

# SegX: Improving Interpretability of Clinical Image Diagnosis with Segmentation-based Enhancement

Yuhao Zhang<sup>1, \*</sup>, Mingcheng Zhu<sup>2, \*</sup>, Zhiyao Luo<sup>2, †</sup>

<sup>1</sup>National University of Singapore, Singapore

<sup>2</sup>Department of Engineering Science, University of Oxford, UK

yuhao-zhang@u.nus.edu, mingcheng.zhu@eng.ox.ac.uk, zhiyao.luo@eng.ox.ac.uk

## Abstract

Deep learning-based medical image analysis faces a significant barrier due to the lack of interpretability. Conventional explainable AI (XAI) techniques, such as Grad-CAM and SHAP, often highlight regions outside clinical interests. To address this issue, we propose Segmentation-based Explanation (SegX), a plug-and-play approach that enhances interpretability by aligning the model’s explanation map with clinically relevant areas leveraging the power of segmentation models. Furthermore, we introduce Segmentation-based Uncertainty Assessment (SegU), a method to quantify the uncertainty of the prediction model by measuring the ‘distance’ between interpretation maps and clinically significant regions. Our experiments on dermoscopic and chest X-ray datasets show that SegX improves interpretability consistently across mortalities, and the certainty score provided by SegU reliably reflects the correctness of the model’s predictions. Our approach offers a model-agnostic enhancement to medical image diagnosis towards reliable and interpretable AI in clinical decision-making. The implementation is publicly available at <https://github.com/JasonZuu/SegX>.

## Introduction

The integration of artificial intelligence (AI) into medical image analysis has become a central area of research due to its potential to streamline workflows and enhance diagnostic accuracy (Pinto-Coelho 2023). Among various AI techniques, deep learning (DL) have shown remarkable efficacy across diverse medical applications, including disease diagnosis and lesion segmentation (Barata et al. 2023; Gu et al. 2018; Ramesh et al. 2021). For instance, on the HAM10000 dataset (Tschandl, Rosendahl, and Kittler 2018), a diagnostic model has achieved a high sensitivity of 87.1% for basal cell carcinoma detection using reinforcement learning (Barata et al. 2023), outperforming the clinicians. Similarly, a model that learns the common patterns from multi-source data achieved an accuracy of 85.8% across multiple diseases on the ChestX-Det10 dataset (Liu, Lian, and Yu 2020), demonstrating DL’s potential in chest X-ray diagnosis (Dai et al. 2024). Despite these advancements, the black-box nature of these models and their opaque decision-making processes remain critical barriers to their adoption in broader clinical environments, where interpretability and trustworthiness are paramount (Abbasian Ardakani et al. 2024).

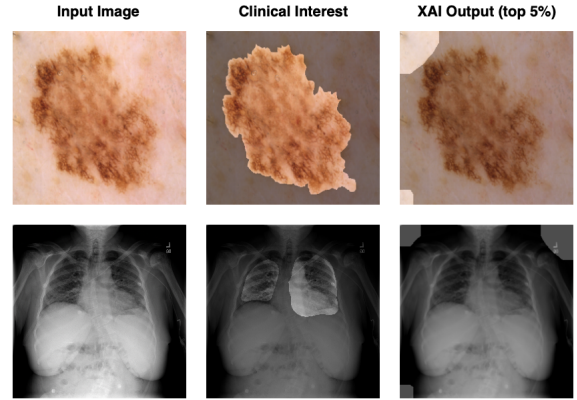


Figure 1: **Negative examples of using XAI to highlight highly important regions.** The first column shows the original images from HAM10000 and ChestX-Det10, respectively. The second column shows the clinical interests of each image. The last column shows the Grad-CAM interpreted results on a learned model. It is clear to see that the model overemphasizes background area for making a diagnosis, which mismatches clinical interests.

Explainable AI (XAI) methods have demonstrated significant value in enhancing model interpretability across various computer vision tasks by providing visual evidence for predictions, which is often intuitive for human users (Hassija et al. 2024). For instance, Grad-CAM (Selvaraju et al. 2017) highlights regions in an image that contribute most to the model’s prediction by leveraging the gradients of target class scores with respect to feature map activations. In medical image classification, methods like Grad-CAM (Selvaraju et al. 2017) and SHAP (Lundberg and Lee 2017) have been widely adopted to enhance the trustworthiness of AI models (Loh et al. 2022; Borys et al. 2023). However, as shown in Fig.1, these methods fail to consistently align their explanation regions with clinically relevant areas (Yuan et al. 2024), which limits their applicability in clinical practice.

Knowledge-guided XAI (KG-XAI) methods have been developed to enhance the clinical relevance of model explanations by integrating clinical knowledge into the interpretability process (Yuan et al. 2024; Jung et al. 2023; Yuan

et al. 2023). These methods extract clinically significant regions from input images using techniques such as image transformations (Yuan et al. 2024) or deep learning-based approaches (Jung et al. 2023; Yuan et al. 2023), which serve as clinical knowledge. The extracted clinical knowledge is then used to align XAI outputs with these regions, resulting in explanation maps that better reflect clinical interest. However, KG-XAI methods overlook regions outside the identified clinical areas, such as background features, which can also carry meaningful information. For example, the background may contain contextual cues or subtle signs of disease progression (Cozzi et al. 2021). masking background regions outside the clinical areas can lead to information loss and potentially mislead clinicians by providing an incomplete understanding of the model’s predictions.

