Interpretability of U-Net Model

1st Zahra Fazel
Department of Computer Science
Western University
London, Canada
zfazel@uwo.ca

2nd Sepehr Ashrafzadeh
Department of Computer Science
Western University
London, Canada
sashra29@uwo.ca

***Abstract*—Recent advances in imaging and compassion have led to a drastic rise in the use of machine learning for medical imaging. The advent of deep learning allows for much higher levels of abstraction for feature selection and discretization. Convolutional neural networks have been shown to learn abstractions obtained from multidimensional medical images and learning features hard to define by humans. This is one of the reasons why convolutional neural networks excel at object recognition and segmentation. However, most of these networks lack interpretability. The interpretability of deep learning models answers the question of why a neural network model provides a particular output. In medical image analysis, this is an essential matter to build confidence while using. For classification models, there are multiple interpretability techniques; however, there is not much research done on the interpretability of segmentation models. In this project, we trained the U-Net model to segment polyps in colonoscopy images and skin cancer lesions. We applied various interpretability techniques to this model and detected important features that the network's decision is based on. Then, using interpretability evaluation metrics, we showed that Grad-CAM has the most reliable outputs among the applied methods.**

***Index Terms*—U-Net Model, Segmentation, Interpretability, Deep Learning, Computer Vision, Medical Image Analysis**

I. Introduction

Image segmentation methods in medical imaging have diverted to deep learning solutions. The U-Net architecture has gained a lot of traction in medical imaging. One disadvantage shared by many neural networks including U-Net is a lack of interpretability. Because these neural networks interface with many convolutional layers simultaneously, it becomes challenging to visualize what features it is learning. This effectively renders the neural net a black box, which poses a challenge when attempting to find the root cause of misclassification, and gives an advantage to potential adversarial attacks.

A lack of model transparency and robustness will hinder its translation into a clinical setting. In medical image analysis, it is desirable to decipher the black-box nature of deep learning models in order to build confidence in clinicians. Interpretability techniques can help understand the model's reasoning.

However, understanding the model's decision criteria is not enough. For both scientific robustness and security reasons, it is critical that these explanations be reliable and to know to what extent can the interpretations be altered by small systematic perturbations to the input data, which might be generated by adversaries or by measurement biases.

In this project, we gain the advantage of existing methods for the interpretability of classification models and extend them to the segmentation problem. Then, we analyze the model's reasoning, each of its layers' focus, and the robustness of its interpretations. We evaluate our model on two public datasets KVASIR-SEG and HAM10000. Consequently, our method not only yields strong performance, it is also robust.

II. Related Work

In this section, we briefly review the U-Net model, several pixel attribution interpretation methods related to this work, and evaluation methods for interpretation.

A. U-Net Model

U-Net consists of a contracting path that captures feature information as well as a symmetric expansive path that enables localization. Moreover, U-Net uses skip connections from encoders to decoders of similar resolutions to pass high-resolution information throughout the network.

In the original proposal of U-Nets, as shown in Fig. 1, each block in the encoder is comprised of two successive normalized 3x3 convolutional layers. Then, a max-pooling layer with a 2x2 kernel of stride 2 is used to down-sample the image in order to obtain greater contextual and spatial information. In the up-sampling path, the same blocks in the encoder are used, except feature maps of similar resolutions from the down-sampling path are concatenated with the feature maps from the up-sampling path after being cropped and aligned to match dimensions before being filtered by a normalized 3x3 convolutional layer. The skip connections are used to obtain more fine-grained information that may have been lost during the intermediate stages. Following each decoder block, an up-sampling is applied to the feature maps to increase the resolution by 2. [2]

B. Gradient-based Methods

In general, pixel attribution methods highlight each pixel according to how much it changed the prediction negatively or positively. Gradient-based methods compute the gradient of prediction with respect to the input features. The difference between them is how they calculate the gradient. [1]

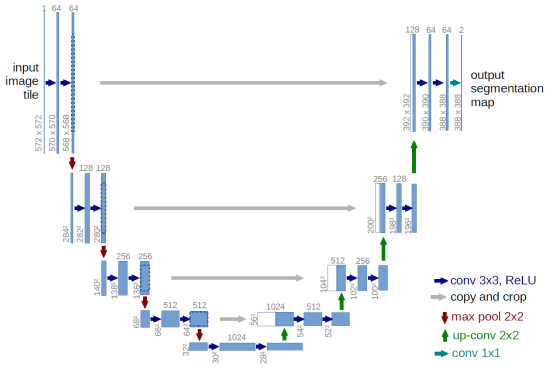

Fig. 1. U-Net Model Architecture [2]

