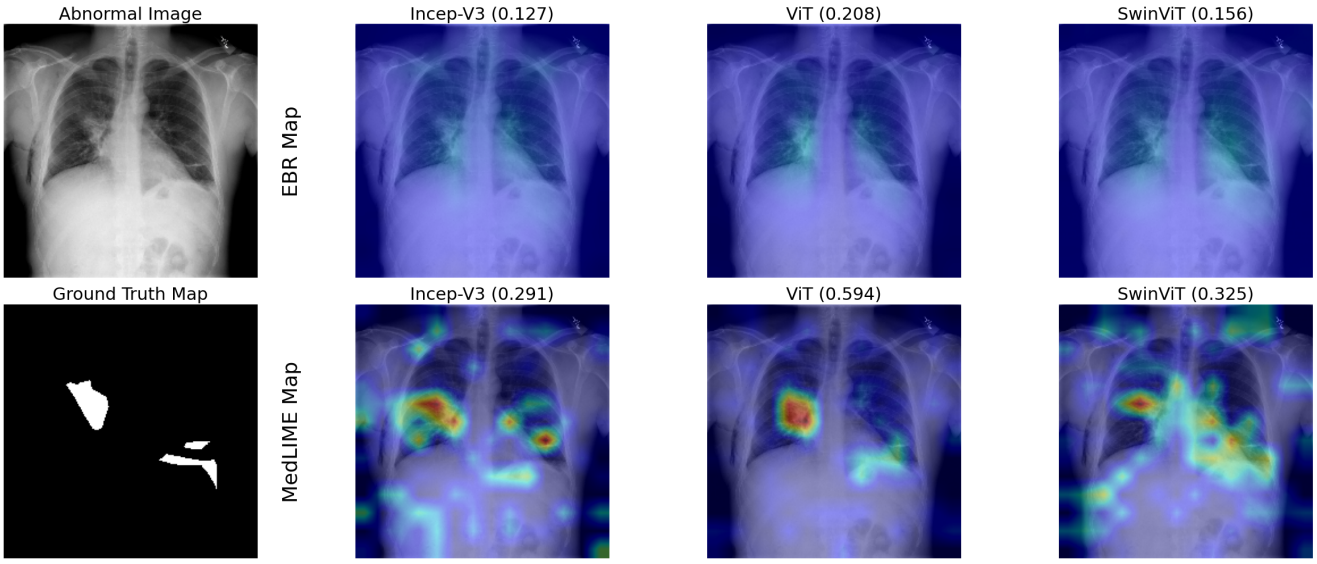

Figure 12. **Visual Comparison of EBR** The figure shows EBR map and the MedLIME map obtained for a image from the ChestX-Det10 dataset for three model architectures. The AUPRC with respect to the ground truth map is reported in parenthesis above the image

8.8. More Visual Results

We present a few more visual results comparing MedLIME with the rest of the baselines in Figure 14 and Figure 15.

