Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bridging the Gap: Explainability Metrics for AI Image-Based Clinical Diagnostics

Albert Sire Langa

alberto.sire@uvic.cat

Universitat de Vic– Universitat Central de Catalunya

Ramon Reig-Bolaño

Universitat de Vic– Universitat Central de Catalunya

Clara Masó-Albareda

Fundació Hospital de la Santa Creu de Vic (FHSC-CHV)

Ariadna Farrés-Serrat

Fundació Hospital de la Santa Creu de Vic (FHSC-CHV)

Sergi Grau Carrion

Universitat de Vic– Universitat Central de Catalunya

Research Article

Keywords: Artificial Intelligence, xAI, Explainable Artificial Intelligence, Wound Diagnosis, Convolutional Neural Networks

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10126184/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) has shown considerable promise in enhancing diagnostic accuracy in clinical practice. However, its integration into healthcare workflows remains hindered by the opaque nature of many AI models, often referred to as “black-box” systems, which lack interpretability and alignment with clinical reasoning. This study introduces a novel evaluation framework that incorporates explainability into the performance assessment of AI-based image classifiers for medical diagnosis. We propose the Explainability Agreement Index (EAI), a new metric designed to quantify the concordance between AI-generated explanations and clinician annotations by measuring the overlap between pixels identified as relevant by both sources, relative to the clinician-defined reference. Clinician annotations were obtained through a consensus-based evaluation by two chronic wounds specialists. Using different convolutional neural networks (CNNs) trained on a public dataset of chronic wound images, we evaluated the model’s diagnostic performance alongside two explainability methods—LIME and saliency maps. While the CNNs achieved classification accuracies of up to 72%, the mean EAI values remained low (ranging approximately between 0.19 and 0.30 across models and methods), highlighting a discrepancy between AI-generated explanations and clinician reasoning. These findings emphasize that classification accuracy does not guarantee clinical trustworthiness or interpretability. The proposed EAI framework provides a practical and quantitative approach to bridge this gap, offering a pathway toward more transparent and clinically reliable AI systems in medicine.

1. Introduction

Artificial intelligence (AI) has rapidly become a cornerstone of decision-support systems across various domains, including high-stakes sectors such as healthcare, the judiciary, and beyond. However, the widespread adoption of AI is often hindered by the opacity of many models, commonly referred to as "black-box" systems [1]. These models, typically based on complex statistical algorithms, produce outputs without offering clear insight into the underlying processes that link input data to results (Fig. 1).

In the healthcare sector, despite the proven potential of AI to enhance clinical decision-making, its adoption remains limited. One major barrier is the perception of AI as a black-box technology, which erodes trust among clinicians and healthcare administrators [2]. Furthermore, a critical gap exists in the alignment of objectives and knowledge between engineers and clinicians—a gap that is especially pronounced in fields like radiology, where AI holds immense promise [3].

This skepticism toward AI is not unwarranted. Most AI algorithms operate by analyzing and processing vast amounts of variables and parameters, producing results through processes that are often inscrutable to healthcare professionals [4]. For instance, consider an AI system that predicts the presence of an infection in a wound based on a visible image captured by a smartphone. Should the patient inquire about the rationale behind this diagnosis, the physician may struggle to provide an explanation, especially when relying on a blackbox model (Fig. 2).

To address these limitations, the field of Explainable Artificial Intelligence (xAI) has emerged, aiming to develop techniques that enhance model interpretability while preserving predictive accuracy. These advancements strive to bridge the gap between complex algorithmic reasoning and human understanding, fostering trust and enabling the effective deployment of AI systems in critical applications [5, 6].

The concept of xAI traces back to academic research milestones, with a pivotal contribution from Johnson et al. in 1994 [7]. Since then, significant progress has been made, culminating in initiatives like the Defense Advanced Research Projects Agency's (DARPA) xAI program. This program underscored the urgent need for transparent AI systems, catalyzing the development of a range of sophisticated tools and methodologies [8].