1) Saliency Maps: Saliency maps or Vanilla Gradient was one of the first pixel attribution approaches. In this method, a forward pass of the image of interest is performed. Then, the gradient of the class score of interest with respect to the input pixels is computed by approximating the score with a first-order Taylor expansion. Finally, the gradients are visualized. [13]

Saliency maps have a saturation problem. When ReLu is used, and the activation goes below zero, then the activation is capped at zero and does not change anymore. The activation is saturated, and the saliency maps indicate that the corresponding neuron was unimportant. [7]

2) Guided Backpropagation: Guided Backpropagation combined saliency maps and another method called DeconvNet, reverses a neural network by using operations reversals of the filtering, pooling, and activation layers. This method zero's out the importance signal at a ReLU if either the input to the ReLU during the forward pass is negative or the importance signal during the backward pass is negative. Guided Backpropagation can be considered equivalent to computing gradients, with the caveat that any gradients that become negative during the backwards pass are discarded at ReLUs. [5]

Since this method zero's out negative gradients, it fails to highlight inputs that contribute negatively to the output. Moreover, this approach still does not address the saturation problem. [7]

3) Grad-CAM: Grad-CAM provides visual explanations for convolutional neural networks. Unlike previous methods, the gradient is not backpropagated all the way back to the image, but to the last convolutional layer to produce a coarse localization map that highlights important regions of the image. The goal is to find the localization map, which is defined as:

$$L^c_{Grad-CAM} \in \mathbb{R}^{u \times v} = ReLU\big(\sum_k (\frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{A^k_{ij}}) A^k\big) \quad (1)$$

Where u is the width, v is the height of the explanation, c is the class of interest, A_k is the kth feature map in the last convolutional layer, and y^c is the score for class before softmax. For visualization, the values are scaled to the interval between 0, and the image is upscaled and overlaid over the original image. Grad-CAM also can be applied to other convolutional layers other than the last convolutional layer. [6]

4) Guided Grad-CAM: The localization in Grad-CAM is very coarse since the last convolutional feature maps have a much coarser resolution compared to the input image. On the other hand, Guided Backpropagation gives much more detailed explanations since the prediction is backpropagated all the way to the input pixels. Guided Grad-CAM combines the last two methods by performing an element-wise product between the scores obtained from Grad-CAM and the scores from Guided Backpropagation to gain the advantages of both techniques. [6]

However, this strategy inherits the limitation of Guided Backpropagation caused by zero-ing out negative gradients during backpropagation. Moreover, both Grad-CAM and Guided Grad-CAM are specific to convolutional neural networks. [7]

C. Path-attribution Methods

These methods compare the current image to a reference image, which can be an artificial 'zero' image. The difference in actual and baseline prediction is divided among the pixels. The choice of the reference image has a big effect on the explanation. [1]

1) DeepLIFT: DeepLIFT as a path-attribution method, explains the difference in output from some reference output in terms of the difference of the input from some reference input. Assuming t represents some target output neuron of interest and let $x_1, x_2, \ldots, x_n$ represent some neurons in some intermediate layer or set of layers that are necessary and sufficient to compute t. Let t^0 represent the reference activation of t. Let $\Delta t = t - t^0$ be the difference-from-reference. DeepLIFT assigns contribution scores $C_{\Delta x_i \Delta t}$ to Δx_i such that:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t \quad (2)$$

Eq. (2) is called the summation-to-delta property. $C_{\Delta x_i \Delta t}$ can b thought of as the amount of difference-from-reference in t that is attributed on the difference-from-reference of x_i. [7]

$C_{\Delta x_i \Delta t}$ can be non-zero even when $\frac{\partial t}{\partial x_i}$ is zero, meaning that a neuron can be signalling meaningful information even if its gradient is zero, which allows DeepLIFT to address a fundamental limitation of gradient-based techniques. Moreover, the difference-from-reference is continuous, which lets DeepLIFT avoid discontinuities caused by bias terms. Like Grad-CAM, DeepLIFT can also be applied to all neural network layers. [7]

Still, DeepLIFT outputs, like other path-attribute techniques, are deeply affected by the choice of the reference image.

D. Evaluation Metrics of Interpretations

Having explanations of neural networks are not enough to establish human trust, but these explanations should be robust. To measure the fragility of interpretation maps, two objective measures were introduced: The infidelity score and the sensitivity score.