To address the limitations of existing KG-XAI, this paper introduces Segmentation-based explanation (SegX) and Segmentation-based Uncertainty Assessment (SegU), two novel approaches that integrate clinical knowledge of disease-specific focus regions to enhance the reliability and interpretability of existing XAI techniques, as shown in Fig 2. Specifically, SegX applies the segmentation masks of clinically significant regions on the XAI output to generate refined explanation maps. SegU calculates the ‘distance’ between the segmentation mask and the original explanation map, therefore measuring uncertainty.

**Designed as flexible, plug-and-play modules, SegX and SegU can be seamlessly integrated into any XAI method for medical image classification models to incorporate clinical knowledge.** Experiments on dermoscopic and X-ray datasets demonstrate that SegX consistently produces refined explanations that closely match clinician annotations. In addition, we found that SegU can distinguish between the correct and incorrect predictions of the classification model. Overall, The enhanced interpretability and uncertainty assessment provide clinicians with a better understanding of AI predictions and improve confidence levels, enabling more informed and cautious clinical decisions (Zhang et al. 2020), thereby facilitating the adoption of AI in clinical practice.

## Related Work

### Explainable AI (XAI)

XAI approaches for image analysis models are designed to enhance transparency by highlighting regions within the input data that contribute significantly to the model’s decision-making process. Among these approaches, class activation maps (CAM) (Zhou et al. 2016) utilize global average pooling (GAP) within CNNs to identify regions within the input associated with specific classes, enabling the model to highlight the areas most relevant for classification. Grad-CAM (Selvaraju et al. 2017) builds upon CAM by backpropagating gradients to specified layers, generating a heatmap that visualizes the most influential regions in the input image; this approach is widely used in tasks such as disease diagnosis for its ability to focus on high-importance areas without altering the network architecture. SHAP (Lundberg and Lee 2017), a feature attribution method based on Shapley val-

ues, assigns importance scores to individual features, capturing the contribution of each pixel or region in complex, non-linear networks; this technique provides a more comprehensive view of feature influence across the network but requires substantial computational resources. While CAM and Grad-CAM primarily highlight relevant regions, SHAP offers pixel-level feature attributions.

Although CNN-based XAI methods such as CAM, Grad-CAM, and SHAP have shown effectiveness in conventional computer vision tasks, their explanations often lack alignment with clinically meaningful regions in medical images (Yuan et al. 2024). These methods may highlight irrelevant background regions due to learning and/or data biases rather than true clinically relevant regions, limiting their utility in practice.

### Knowledge-guided XAI

Recent research focuses on aligning XAI outputs with clinical interests to enhance their relevance for clinical decision-making. These methods can be broadly categorized based on how they model clinical knowledge. One type involves extracting clinically significant regions through image transformation. For instance, Yuan et al. (Yuan et al. 2024) proposed a clinical template for pneumothorax by employing image transformation techniques such as flipping and dilation. This template is then applied post hoc to refine XAI outputs. While this template-based method improves the alignment effectively for the pneumothorax in chest X-ray images, its static design may limit its generalisation across different diseases and imaging modalities.

The second type of clinical knowledge modelling leverages deep learning (DL) methods to extract clinically significant regions. For instance, Jung et al. (Jung et al. 2023) introduced a spatial attention mechanism that identifies potential disease areas through learning, with the attention map then used to refine CAM. However, this method tightly couples the attention mechanism with the prediction process, resulting in a high dependency between clinical knowledge modelling and the model’s predictions. Consequently, it fails to provide precise explanations for incorrect predictions, limiting its interpretability. In contrast, Yuan et al. (Yuan et al. 2023) proposed a segmentation-based approach that decouples knowledge modelling from disease diagnosis. Their method employs segmentation models to create clinical masks directly from original images, refining XAI outputs independently of the prediction process. However, this approach is task-specific, as the segmentation model is designed for pneumothorax and lacks the flexibility to generalize to multi-disease scenarios. Additionally, by focusing on a narrow region only for pneumothorax, the method may overlook other relevant areas, potentially leading to significant information loss and limiting the comprehensiveness of the explanation.

While these methods improve alignment between XAI outputs and clinically significant areas, they have notable drawbacks. By focusing solely on clinically significant regions, they often mask out surrounding areas, including background regions, which can still carry important contextual information. Background features may provide subtle