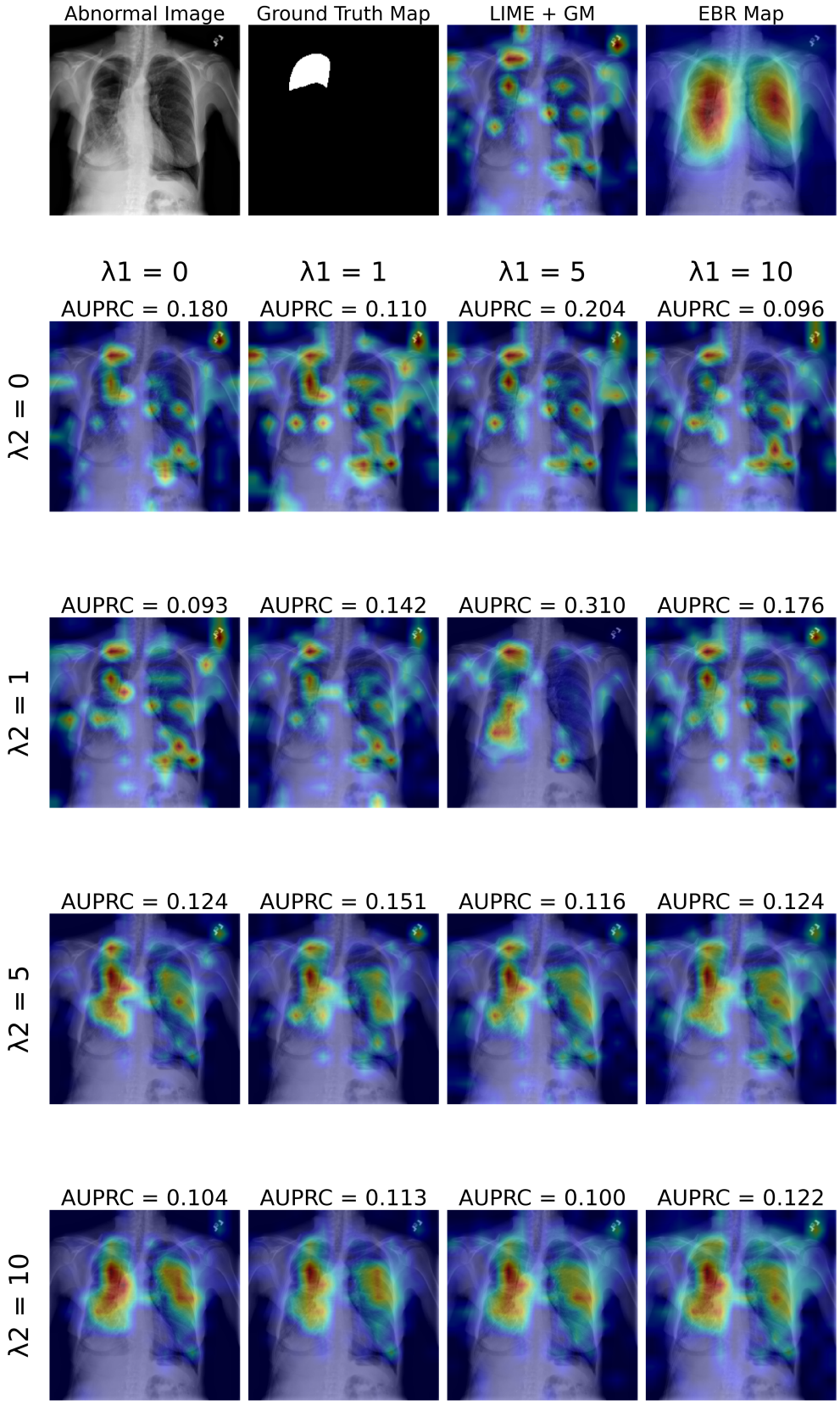

Figure 13. **Visual Comparison for different λ_1 and λ_2 values** Figure shows saliency maps obtained for different values of the hyperparameters in MedLIME estimation