Today, xAI encompasses a diverse array of techniques, each aiming to make AI more interpretable and actionable. These techniques are often categorized based on their approach to explanation, their scope, and their applicability across different model types (Table 1) [9–10].

Table 1
General classification of xAI techniques.

According to the type of explanation	
Intrinsic Interpretable Model	An Intrinsic Interpretable Model is a type of machine learning model designed with transparency, where humans can directly understand its internal workings, parameters, and predictions. Examples include linear regression, decision trees, or rule-based systems, as they allow for straightforward explanations of their outputs due to their simplicity and logical structure
Post Hoc Explanation	Post Hoc Explanation refers to techniques used to interpret machine learning predictions without providing direct insights into how the underlying model works. These methods extract additional information after the model has made its prediction, using tools like saliency maps, feature importance visualizations, or example-based explanations.
According to the scope of the explanation	
Global Explanation	A Global Explanation method provides an overarching understanding of how a machine learning model operates across its entire input space.
Local Explanation	A Local Explanation method focuses on explaining the predictions of a machine learning model in a specific region of the input space, rather than describing the global behavior of the model.
According to the level of model specificity	
Model-Agnostic Explanation	A Model-Agnostic explanation technique is a post-hoc interpretability approach that can explain the predictions of any machine learning model, regardless of its type or structure.
Model-Specific Explanation	It refers to an interpretability technique designed specifically for a particular type of model.
According to the format of the explanation	
Text Explanation	Text Explanation method provides interpretability by delivering explanations in a textual format that describes the reasoning behind a model’s predictions.
Visualization	Visualization method refers to an interpretability method designed to provide insights into the internal workings of machine learning models through visual representations.
Explanation by Example	Explanation by Example technique involves providing explanations for a model's prediction by presenting similar instances from the training dataset or previously seen examples.

Based on the classification criteria above, Table 2 presents examples of different xAI techniques.

Table 2
Examples of xAI techniques.

Method	Description	Type of xAI method
LIME (Local Interpretable Model-Agnostic Explanations)	This technique generates a local explanation of the model for a specific prediction, showing which input features had the greatest impact on the output [11].	• Post-hoc Explanation • Model-Agnostic • Local explanation • Visualization and Text Explanation
Saliency maps	Saliency maps highlight the regions of an input image that significantly contribute to the model's prediction. These maps are generated by visualizing gradients flowing back to the input layer, thereby identifying areas of the image most influential in the decision-making process. They provide an intuitive way to interpret and validate the model's predictions in tasks such as image classification. [12].	• Post-hoc Explanation • Model-Specific • Local explanation • Visualization
Grad-CAM (Gradient-weighted Class Activation Mapping)	Grad-CAM generates visual explanations for deep learning models by utilizing gradients from the final convolutional layer. These gradients are used to create heatmaps that highlight regions in the input image that are most influential for the model's decision. This approach offers localized and interpretable insights into the model's predictions [13].	• Post-hoc Explanation • Model-Specific • Local explanation • Visualization
SHAP (SHapley Additive exPlanations)	SHAP is a unified framework for interpreting model predictions based on cooperative game theory. It assigns an importance value to each feature by quantifying its contribution to the model's prediction. SHAP ensures consistent, interpretable, and additive explanations, aligning with theoretical properties like local accuracy and consistency [14].	• Post-hoc Explanation • Model-Agnostic • Local explanation • Visualization
XGradCAM (eXplanation with Gradient-weighted Class Activation Mapping)	XGradCAM is an enhanced explainability method built upon the Grad-CAM framework. It utilizes generalized gradients and a feature importance weighting mechanism to generate robust and accurate class activation maps[15].	• Post-hoc Explanation • Model-Specific • Local explanation

Method	Description	Type of xAI method
		• Visualization

In the healthcare sector, particularly in clinical diagnostics, the adoption and integration of AI models into routine practice depends heavily on their explainability and transparency [16]. Explainable AI plays a pivotal role in bridging the gap between complex algorithms and clinical usability, enabling healthcare professionals to trust and effectively utilize Artificial Intelligence Models. For instance, Fig. 3 demonstrates the application of an xAI model in diagnosing infections in chronic wounds, emphasizing how explainability can support clinical decision-making.