1) Infidelity Score: Given a black-box function f, explanation function ϕ, a random variable $I \in \mathbb{R}^d$ with probability measure μ_I, which represents meaningful perturbations of interest (like a gaussian noise), the infidelity of ϕ is defined as: [3]

$$\text{INFD}(\phi, f, x) = \mathbb{E}_{I \sim \mu_I}\left[(I^T \phi(f, x) - (f(x) - f(x - I)))^2\right] \quad (3)$$

2) Sensitivity Score: Given a black-box function f, explanation function ϕ, input x, and perturbation radius r, the sensitivity of ϕ is defined as: [4]

$$SENS(\phi, f, x, r) = \max_{||y-x|| \leq r} ||\phi(f, y) - \phi(f, x)|| \quad (4)$$

III. Datasets

In this study, we conducted our experiments on two medical datasets: KVASIR-Seg and HAM10000.

A. KVASIR-SEG

The KVASIR-SEG dataset is an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated and verified by an experienced gastroenterologist. This dataset contains 1000 polyp images and their corresponding ground truth mask. The resolution of the images contained in KVASIR-SEG varies from 332x487 to 1920x1072 pixels. [10] The dataset was divided into 800 images for training, 100 images for validation and 100 images for testing. An example of data is shown in Fig. 2.

B. HAM10000

The KVASIR-SEG dataset is an open-access dataset of dermatoscopic skin lesions images and corresponding segmentation masks. This dataset contains 10000 skin lesion images and their corresponding ground truth mask. The resolution of the images contained in HAM10000 is 600x450 pixels. [11], [12] The dataset was divided into 8000 images for training, 1000 images for validation and 1000 images for testing. An example of data is shown in Fig. 3.

C. Data Preprocessing

KVASIR-SEG and HAM10000 dataset images were resized to 400x400 pixels and 256x256 pixels, respectively. Then, a set of random augmentations, including flips, gaussian blur, transpose, zoom, rotations and transport with a probability of 0.5 using the Albumentation library, were applied.

IV. Methods

The project aims to interpret the U-Net model outputs in the segmentation of medical images using existing classification interpretability methods. To extend these approaches to the segmentation problem, we thought of the segmentation problem as a binary classification problem, where the goal is to classify each pixel of the input image to either the background class or the segment class. Formally, if we have N images, each having WxH pixels, the segmentation problem for these images can be considered as the classification of NxWxH pixels into two classes: Background and Segment.
An object's shape in an image only depends on its border pixels and not its inner pixels. We define the border of a segment as the set of pixels, where at least one of their four neighbour pixels is classified as a background pixel.
In the interpretation of segmentation models, we are looking for the model's decision criteria for finding the segment. As we discussed, the segment can be defined uniquely by its border pixel. Therefore, we should only explore the model's decision criteria for indicating this set of pixels.
In this study, we applied Guided Backpropagation, Guided Grad-CAM, and DeepLIFT to obtain interpretation maps of the model. We also used Grad-CAM and DeepLIFT on each layer of the model to get the interpretation maps of each layer. First, the model is trained on the training data and using the validation data, the model with the highest IoU score is saved. The test data is given to the model, and the outputs are obtained. For each test output, the set of border pixels is computed. Then, the interpretation map for each border pixel is calculated. Each pixel has a score in these interpretation maps. The interpretation map produced by the mean of interpretation maps of border pixels is considered the total interpretation map of the test output.

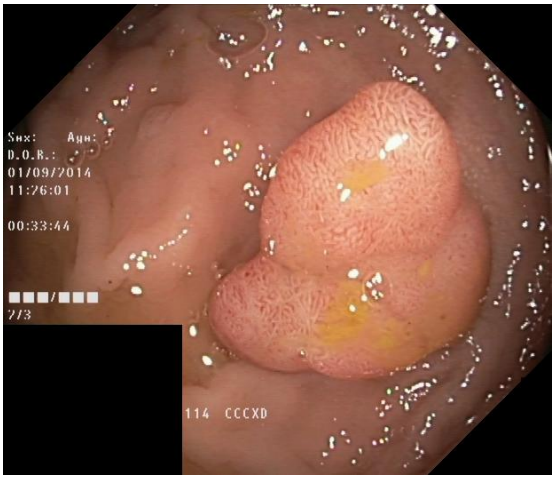

Fig. 2. An Example of KVASIR-SEG Dataset [10]

Fig. 3. An Example of HAM10000 Dataseet [11], [12]

A. Training

We used the pre-trained U-Net model provided by [9] in this study. This model, which is shown in Fig. 4 is slightly different from the original U-Net architecture. In our model, the number of channels in convolutional layers is half of the ones in the original paper and also a sigmoid layer was added as the last layer of the neural network. To implement training and test processes, we used the PyTorch library, and to perform interpretation techniques and its evaluation metrics, we employ the Captum library [8]. We ran our experiments on the Kaggle website using its NVIDIA Telsa P100 GPU with 16-gigabyte memory and 13 gigabytes of RAM. We used the Adam optimizer with a learning rate of 0.0005 for both datasets.