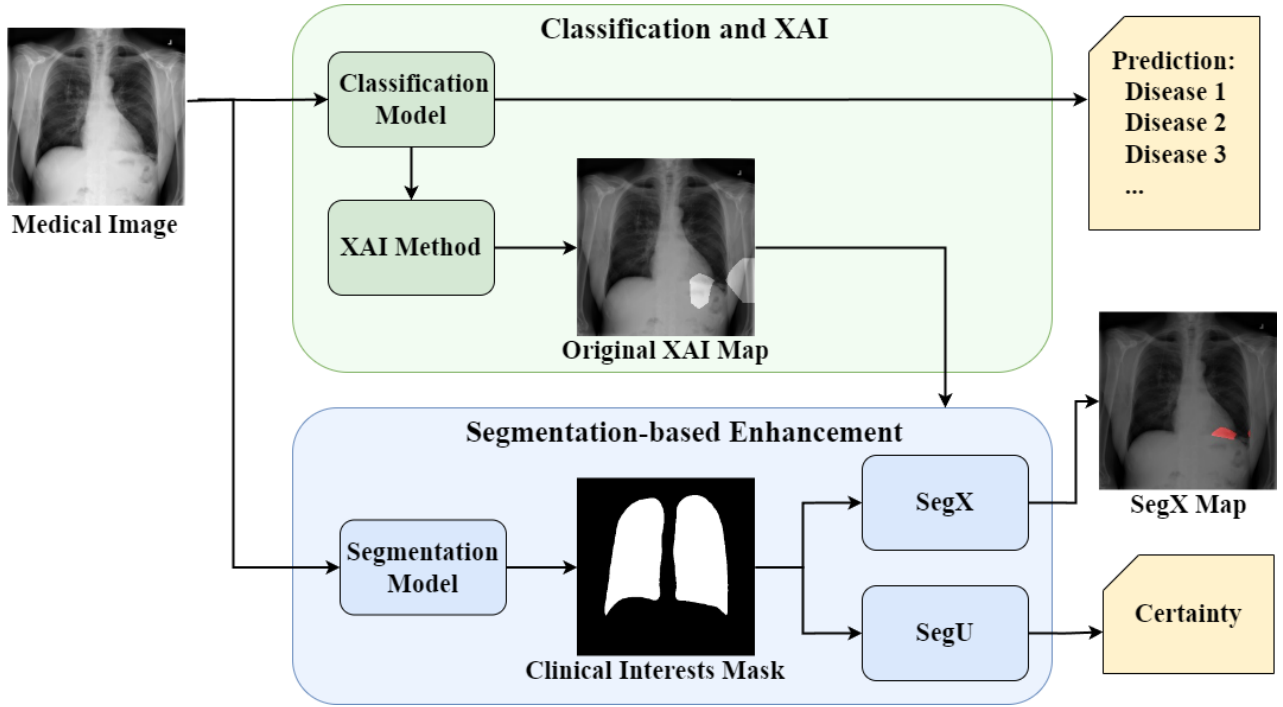


Figure 2: **Overview of the proposed method for improving XAI outputs with segmentation-based improvement.** The process begins with the classification and XAI pipeline (green), where a medical image is processed by a classification model to generate predictions and an XAI method produces the original explanation map. Afterwards, the segmentation-based enhancement (blue) starts with a segmentation model that generates a clinical interests mask to align the XAI outputs with clinically significant regions using the SegX module, producing a refined SegX map. Additionally, the SegU module quantifies the certainty of the prediction by evaluating the alignment between the explanation map and the clinical interests mask. Together, SegX and SegU enhance the interpretability and reliability of the original process, supporting better decision-making in clinical applications.

cues or additional context for understanding the model’s predictions. Masking these regions entirely can lead to information loss and may mislead clinicians by providing an incomplete view of the model’s decision-making process, potentially causing them to overestimate the model’s reliability.

## Methodology

### Problem Formulation

**Disease Diagnosis from Medical Images.** Given an input image  $\mathbf{x}_i \in \mathbb{R}^{W \times H \times C}$  with width  $W$ , height  $H$ , and channel  $C$ , the classification model  $f_{\text{cls}}$  generates a probability distribution over multiple labels, represented as

$$\mathbf{p}_i = f_{\text{cls}}(\mathbf{x}_i), \quad (1)$$

where  $\mathbf{p}_i = \{p_{ij} \mid j = 1, \dots, N\}$  is the set of probability scores for each of  $N$  classes. Each element  $p_{ij}$  represents the probability of class  $j$  for the input image  $\mathbf{x}_i$ .

The multi-label prediction  $\hat{\mathbf{y}}_i$  is determined by a class-wise threshold  $\tau_j$ . Specifically,

$$\hat{\mathbf{y}}_i = \{j \mid p_{ij} > \tau_j\}. \quad (2)$$

The objective of the classification model is to minimize the overall prediction error across the dataset by optimizing

the model parameters  $\theta$ . Specifically, the objective function  $\mathcal{O}_{\text{cls}}$  is defined as:

$$\mathcal{O}_{\text{cls}} = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N \mathcal{E}(\hat{\mathbf{y}}_i, \mathbf{y}_i), \quad (3)$$

where  $N$  is the number of samples,  $\mathbf{y}_i$  is the corresponding ground truth label set, and  $\mathcal{E}$  represents a general error metric that quantifies the difference between predictions and ground truth.

**Explainable AI (XAI).** XAI method  $f_{\text{xai}}$  is applied to the classification model to produce the interpretability mask  $\mathbf{m}_{i,j} \in \mathbb{R}^{H \times W}$  for each label  $j \in \hat{\mathbf{y}}_i$ .

$$\mathbf{m}_{i,j}^{\text{xai}} = f_{\text{xai}}(f_{\text{cls}}, X_i, j), \quad (4)$$

$\mathbf{m}_{i,j}$  highlights the regions in  $x_i$  that makes the model  $f_{\text{cls}}$  to output  $p_{i,j}$ .