Traditionally, the performance of AI-based classification models in clinical environments is assessed using statistical metrics such as accuracy, sensitivity, specificity, and other related indicators (Table 3) [17].

Table 3
List of metrics for measuring diagnostic accuracy (TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, P = Total Positives, N = Total Negatives)

Performance metrics	Description
Accuracy	Accuracy is defined as the proportion of correctly classified instances out of the total number of evaluated cases. It provides an overall measure of the model's diagnostic performance.
Sensitivity	Sensitivity refers to the proportion of true positive cases (TP)—patients with the disease correctly identified by the model—relative to the total number of actual positive cases (TP + false negatives, FN). It is calculated as TP / (TP + FN) and reflects the model's ability to accurately detect the presence of the condition in affected individuals.
Specificity	Specificity is defined as the proportion of true negative cases (TN)—patients without the disease who are correctly classified as negative by the model—relative to the total number of actual negative cases (TN + false positives, FP). It is calculated as TN / (TN + FP) and indicates the model's ability to accurately identify individuals who are free of the condition.
Precision or True Positive Rate (TPR)	The True Positive Rate (TPR), also referred to as precision, is the proportion of true positive cases (TP) relative to all positive predictions made by the model (TP + false positives, FP). It is calculated as TP / (TP + FP) and represents the likelihood that a patient classified as positive by the model truly has the disease.
True Negative Rate (TNR)	The True Negative Rate (TNR) is the proportion of true negative cases (TN) relative to all instances predicted as negative by the model (TN + false negatives, FN). It is calculated as TN / (TN + FN) and reflects the probability that a patient classified as negative truly does not have the disease.

These metrics focus on quantifying the model's predictive correctness. However, while these measures are essential, they overlook a critical aspect: the explainability of the model. This omission raises an

important question—what happens if the explanation provided by an AI model for a clinical diagnosis does not align with the reasoning of a clinician?

Addressing this gap, Venugopal et al. [18] proposed a novel approach to evaluating AI performance in clinical diagnostics by integrating explainability metrics. Their approach integrates traditional statistical metrics with insights derived from explainable AI (xAI) techniques, introducing two novel concepts: Clinical Explainability Failure (CEF) and Explainability Failure Ratio (EFR). CEF refers to situations in which the AI model's classification agrees with the clinician's diagnosis, yet the explanation provided by the xAI method does not align with the clinician's underlying reasoning. The EFR serves as a quantitative measure of how often these discrepancies occur between the AI-generated explanations and the justifications given by clinicians.

This study aims to develop and validate novel metrics to assess the performance of AI algorithms for image-based diagnostic by incorporating explainability as a critical factor. To this end, several convolutional neural networks were trained to diagnose infections in chronic wounds based on images of the wounds. Subsequently, LIME and saliency maps explainability techniques was applied to a set of images from real patients with infected chronic wounds, identifying the pixels that had a positive influence on the diagnosis. A clinician then reviewed the same images and marked the pixels that, according to their clinical judgment, indicated the presence of infection in the wound. Finally, a new explainability agreement index was calculated to assess the performance of the image-based diagnostic neural networks, considering their explainability and offering a more relevant evaluation framework.

In this study, LIME and saliency maps were selected as representative explainability methods due to their widespread use and complementary characteristics. LIME is a model-agnostic approach that provides local, interpretable explanations based on perturbations, while Saliency Maps are gradient-based methods that highlight pixel-level relevance directly from the model. The objective of this work is not to benchmark different explainability methods, but to propose and validate a general framework for assessing the alignment between AI explanations and clinical reasoning. Therefore, these two well-established methods were chosen to illustrate the applicability of the proposed metric, while maintaining methodological clarity and focus.

2. Materials and Methods

A staged approach was adopted to achieve the study objectives.