B. Loss Function

To train the model, we used a linear combination of binary cross-entropy loss, IoU loss, and dice loss. Formally, our loss function is defined as follows:

$$\begin{aligned}\mathcal{L}(T^k, P^k) &= \mathcal{L}_{BCE}(T^k, P^k) + \mathcal{L}_{IoU}(T^k, P^k) + \mathcal{L}_{Dice}(T^k, P^k) \\ &= \mathcal{L}_{BCE}(T^k, P^k) + (1 - IoU(T^k, P^k)) + (1 - Dice(T^k, P^k)) \\ &= 2 + \mathcal{L}_{BCE}(T^k, P^k) - IoU(T^k, P^k) - Dice(T^k, P^K))\end{aligned} \quad (5)$$

Where T^k and P^k are the true mask and predicted mask of kth image, respectively. The total loss is the mean of loss for all images. This loss function optimizes for precise segmentations and promotes the learning of shape. We also tried different weights for each term of the loss function; however, the best result was attained when they had equal weights.

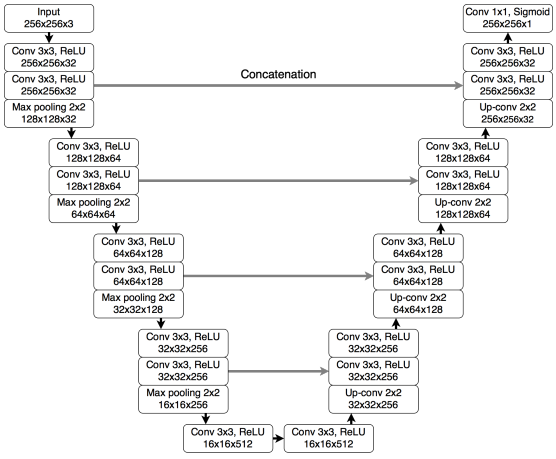

Fig. 4. Pre-trained U-Net Model Used in the Experiments [9]

V. Results

We used MIoU and MDice to evaluate our model and the infidelity score to evaluate the fragility of interpretation methods.

A. KVASIR-SEG

For KVASIR-SEG, we trained our model for 500 epochs with a batch size of 8. No hyper-parameter tuning was done, and none of the layers was frozen. Table I shows the MIoU and MDice of our model compared to other models.
The interpretation maps of the model and layers for this dataset are shown in Figs. 5, 6, and 7. From Fig. 5, we can see that border pixels have positive effects and inner pixels of polyp have negative effects in the model's decision. The reason for this could be the similarity of the colon's and polyp's surfaces. We also can find out that the model is learning to find protuberances in the images. Moreover, we observe that DeepLIFT shows more pixels with an effect on the output than the other methods.
Figs. 6 and 7 show that as the encoder becomes deeper, the model focuses more on the border pixels and they get higher scores. Fig. 6 also shows that in the bottleneck layer, the pixels around the segment have a high negative influence on the model's learning. In the decoder layers, segment pixels have positive effects and surrounding pixels have negative effects on the learning process. These positive and negative influences are increasing until the decoder layer Decoder2 and after that are decreasing, meaning this layer has the highest effect on the model's learning.
Table II and III show the fragility of the interpretation methods, where we can observe that Guided Grad-CAM and Grad-CAM are the most robust interpretation methods on this dataset and their results are more reliable.

TABLE I
COMPARISON OF MIOU AND MDICE OF OUR MODEL WITH OTHER MODELS ON KVASIR-SEG DATASET

Model	MDice	MIoU
Standard U-Net	0.8180	-
ColonSegNet	0.8206	0.7239
FANet	0.8803	-
MSRF-Net	0.9217	0.8914
Our Model	0.08412	0.7598

TABLE II
FRAGILITY OF INTERPRETATION METHODS FOR THE MODEL ON KVASIR-SEG DATASET

Method	Infidelity Score
Guided Backpropagation	0.2415
Guided Grad-CAM	0.0265
DeepLIFT	0.0907

TABLE III
FRAGILITY OF INTERPRETATION METHODS FOR LAYERS ON KVASIR-SEG DATASET

Method	Infidelity Score
Grad-CAM	0.3606
DeepLIFT	2.4576

B. HAM10000

For HAM10000, we trained our model for 100 epochs with a batch size of 64. No hyper-parameter tuning was done, and none of the layers was frozen. Table IV shows the MIoU and MDice of our model compared to other models.

The interpretation maps of the model and layers for this dataset are shown in Figs. 8, 9, and 10. From Fig. 8, we can see that only border pixels contribute to the model's decision, meaning the model decides based on the difference in colour between the skin and the lesion.