**Segmentation-guided XAI.** A segmentation model  $f_{\text{seg}}$  takes the input image  $\mathbf{x}_i$  to generate a segmentation mask  $\mathbf{m}_i^{\text{seg}} \in \mathbb{R}^{H \times W}$  of the clinically relevant regions.

$$\mathbf{m}_i^{\text{seg}} = f_{\text{seg}}(X_i), \quad (5)$$

For each label  $j \in \hat{\mathbf{y}}_i$ , a segmentation-based interpretability mask is generated by overlaying the interpretability mask  $\mathbf{m}_{i,j}^{\text{xai}}$  with the segmentation mask  $\mathbf{m}_i^{\text{seg}}(\mathbf{x}_i)$ :

$$\mathbf{m}_{i,j}^{\text{segX}} = \mathbf{m}_{i,j}^{\text{xai}} \cap \mathbf{m}_i^{\text{seg}}. \quad (6)$$

## Disease Diagnosis and XAI

This study implemented CNN models (He et al. 2016; Huang et al. 2017) for disease diagnosis from different types of medical images, including dermoscopic images (Tschandl, Rosendahl, and Kittler 2018) and X-ray images (Liu, Lian, and Yu 2020). For an input image  $\mathbf{x}$ , the CNN-based model scans and analyzes the image to extract disease-related features. These features are then processed to generate a probability distribution over potential diseases.

To interpret the model’s prediction, we employ XAI approaches (Selvaraju et al. 2017; Lundberg and Lee 2017) to identify regions contributing to predictions. Specifically, the XAI method calculates the influence of each feature map on the model’s output by weighting feature activations according to their gradient values. This weighted sum of feature maps is processed through an activation function  $f_{\text{act}}(\cdot)$ , producing interpretability output. Following previous studies (Yuan et al. 2023, 2024), we extract the top 5% saliency map  $\mathbf{m}^{\text{xai}}$  as the explanation map from the XAI output. This map reveals spatial and feature-specific insights, offering a clearer understanding of the model’s decision-making process.

## Segmentation-based Explanation (SegX)

The SegX approach enhances interpretability by aligning the XAI’s output for the classification model with clinically relevant regions yielded by a segmentation model. By constraining the XAI output to the critical clinical regions, this approach provides a more detailed and reliable explanation of the CNN model’s decision-making process, making the interpretability output more meaningful and useful for clinical applications.

Specifically, given an input image  $\mathbf{x}_i$ , the segmentation model generates a mask  $\mathbf{m}_{i,j}^{\text{seg}}$ , where each pixel  $\mathbf{m}_{i,j}^{\text{seg}} \in \{0, 1\}$ . The segmentation mask  $\mathbf{m}_{i,j}^{\text{seg}}$  indicates the region of clinical interest for the specific disease  $j$ . To enhance the interpretability of XAI, we overlaid the segmentation mask  $\mathbf{m}_{i,j}^{\text{seg}}$  with the original XAI’s saliency map  $\mathbf{m}_{i,j}^{\text{xai}}$  for the classification model to create the segmentation-guided XAI mask  $\mathbf{m}_{i,j}^{\text{segX}}$ :

$$\mathbf{m}_{i,j}^{\text{segX}} = \mathbf{m}_{i,j}^{\text{seg}} \cdot \mathbf{m}_{i,j}^{\text{xai}}. \quad (7)$$

As the  $\mathbf{m}_{i,j}^{\text{xai}}$  provides the classification model’s focus and  $\mathbf{m}_{i,j}^{\text{seg}}$  indicates the clinical regions, the overlap mask  $\mathbf{m}_{i,j}^{\text{segX}}$  is able to highlight regions where the classification model’s focus aligns with clinically critical regions.

Since the segmentation-guided XAI masks rely on the output of the segmentation model, it is essential for the segmentation model to accurately capture clinically relevant regions. To optimize segmentation performance and ensure

alignment with true lesion areas, we employ a composite loss function:

$$\mathcal{L}_{\text{seg}} = \lambda \mathcal{L}_{\text{CE}} + (1 - \lambda) \mathcal{L}_{\text{Dice}}, \quad (8)$$

where  $\lambda$  controls the balance between cross-entropy loss (Zhu, Zhang, and Wang 2022) and Dice loss (Kato and Hotta 2024). By combining these two losses, we ensure that the model achieves both pixel-level precision and accurate clinical region localization, which are critical for reliable XAI explanations (Yeung et al. 2022).

## Segmentation-based Uncertainty Assessment (SegU)

To address the potential issue that knowledge-guided XAI may inadvertently obscure the model’s focus on clinically irrelevant regions, we propose a segmentation-based uncertainty assessment to alert clinicians to the model’s uncertain decisions. We hypothesized that the correction predictions for biomedical images are more likely to be made from the clinically significant regions. Thus, we use the Intersection over Union (IoU) between the classification model’s focus mask  $M_{\text{cls}}(X_i)$  and the segmentation mask  $M_{\text{seg}}(X_i)$  as an uncertainty metric. The segmentation mask, when trained appropriately, is hypothesized to serve as a clinically meaningful reference. A low IoU score would indicate insufficient alignment between the classification model’s attention and clinically important regions, suggesting that the interpretability output may be unreliable.