2.1. Step 1: Baseline performance evaluation

Chronic wound infections occur when open wounds fail to heal due to persistent bacterial colonization, biofilm formation, and excessive inflammation. These infections delay healing, complicate treatment due to antimicrobial resistance, and often require debridement, antiseptics, or systemic antibiotics for management [19].

In the study, different Convolutional Neural Networks (CNN) models was developed to diagnose infections in chronic wounds based on an image of the wounds. Convolutional Neural Networks (CNNs), widely used for image classification tasks, were selected due to their capability to extract and process high-level features from input images [20]. Additionally, these networks support Transfer Learning techniques, allowing the development of high-performance models by leveraging data from different domains [21].

The CNNs models were trained using a public Chronic Wounds dataset [22], which included 545 images of infected wounds and 697 images of non-infected wounds, reflecting a moderately balanced class distribution. All images are labeled at the image level according to the presence or absence of infection.. The images were acquired under real-world clinical conditions, resulting in variability in lighting, background, and acquisition devices, which reflects realistic use scenarios. No pixel-level annotations were available within the original dataset; therefore, expert clinician annotations were generated specifically for this study to support the explainability analysis. The dataset was partitioned into training and validation sets using an 80:20 split ratio. Data augmentation was applied exclusively to the training set to increase the number of samples and enhance model generalization, while the validation set remained unaltered to ensure an unbiased evaluation. An independent test set, consisting of 200 images (100 infected and 100 no infected), was used to assess the final performance of the models.

Developed CNNs utilized transfer learning leveraging different base models pre-trained on ImageNet: ResNet-50 [23], EfficientNet V2B0 [24] and DenseNet121 [25].

The model architecture starts with an input layer designed to receive RGB images of a specified size (224 × 224 × 3). Initially, the image is processed using a GAN-CN-CC network [26], a Pix2Pix-based generative adversarial network that enables the normalization of image illumination conditions by simulating the use of a color checker. Each input image was processed through the network's U-Net generator, which applies an illumination correction learned from pairs of images normalized using a reference color checker. This approach produces normalized images that are independent of both lighting conditions and image acquisition devices, without the need to include a physical marker in the scene.

Subsequently, the base model was loaded without its original classification head and initially frozen, except for the top 25% of its layers. This setup enabled fine-tuning of the higher-level abstract features while retaining the ability to extract low-level features.

After the base CNN, a global average pooling layer was applied to reduce spatial dimensions and summarize feature maps. To mitigate overfitting, a dropout layer with a dropout rate of 30% was incorporated. The final classification component consisted of a fully connected dense layer with a sigmoid activation function, suitable for binary classification tasks (see Fig. 4).

The model was compiled using the Adam optimizer with an initial learning rate of 1×10^{-3} and binary cross-entropy as the loss function. Two callback mechanisms were implemented during training: early

stopping with a patience of 8 epochs, and a learning rate reduction strategy that decreased the learning rate by a factor of 0.1 if validation accuracy plateaued.

A fine-tuning stage was executed. In this phase, all layers of the base model were unfrozen, except for the first 50, and training resumed with a lower learning rate of 1×10^{-4}. The training was extended by an additional 10 epochs beyond the original limit.

The performance of the AI-based diagnostic algorithm was initially evaluated without considering explainability. Standard metrics such as accuracy, sensitivity, specificity, and related indicators were calculated to establish a baseline for the diagnostic model's accuracy.

2.2. Step 2: Explainability analysis with saliency maps

As previously discussed, Saliency maps [12] are a technique used in the field of explainability for CNNs, that enables the visualization of which specific regions of an input image are most decisive for the prediction made by the model. The central idea is to quantify the importance of each pixel in the image with respect to the neural network's output, that is, to identify which pixels exert the greatest influence on the probability assigned to a given class.