Figs. 8 and 9 show that like the KVASIR-SEG dataset as the deeper the encoder becomes, the model focuses more on the border pixels and they get higher scores.

Table V and VI show the fragility of the interpretation methods, where we can see like the KVASIR-SEG dataset, Guided Grad-CAM and Grad-CAM are the most robust interpretation methods on this dataset and their results are more reliable.

TABLE IV
COMPARISON OF MIOU AND MDICE OF OUR MODEL WITH OTHER MODELS ON HAM10000 DATASET

Model	MDice	MIoU
Polar Res-U-Net++	0.9253	-
DoubleU-Net	0.8962	-
BAT	0.912	0.843
MSRF-Net	0.8813	-
Our Model	0.9408	0.8976

TABLE V
FRAGILITY OF INTERPRETATION METHODS FOR THE MODEL ON HAM10000 DATASET

Method	Infidelity Score
Guided Backpropagation	0.0121
Guided Grad-CAM	0.0068
DeepLIFT	0.0191

TABLE VI
FRAGILITY OF INTERPRETATION METHODS FOR LAYERS ON HAM10000 DATASET

Method	Infidelity Score
Grad-CAM	0.2507
DeepLIFT	0.2704

VI. CONCLUSIONS

In this work, we present a way to extend classification interpretation methods to the segmentation models. We showed that the U-Net's decision criteria are clinically valid and robust for two datasets. Furthermore, we demonstrated although DeepLIFT gives us more detailed interpretation maps, Guided Grad-CAM and Grad-CAM outputs are much more robust. In future studies, experiments in this project can be extended to other segmentation models and their interpretability and fragility of their interpretations can be explored.

REFERENCES

[1] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., christophm.github.io/interpretable-ml-book/, 2022.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U- net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234-241.

[3] C. K. Yeh, C. Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in NeurIPS, 2019.

[4] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in AAAI, vol. 33, 2019, pp. 3681-3688.

[5] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in CoRR, abs/1412.6806, 2015.

[6] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Visual explanation from deep networks via gradient-based localization.," in International Journal of Computer Vision, vol. 128, 2019, pp. 336-359.

[7] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagation activation differences," in ICML, 2017.

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

[9] P. Yakubovskiy, "Segmentation models pytorch," github.com/qubvel/segmentation_models.pytorch, 2020.

[10] D. Jha, P. H. Smedsrud, M. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in ArXiv, abs/1911.07.069, 2020.

[11] N. C. F. Codella, V. M. Rotemberg, P. Tschandl, M. E. Celebi, S. W. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. A. Marchetti, H. Kittler, and A. C. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," in ArXiv, abs/1902.03368, 2019.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," in Scientific Data, vol. 5, 2018.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in CoRR, abs/1312.6034, 2014.