We propose two novel certainty scores to assess the model’s certainty for its prediction based on the alignment between the XAI maps and clinically significant segmentation masks. The first is the **IoU-based certainty score** ( $c_{\text{IoU}}$ ), which quantifies the overlap between the XAI map  $\mathbf{m}_{i,j}^{\text{xai}}$  and the corresponding segmentation mask  $\mathbf{m}_{i,j}^{\text{seg}}$ . It is calculated as follows:

$$c_{\text{IoU}} = \frac{|\mathbf{m}_{i,j}^{\text{xai}} \cap \mathbf{m}_{i,j}^{\text{seg}}|}{|\mathbf{m}_{i,j}^{\text{xai}} \cup \mathbf{m}_{i,j}^{\text{seg}}|}, \quad (9)$$

The second is based on the **Area Under the IoU-Threshold Curve (AUITC)**, which provides a comprehensive measure of alignment by evaluating the IoU scores across varying thresholds used to generate the explanation map. The AUITC-based certainty score  $c_{\text{AUITC}}$  reflects the overall consistency between the XAI outputs and the segmentation mask across different saliency levels. Formally, it is defined as:

$$c_{\text{AUITC}} = \int_0^1 c_{\text{IoU}}(\tau) d\tau, \quad (10)$$

where  $\tau$  denotes the threshold applied to the XAI output to generate the explanation map  $\mathbf{m}^{\text{xai}}$ , and  $c_{\text{IoU}}(\tau)$  is the IoU score calculated at a given threshold  $\tau$ .

## Experiment

### Datasets

To evaluate the effectiveness of our proposed SegX and SegU across various types of medical images, we as-

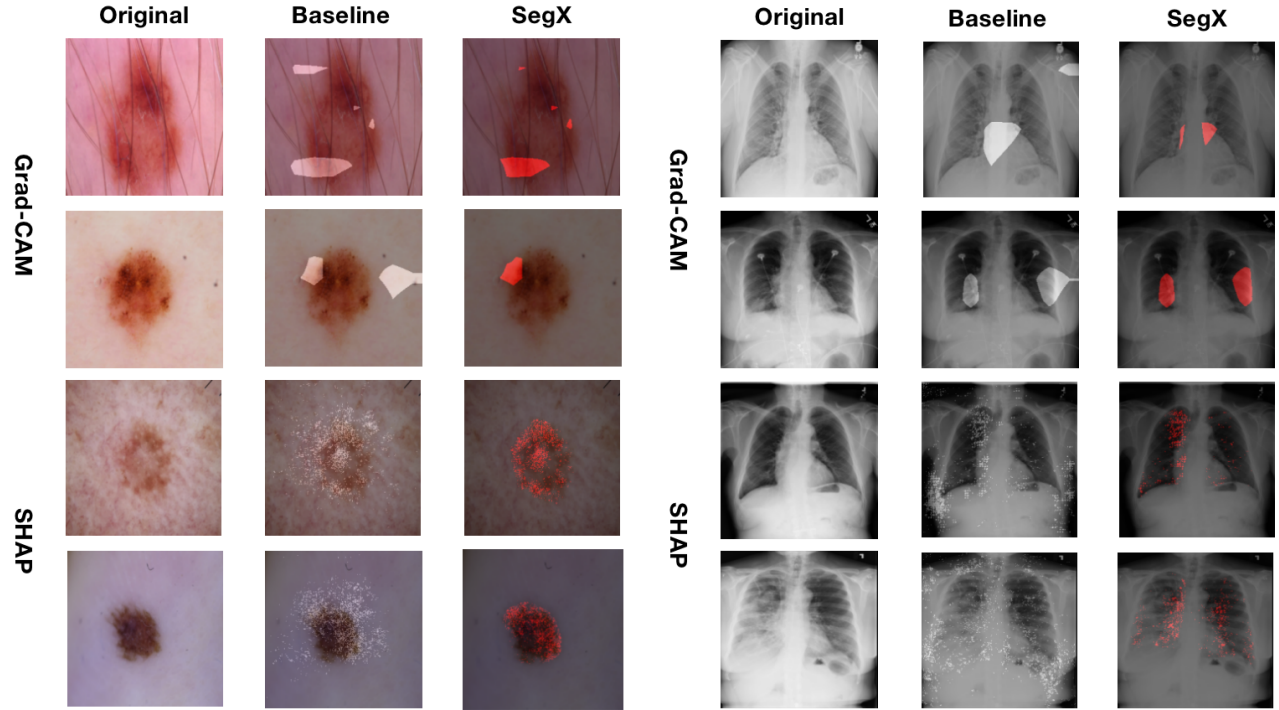


Figure 3: Examples of original XAI masks (baseline) versus SegX masks on the HAM10000 and ChestX-Det10 test sets. The model’s top 5% focus region is highlighted in white for the baseline and in red for the SegX method.

essed its performance using two real-world biomedical image datasets: HAM10000 (Tschandl, Rosendahl, and Kittler 2018), a dermoscopic image dataset; and ChestX-Det10 (Liu, Lian, and Yu 2020), a chest X-ray dataset.

The HAM10000 (Tschandl, Rosendahl, and Kittler 2018) dataset is a publicly available collection of 10,015 dermoscopic images for skin lesion classification. Collected over 20 years from two sites—the Department of Dermatology at the Medical University of Vienna, Austria, and the skin cancer practice of Cliff Rosendahl in Queensland, Australia—this dataset includes images of 7 distinct types of skin lesions, representing both benign and malignant cases. The statistics of the HAM10000 dataset are shown in Table 1.