From a mathematical perspective, a Saliency map is obtained by computing the gradient of the model's output for the target class with respect to the input image. In other words, it involves calculating the partial derivative of the prediction with respect to the value of each pixel. The magnitude of this gradient measures how much the prediction would change if that pixel were slightly altered, such that high gradient values indicate pixels with greater influence on the model's decision.

When working with RGB images, the computation is performed for each color channel, and subsequently, the maximum absolute gradient value across the three channels is taken for each pixel, resulting in a single grayscale map that indicates the spatial relevance of each region. In our case, we additionally applied a threshold, above which a pixel was considered influential for the prediction, while pixels below this threshold were regarded as non-influential. A threshold value of 0.1 yielded the best results in our experiments.

Therefore, Saliency maps offer the advantage of providing a clear visual explanation of why the model has made a specific decision. Moreover, it aligns with LIME in that both techniques are local and based on visual explanations, which are key features necessary for conducting this study.

The Saliency Maps technique was applied to the set of 10 infected wound images, identifying the pixels that positively influenced the AI-based diagnosis.

2.3. Step 3: Explainability analysis with LIME

In this part of the study, we employed LIME (Local Interpretable Model-agnostic Explanations) [11] as our explainable AI (xAI) technique, due to its capability to provide local interpretability of individual predictions from image classification models (e.g. CNN) through simple surrogate models.

The first step in applying LIME is to represent the input image using features that are inherently interpretable. Instead of evaluating individual pixels—given that isolated pixels rarely significantly influence model predictions—the input image is segmented into superpixels. Superpixels are defined as contiguous groups of pixels exhibiting similar color or textural properties and are generated through visual segmentation algorithms. LIME segments the input image into superpixels using a segmentation algorithm that groups neighboring pixels based on similarity in color and texture. In our implementation, this segmentation is performed using the SLIC (Simple Linear Iterative Clustering) algorithm, which clusters pixels in the combined color–spatial domain to generate compact and perceptually meaningful regions. The number of superpixels and compactness parameters were set accordingly to ensure a balance between spatial coherence and boundary adherence.

Each superpixel serves as an interpretable elementary unit (a "feature" at the explanation level) that can be either activated (present as-is in the original image) or deactivated (masked or hidden) in the local experiments. This segmentation transforms the image into a set of semantically meaningful components, thus laying the foundation for generating intuitive explanations regarding which regions influence the classification decision (Fig. 5).

Once the image for which the prediction is to be explained is segmented into superpixels, LIME generates a set of perturbed samples derived from this segmented image. Perturbations are created by randomly turning off or removing different subsets of superpixels from the segmented image. Practically, "turning off" a superpixel means replacing all its constituent pixels with a neutral value (e.g., a uniform gray color), thus eliminating the informative content of that region. By repeating this process multiple times using various combinations of activated and deactivated superpixels, numerous slightly modified versions of the original image are produced. Each perturbed image retains certain features of the original scene but omits specific regions, allowing the method to probe the importance of each region to the model’s prediction. Essentially, LIME explores the local behavior of the classifier around the given instance, constructing a synthetic dataset composed of perturbations (represented by binary vectors indicating the presence or absence of each superpixel) and their corresponding predictions from the black-box model. Each modified image is subsequently classified by the CNN, and its output prediction is recorded. In this manner, LIME compiles a table representing "perturbed inputs versus model outputs," which serves as the basis for adjusting the local explanation.

Having obtained predictions from the black-box model for all perturbed versions, LIME proceeds by weighting each perturbed example according to its similarity to the original image. The rationale is that perturbed samples closely resembling the original image should exert greater influence on the explanation, thus ensuring the local nature of the interpretation. In the context of images, a common measure of proximity is the fraction of differing superpixels compared to the original image.

With the set of perturbed data and their appropriately weighted predictions, LIME trains an intrinsically interpretable model (in our case, a regression model) to approximate the behavior of the complex classifier in the vicinity of the considered instance. During training of this local model, perturbed

instances contribute proportionally to their weights, meaning samples closer to the original image have a greater influence on model fitting. The result is a linear model intended to predict the original classifier's output using binary variables indicating the presence of each superpixel. The coefficients obtained through this local linear regression indicate the influence of each superpixel on the prediction: a large positive coefficient implies that the presence of that region increases the probability of the target class, whereas a negative coefficient indicates that the presence of that region reinforces a different class or reduces confidence in the target class.