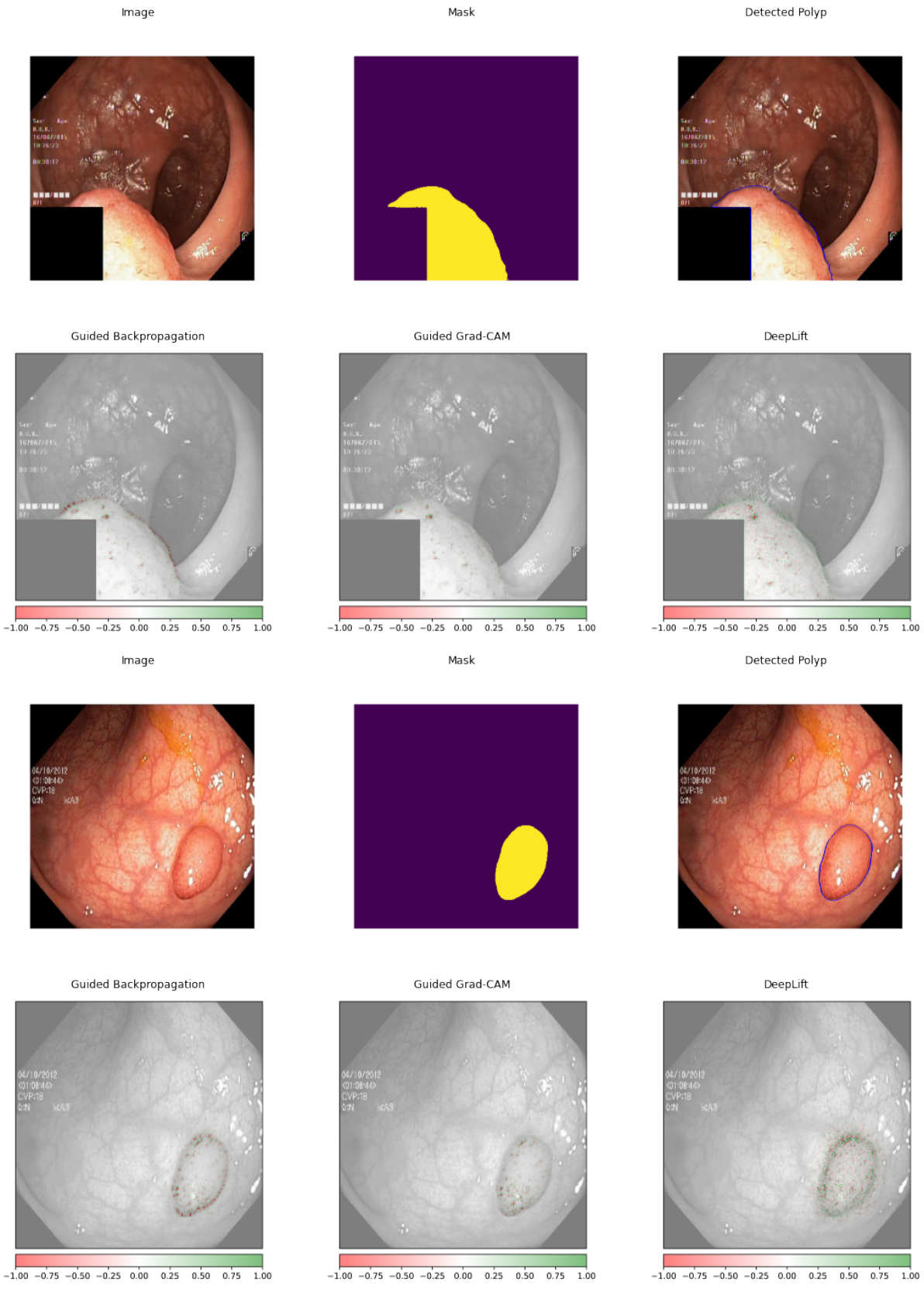

Fig. 5. Interpretation Maps of Model for KVASIR-SEG Test Data

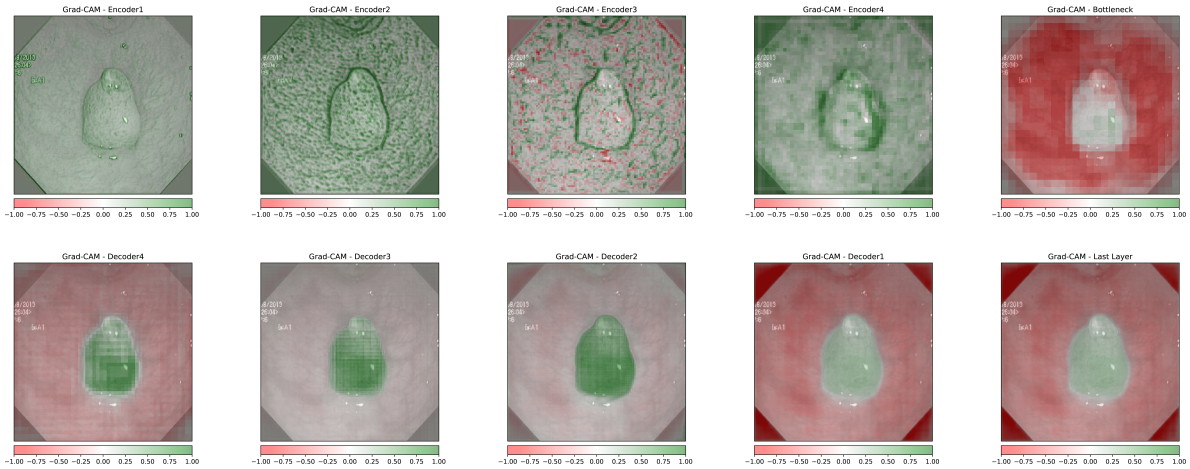

Fig. 6. Guided Grad-CAM Interpretation Maps of Layers for KVASIR-SEG Test Data