Table 1: Statistics for the HAM10000 Dataset

Lesion Type	Count
Actinic Keratoses and Intraepithelial Carcinoma (akiec)	327
Basal Cell Carcinoma (bcc)	514
Benign Keratosis-like Lesions (bkl)	1,099
Dermatofibroma (df)	115
Melanoma (mel)	1,113
Melanocytic Nevi (nv)	6,705
Vascular Lesions (vasc)	142
Total Images	10,015

ChestX-Det10 (Liu, Lian, and Yu 2020) is an X-ray image

dataset designed for chest disease classification. It consists of 3,583 chest X-ray images annotated with labels for 10 common chest diseases or abnormalities. As each image can contain multiple co-occurring diseases, the dataset is structured for multi-label classification. The detailed statistics of ChestX-Det10 are shown in Table 2.

Table 2: Statistics for the ChestX-Det10 Dataset

Disease	Count
Atelectasis	340
Calcification	347
Consolidation	2,537
Effusion	2,092
Emphysema	298
Fibrosis	738
Fracture	661
Mass	160
Nodule	955
Pneumothorax	211
Total Image	3,583

## Experiment Settings

We randomly divided the HAM10000 and ChestX-Det10 datasets into training, validation, and test sets in the ratio of 70:10:20, respectively.

In this experiment, we implement a range of models and methods to evaluate the effectiveness and the generalisation

of our proposed method:

- **Classification models:** two CNN architectures were used for the classification task, including ResNet (He et al. 2016) and DenseNet (Huang et al. 2017), which have been proven excellent in image classification tasks and are widely used on biomedical images (Mall et al. 2023).
- **XAI methods:** To provide interpretability of the classification models, we applied two common interpretability methods to the classification model, including SHAP (Lundberg and Lee 2017) and Grad-CAM (Selvaraju et al. 2017).
- **Segmentation model:** a UNet (Ronneberger, Fischer, and Brox 2015) is used to segment clinically significant regions.

### Results of SegX’s Effectiveness Evaluation

To evaluate the effectiveness of SegX in improving the alignment between the explanation map and the clinically significant regions, we use two metrics: the IoU and the AUITC. The IoU metric measures the overlap between two masks, defined as:

$$\text{IoU}(\mathbf{m}_1, \mathbf{m}_2) = \frac{|\mathbf{m}_1 \cap \mathbf{m}_2|}{|\mathbf{m}_1 \cup \mathbf{m}_2|}, \quad (11)$$

where  $\mathbf{m}_1$  and  $\mathbf{m}_2$  represent any two binary masks.

The value of IoU ranges from 0 to 1. The closer the value is to 1, the higher the alignment between the model’s explanation and the clinical interest, which means that the XAI method has better interpretability in revealing the rationale of its decision-making.

The AUITC metric provides a comprehensive assessment of alignment by evaluating IoU scores across varying thresholds  $\tau$  applied to XAI outputs. It is computed as:

$$\text{AUITC} = \int_0^1 \text{IoU}(\mathbf{m}^{\text{gt}}, \mathbf{m}(\tau)) d\tau, \quad (12)$$

where  $\mathbf{m}^{\text{gt}}$  is the ground truth clinically significant mask, and  $\mathbf{m}(\tau)$  is the binary mask generated by applying a threshold  $\tau$  to the XAI output.

**SegX Consistently Improves Alignment.** Table 3 compares the original XAI methods with SegX on the HAM10000 and ChestX-Det10 datasets, using ResNet and DenseNet classification models. The results show that SegX consistently outperforms the original XAI methods in terms of IoU across all datasets and model-XAI combinations. On the HAM10000 dataset, SegX achieved improvements in both IoU and AUITC for both SHAP and Grad-CAM when applied to ResNet and DenseNet. Similarly, on the ChestX-Det10 dataset, SegX demonstrated smaller but noticeable gains in both metrics, reinforcing its ability to enhance alignment between interpretability outputs and clinically significant regions.

These findings highlight the effectiveness of the segmentation guidance mechanism in improving the interpretability of the model, ensuring closer alignment between model explanations and clinically relevant regions, thereby increasing the usability of interpretability in medical applications.

**SegX Excludes Misleading Explanations.** Fig. 3 illustrates interpretability masks generated by the original XAI methods and our SegX approach on the HAM10000 and ChestX-Det10 test sets. The results demonstrate that baseline XAI methods often produce coarser masks that highlight clinically irrelevant regions. In contrast, SegX refines these outputs by focusing more accurately on critical areas of interest, such as lesions or disease-relevant regions.

This refinement, achieved by incorporating clinically significant regions through segmentation, reduces the influence of irrelevant areas in model explanations. By providing clearer and more focused interpretability maps, SegX enhances trust in the model’s decision-making process, particularly in scenarios requiring precise identification of clinically significant regions, making it a valuable tool for clinical applications.

**SegX’s Improvement is Effected by Original XAI.** Table 3 demonstrates that SegX consistently improves alignment between XAI outputs and clinically significant regions on both HAM10000 and ChestX-Det10 datasets. However, the improvements in IoU and AUITC are more pronounced on the HAM10000 dataset. Specifically, SegX achieves a 1-2% increase in IoU and 2-3% improvement in AUITC on HAM10000 across both ResNet and DenseNet models. In contrast, the gains on the ChestX-Det10 dataset are relatively smaller, with improvements in IoU and AUITC limited to less than 1% for most model-XAI combinations.