Finally, LIME visually communicates the explanation by highlighting on the original image those regions (superpixels) most influential to the model's prediction. Based on the coefficients derived from the local surrogate model, superpixels with the greatest positive influence on the class of interest (those with the highest positive regression coefficients) are identified and highlighted accordingly.

The LIME technique was applied to the set of 10 infected wound images, identifying the pixels that positively influenced the AI-based diagnosis.

2.4. Step 4: Explainability analysis by clinicians

For the set of 10 infected wound images from the training dataset, two clinicians from the Fundacio Hospital de la Santa Creu de Vic (Catalonia, Spain) specialized in chronic wounds were asked to review each image and reach a consensus on the regions they believed contributed positively to the diagnosis. This process mirrored the principles of LIME/Saliency maps by focusing on feature importance through visual interpretation. To implement the proposed process while eliminating the need for manual annotations, clinicians utilized a purpose-built software that allowed users to mark the different areas of the wound indicating an infection.

2.5. Step 5: Explainability Agreement Index calculation

To quantify the overlap between the explanations provided by LIME/Saliency maps and those from clinicians, an Explainability Agreement Index was calculated. This index quantifies the degree of concordance between the clinician’s explanations and those generated by LIME/saliency maps. Explanations were evaluated at the image level, focusing exclusively on true positive cases in which both the AI model and clinicians agreed on the presence of infection. For each image, the Explainability Agreement Index (EAI) was computed by comparing the set of pixels identified as having a positive influence on the diagnosis by the explainability method and those annotated by clinicians. Specifically, to calculate EAI, the number of pixels that had a positive influence on the diagnosis according to the clinician and also had a positive influence according to LIME/Saliency maps is computed. This value is then divided by the total number of pixels that had a positive influence according to the clinician.

It is important to note that true negative cases are not considered in this analysis because, in such instances, the explanation provided by the AI would have to be compared against the entire image (in clinical practice, a negative diagnosis indicates that no pixel in the wound image suggests the presence

of a possible infection). Such a comparison would offer no practical value, as it would consistently yield an EAI of 1.0.

To ensure consistency across explainability methods, a standardized thresholding strategy was applied to all explanation maps. For saliency maps, pixels exceeding a fixed gradient magnitude threshold were considered to have a positive influence, while for LIME, superpixels with positive contribution weights were selected. This approach ensures that the evaluation is performed under consistent criteria, enabling a fair comparison between different explainability techniques.

3. Results

3.1. Results step 1: Baseline performance evaluation

The CNN models were trained using the prepared dataset, and baseline performance metrics were calculated. The performance metrics on the test dataset are shown in Table 4. This dataset contains 100 images of infected wounds and 100 images of non-infected wounds. It is worth noting that the test dataset had not been seen by the model during its training.

Table 4
List of CNN metrics for measuring diagnostic accuracy (test dataset).

Performance Metrics	ResNet-50	EfficientNet V2 B0	DenseNet121
Accuracy	0.7100	0.7200	0.6600
Sensitivity	0.6800	0.6400	0.4700
Specificity	0.7400	0.8000	0.8500
Precision or True Positive Rate (TPR)	0.7234	0.7619	0.7581
True Negative Rate (TNR)	0.6981	0.6897	0.6159

3.2. Results Step 2: Explainability analysis by saliency maps

Following the methodology, the Saliency maps explainability technique was applied to the same 10 training images described before to generate LIME-derived explanations. Figure 6 shows the Saliency maps-generated results for some of these images.

3.3. Results step 3: Explainability analysis with LIME

Following training, the LIME explainability technique was applied to 10 training images to generate AI-derived explanations (it is important to note that all models correctly classified these wound images as infected wounds). These images represented infected wounds with matching diagnoses from the clinician and AI. Figure 7 shows the LIME-generated results for some of these images.