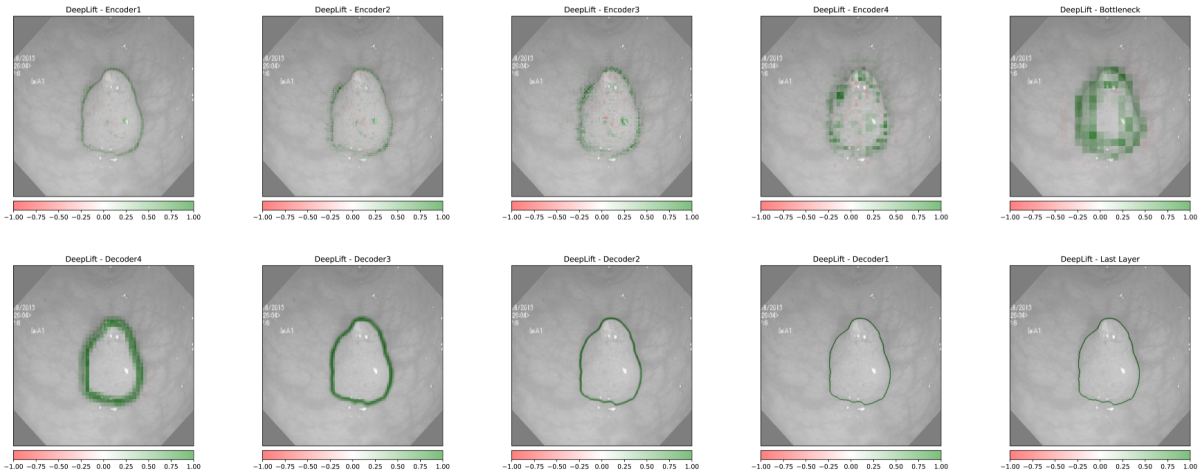

Fig. 7. DeepLIFT Interpretation Maps of Layers for KVASIR-SEG Test Data

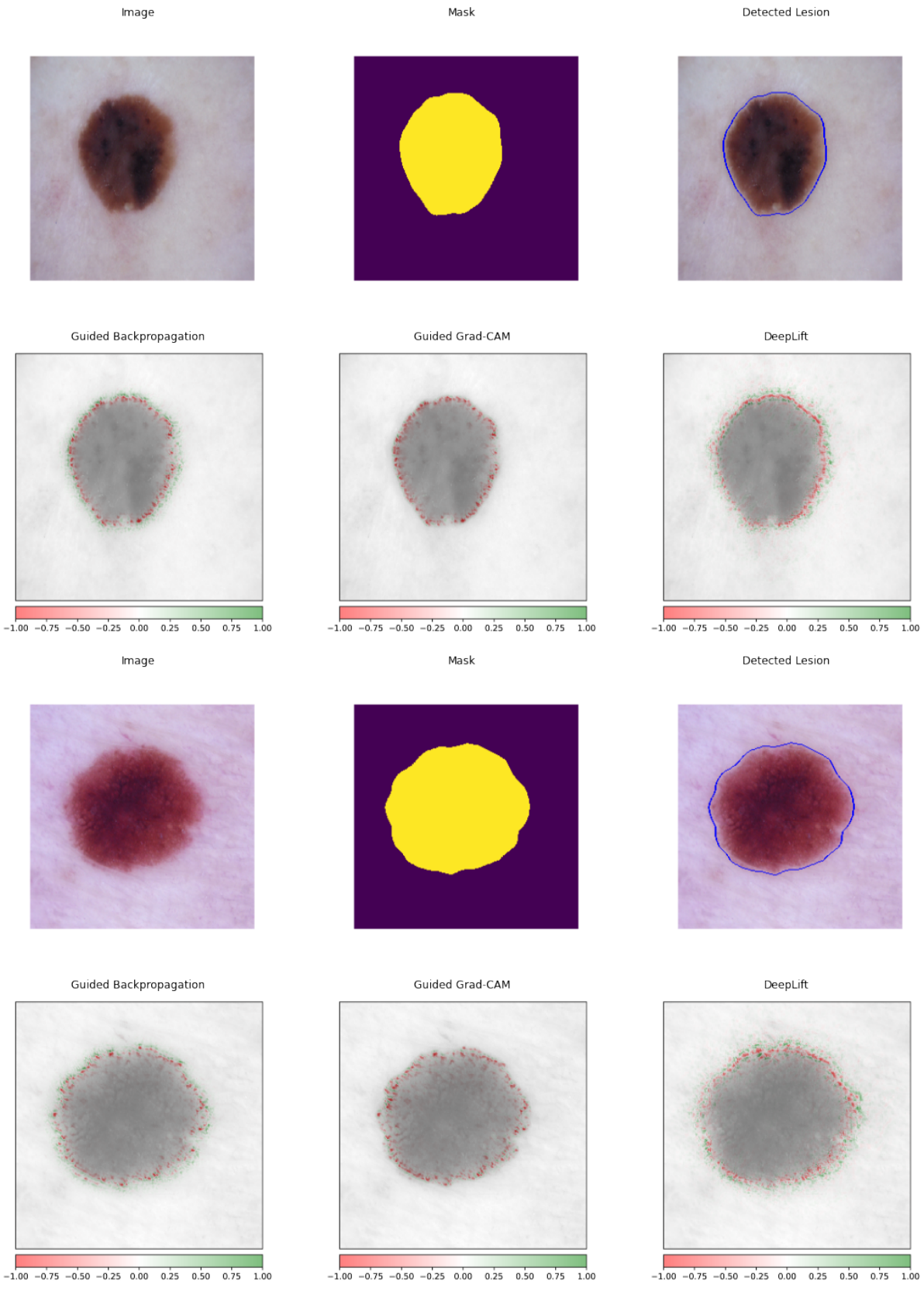

Fig. 8. Interpretation Maps of Model for HAM10000 Test Data