The greater effectiveness of SegX on the HAM10000 dataset may be influenced by the baseline performance of the original XAI methods. As shown in Table 3, the original XAI methods yield higher IoU and AUITC scores on the HAM10000 dataset compared to the ChestX-Det10 dataset, indicating stronger baseline performance for dermoscopic images. SegX builds upon the outputs of the original XAI methods, and its ability to enhance interpretability relies on the quality of these baseline explanations. On ChestX-Det10, where the original XAI methods perform worse, the improvement achieved by SegX is inherently limited. This suggests that the effectiveness of SegX is related to the initial quality of the XAI outputs it seeks to refine.

### Results of SegU Evaluation

In this experiment, we evaluate the effectiveness of SegU in distinguishing between correct and incorrect predictions of the classification model using the IoU-based score  $c_{\text{IoU}}$  and the AUITC-based score  $c_{\text{AUITC}}$ . By comparing the certainty scores between these groups, we assess whether the SegU approach can effectively differentiate accurate predictions from inaccurate ones.

**SegU Effectively Distinguishes Correct and Incorrect Predictions.** The evaluation results in Table 4 shows that SegU achieves consistently higher  $c_{\text{IoU}}$  and  $c_{\text{AUITC}}$  scores for the correct predictions compared to incorrect predictions across all datasets, models, and XAI methods. For instance, on the HAM10000 dataset, Grad-CAM with ResNet achieves a  $c_{\text{IoU}}$  of 0.112 for correct predictions but only 0.015 for incorrect predictions, with a similar trend in  $c_{\text{AUITC}}$  (0.142 vs. 0.022). A similar pattern is observed for the



Table 3: Comparison between XAI and SegX on HAM10000 and ChestX-Det10

Dataset	Model	XAI	IoU@5% $\uparrow$		AUITC $\uparrow$	
			Original	SegX	Original	SegX
HAM10000	ResNet	SHAP	0.125	<b>0.136</b>	0.128	<b>0.153</b>
		Grad-CAM	0.084	<b>0.093</b>	0.088	<b>0.109</b>
	Densenet	SHAP	0.121	<b>0.131</b>	0.128	<b>0.147</b>
		Grad-CAM	0.176	<b>0.188</b>	0.178	<b>0.214</b>
ChestX-Det10	ResNet	SHAP	0.076	<b>0.084</b>	0.076	<b>0.084</b>
		Grad-CAM	0.033	<b>0.035</b>	0.036	<b>0.039</b>
	Densenet	SHAP	0.061	<b>0.068</b>	0.061	<b>0.070</b>
		Grad-CAM	0.052	<b>0.055</b>	0.054	<b>0.058</b>

Table 4: Evaluation of SegU on correct and incorrect prediction groups

Dataset	Model	XAI	$c_{IoU} \uparrow$		$c_{AUITC} \uparrow$	
			Correct	Incorrect	Correct	Incorrect
HAM10000	ResNet	SHAP	<b>0.133</b>	0.122	<b>0.135</b>	0.122
		Grad-CAM	<b>0.112</b>	0.015	<b>0.142</b>	0.022
	Densenet	SHAP	<b>0.124</b>	0.123	<b>0.124</b>	0.112
		Grad-CAM	<b>0.183</b>	0.092	<b>0.183</b>	0.093
ChestX-Det10	ResNet	SHAP	<b>0.055</b>	0.010	<b>0.079</b>	0.071
		Grad-CAM	<b>0.079</b>	0.072	<b>0.061</b>	0.014
	Densenet	SHAP	<b>0.072</b>	0.055	<b>0.073</b>	0.055
		Grad-CAM	<b>0.086</b>	0.025	<b>0.086</b>	0.028

ChestX-Det10 dataset, where SHAP with ResNet achieves a  $c_{IoU}$  of 0.055 for correct predictions compared to 0.010 for incorrect predictions. This consistent disparity in scores highlights SegU’s ability to assess the quality of interpretability outputs relative to model prediction correctness.

These findings suggest that SegU provides a reliable means to quantify the alignment between saliency maps and clinically significant regions, effectively distinguishing correct predictions from incorrect ones. This distinction is critical for clinical decision-making, as it allows clinicians to identify predictions that require additional scrutiny. By quantifying the uncertainty in interpretability outputs, SegU enhances trust in AI-assisted diagnosis and provides an additional layer of validation to improve model reliability in clinical practice.

#### SegU Struggles with SHAP on Dermoscopic Images.

The results in Table 4 reveal that SegU exhibits only marginal differences in  $c_{IoU}$  and  $c_{AUITC}$  scores between correct and incorrect predictions for SHAP on the HAM10000 dataset. For example, the  $c_{IoU}$  scores are 0.133 and 0.122 for correct and incorrect predictions, respectively, while the  $c_{AUITC}$  scores are 0.135 and 0.122. These narrow score gaps indicate that SegU struggles to distinguish between the two groups when SHAP is used as the XAI method.

SHAP explanations assume feature independence (Aas, Jullum, and Løland 2021), resulting in the explanation maps that highlight pixels distributed across both clinically relevant and irrelevant areas. For dermoscopic images, where lesions are typically clear and localized in the centre, this distributed saliency can lead to similar alignment scores for both correct and incorrect predictions, thereby diminishing

SegU’s ability to distinguish between them effectively.