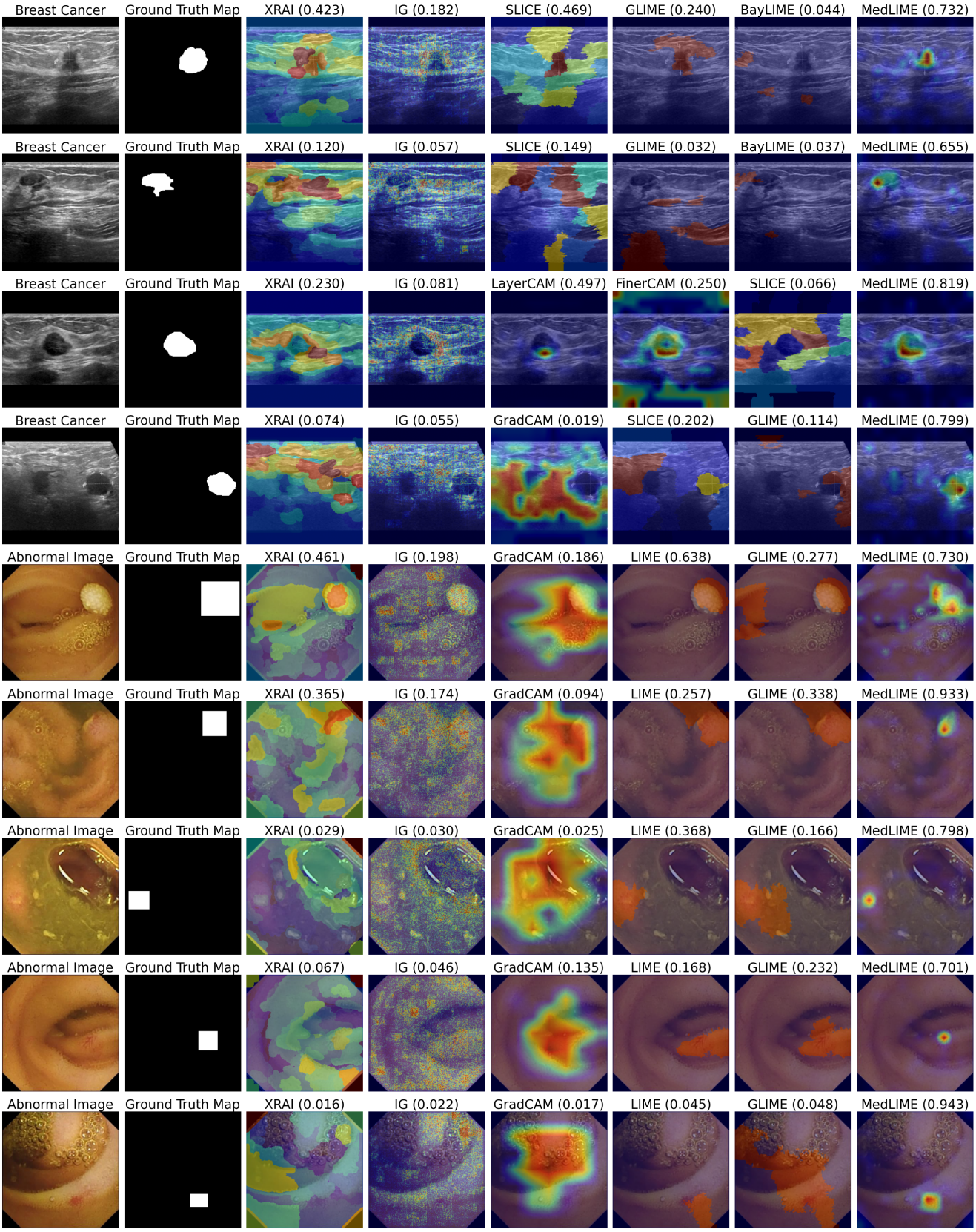

Figure 14. **Visual Results.** Comparison of MedLIME against baselines on Kvasir-Capsule and BUID datasets. Top 6 best performing methods are reported for each sample

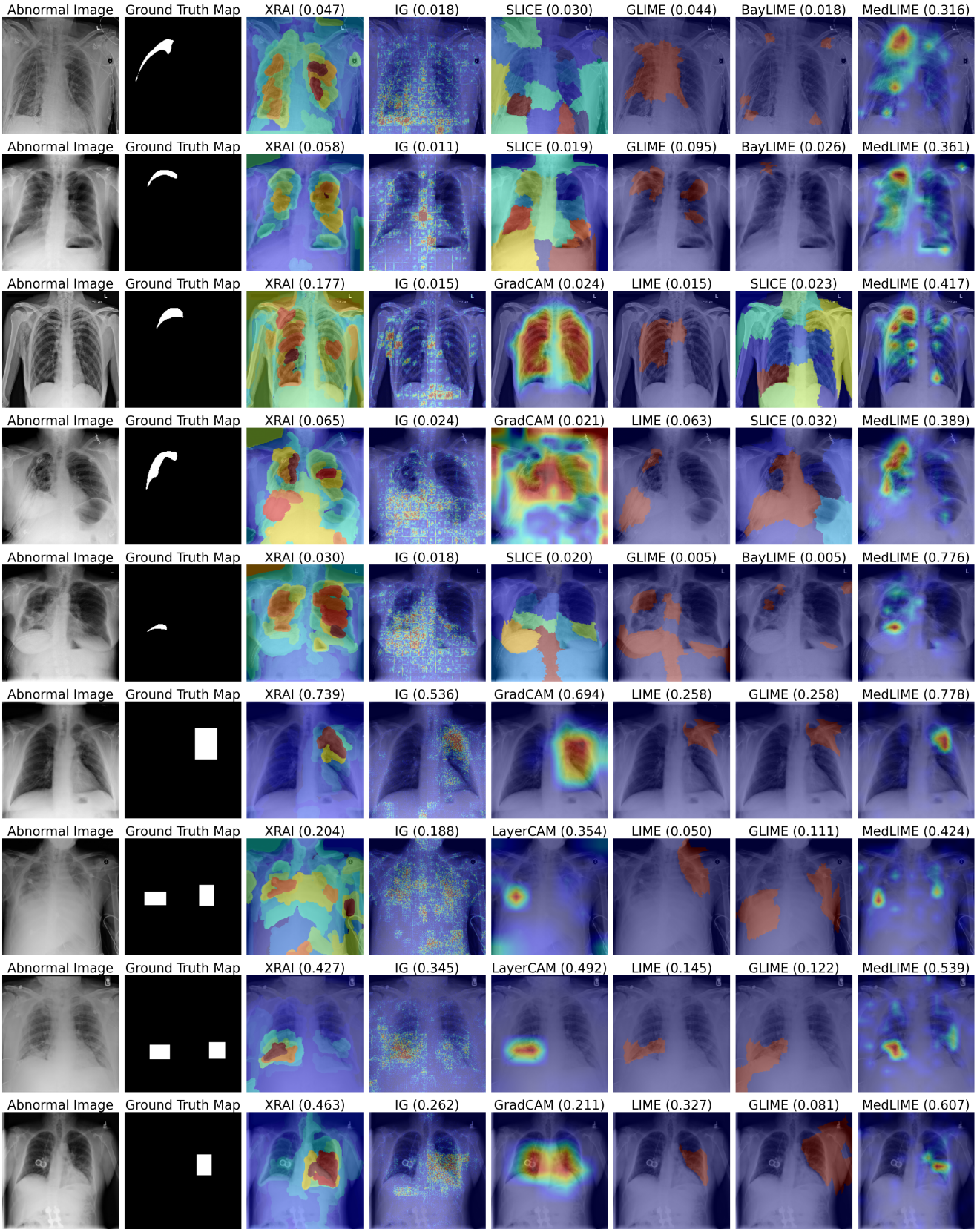

Figure 15. **Visual Results.** Comparison of MedLIME against baselines on chest X-ray datasets datasets. Top 6 best performing methods are reported for each sample