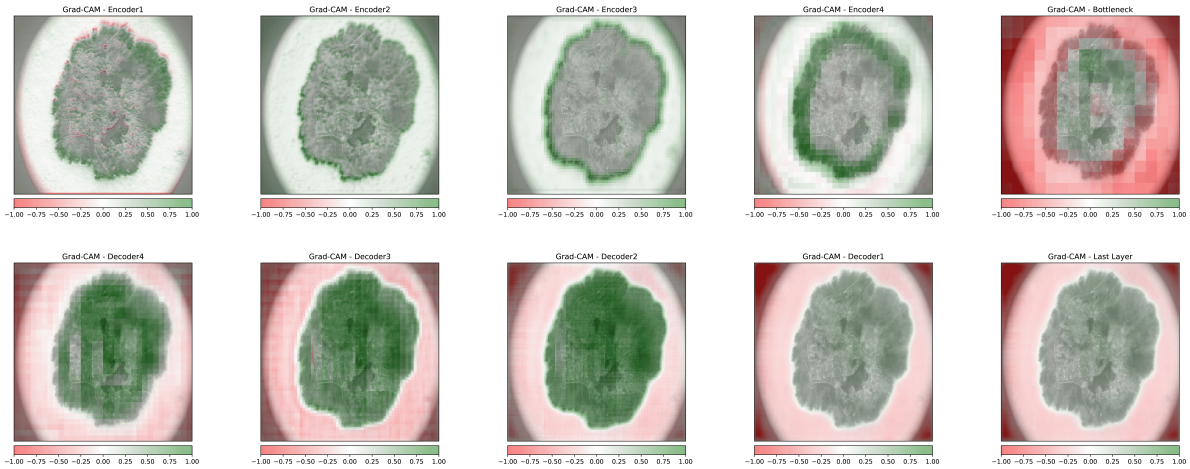

Fig. 9. Guided Grad-CAM Interpretation Maps of Layers for HAM10000 Test Data

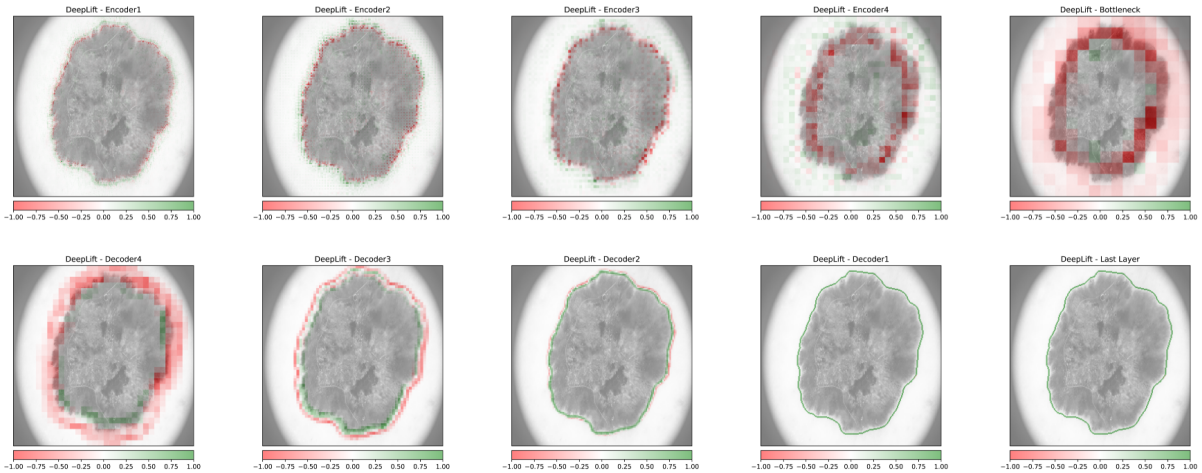

Fig. 10. DeepLIFT Interpretation Maps of Layers for HAM10000 Test Data