## Conclusion

In this study, we proposed SegX and SegU to mitigate the limitations of existing KG-XAI methods for medical image analysis, specifically SegX’s effectiveness in multi-disease scenarios and SegU’s potential to improve the reliability of classification models. Our experiments show that SegX consistently improves the relevance of explanation maps to clinical interests across diverse imaging modalities. Furthermore, to ensure that SegX does not obscure incorrect model predictions, we introduced SegU to measure model reliability and provide clinicians with deeper insights into the trustworthiness of predictions. Both SegX and SegU are model-agnostic, making them versatile tools for enhancing medical decision-making by offering improved explanations and robust uncertainty quantification.

Certain limitations remain while our method demonstrates significant potential. First, the IoU may not be the best measurement to calculate the distance between clinically relevant regions and XAI outputs, as IoU is insensitive to small shifts, shape differences and disproportionately penalties on small regions. Second, SegX and SegU require the accuracy of the segmentation model, which may not always capture clinically significant regions with sufficient precision. We believe this is not a critical issue thanks to the advancement of (zero-shot) vision foundation models. Thirdly, we validated SegX and SegU on a chest ray diagnosis dataset and a skin cancer recognition dataset. Future work could explore SegX and SegU on diverse imaging modalities, such as MRI and ultrasound, to help validate their generalizability.

## References

- Aas, K.; Jullum, M.; and Løland, A. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298: 103502.
- Abbasian Ardakani, A.; Airom, O.; Khorshidi, H.; Bureau, N. J.; Salvi, M.; Molinari, F.; and Acharya, U. R. 2024. Interpretation of Artificial Intelligence Models in Healthcare: A Pictorial Guide for Clinicians. *Journal of Ultrasound in Medicine*, 43(10): 1789–1818.
- Barata, C.; Rotemberg, V.; Codella, N. C.; Tschandl, P.; Rinner, C.; Akay, B. N.; Apalla, Z.; Argenziano, G.; Halpern, A.; Lallas, A.; et al. 2023. A reinforcement learning model for AI-based decision support in skin cancer. *Nature Medicine*, 29(8): 1941–1946.
- Borys, K.; Schmitt, Y. A.; Nauta, M.; Seifert, C.; Krämer, N.; Friedrich, C. M.; and Nensa, F. 2023. Explainable AI in medical imaging: An overview for clinical practitioners—Beyond saliency-based XAI approaches. *European journal of radiology*, 162: 110786.
- Cozzi, D.; Cavigli, E.; Moroni, C.; Smorchkova, O.; Zantonelli, G.; Pradella, S.; and Miele, V. 2021. Ground-glass opacity (GGO): a review of the differential diagnosis in the era of COVID-19. *Japanese journal of radiology*, 39(8): 721–732.
- Dai, T.; Zhang, R.; Hong, F.; Yao, J.; Zhang, Y.; and Wang, Y. 2024. UniChest: Conquer-and-Divide Pre-training for Multi-Source Chest X-Ray Classification. *IEEE Transactions on Medical Imaging*.
- Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. 2018. Recent advances in convolutional neural networks. *Pattern recognition*, 77: 354–377.
- Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; and Hussain, A. 2024. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1): 45–74.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Jung, H.-G.; Nam, W.-J.; Kim, H.-W.; and Lee, S.-W. 2023. Weakly supervised thoracic disease localization via disease masks. *Neurocomputing*, 517: 34–43.
- Kato, S.; and Hotta, K. 2024. Adaptive t-vMF dice loss: An effective expansion of dice loss for medical image segmentation. *Computers in Biology and Medicine*, 168: 107695.
- Liu, J.; Lian, J.; and Yu, Y. 2020. ChestX-Det10: chest x-ray dataset on detection of thoracic abnormalities. *arXiv preprint arXiv:2006.10550*.
- Loh, H. W.; Ooi, C. P.; Seoni, S.; Barua, P. D.; Molinari, F.; and Acharya, U. R. 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 226: 107161.
- Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.
- Mall, P. K.; Singh, P. K.; Srivastav, S.; Narayan, V.; Paprzycki, M.; Jaworska, T.; and Ganzha, M. 2023. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, 100216.
- Pinto-Coelho, L. 2023. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering*, 10(12): 1435.
- Ramesh, K.; Kumar, G. K.; Swapna, K.; Datta, D.; and Rajest, S. S. 2021. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(27): e6–e6.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, 234–241. Springer.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
- Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9.
- Yeung, M.; Sala, E.; Schönlieb, C.-B.; and Rundo, L. 2022. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95: 102026.
- Yuan, H.; Hong, C.; Jiang, P.-T.; Zhao, G.; Tran, N. T. A.; Xu, X.; Yan, Y. Y.; and Liu, N. 2024. Clinical domain knowledge-derived template improves post hoc AI explanations in pneumothorax classification. *Journal of Biomedical Informatics*, 104673.
- Yuan, H.; Hong, C.; Tran, N. T. A.; Xu, X.; and Liu, N. 2023. Leveraging Anatomical Constraints with Uncertainty for Pneumothorax Segmentation. *arXiv preprint arXiv:2311.15213*.
- Zhang, X.; Qian, B.; Cao, S.; Li, Y.; Chen, H.; Zheng, Y.; and Davidson, I. 2020. INPREM: An interpretable and trustworthy predictive model for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 450–460.
- Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization.



tion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhu, M.; Zhang, R.; and Wang, H. 2022. Recognizing irrelevant faces in short-form videos based on feature fusion and active learning. *Neurocomputing*, 501: 694–704.