3.4. Results Step 4: Explainability analysis by clinicians

At the same time, a group of two chronic wound specialists jointly reviewed the same 10 images from the training dataset and reached a consensus by marking the regions they considered to have a positive influence on the diagnosis (Fig. 8).

3.5. Results step 5: Explainability Agreement Index calculation

Finally, an Explainability Agreement Index was computed for each image to measure the overlap between the LIME/Saliency maps-generated explanations and clinician-derived explanations (see Fig. 9).

Tables 5 and 6 show the different values of the Explainability Agreement Index, as well as its mean value.

Table 5
LIME - Explainability Agreement Index per test image.

Image	ResNet-50	EfficientNet V2 B0	DenseNet121
1	0.0495	0.0633	0.2064
2	0.0750	0.0172	0.1755
3	0.0579	0.0000	0.1257
4	0.2016	0.2739	0.2396
5	0.4057	0.4596	0.3078
6	0.2165	0.4935	0.4022
7	0.3308	0.0113	0.1177
8	0.0955	0.3529	0.3548
9	0.4852	1.0000	0.0790
10	0.2021	0.3280	0.2496
Mean	0.2119	0.2999	0.2258

Table 6
Saliency maps - Explainability Agreement Index per test image.

Image	ResNet-50	EfficientNet V2 B0	DenseNet121
1	0.3875	0.1880	0.0824
2	0.1796	0.3932	0.3949
3	0.1907	0.2326	0.2382
4	0.2540	0.1508	0.0311
5	0.2322	0.3375	0.2918
6	0.1562	0.1997	0.2976
7	0.2459	0.0870	0.1396
8	0.3323	0.3191	0.3100
9	0.6901	0.5210	0.0210
10	0.1430	0.1096	0.1397
Mean	0.2811	0.2538	0.1946

The mean value of the Explainability Agreement Index allows for a proper weighting of the CNN's performance indicators. The multi-step approach used ensures a comprehensive evaluation of the AI model's diagnostic performance, integrating both statistical performance metrics and explainability into a unified framework.

4. Discussion

The findings of this study highlight a fundamental challenge in deploying AI systems for medical image analysis: the gap between algorithmic performance, measured through traditional metrics such as accuracy, and the interpretability necessary for clinical trust and adoption. While the CNN models developed for this study achieved reasonable classification performance—up to 72% accuracy, 64% sensitivity and 80% specificity with the EfficientNet V2 B0 architecture—the alignment between AI-generated explanations and clinical reasoning, as captured by the Explainability Agreement Index (EAI), was substantially lower. This disparity underscores that predictive performance alone is insufficient to ensure that AI systems will be accepted and trusted by clinicians in routine practice.

Our results demonstrate that explainability metrics such as the EAI can reveal insights into how well the reasoning of an AI model aligns with human expert interpretations. The low EAI scores suggest that although the models may identify correct diagnoses, the regions of interest highlighted by their visual explanations often differ from those that clinicians consider diagnostically relevant. This misalignment can have critical implications, as clinicians require not only correct predictions but also meaningful justifications for those predictions to integrate AI safely and confidently into decision-making processes.

Our findings are consistent with recent studies reporting limitations of current explainability methods in medical imaging, particularly regarding their robustness and alignment with clinically relevant features. While previous work has focused on improving explainability techniques or proposing hybrid approaches combining deep features with clinically interpretable descriptors, our study takes a complementary perspective by introducing a quantitative framework to evaluate the agreement between AI explanations and clinician reasoning. This positions the proposed EAI as a tool that can be used alongside existing explainability methods to assess their clinical validity.

Moreover, the study provides evidence that the two explainability methods assessed—LIME and saliency maps—are both limited in capturing clinically relevant features with sufficient precision. While both techniques produced localized visual explanations, neither consistently matched the regions annotated by clinicians, which highlights the need for further research into explainability methods tailored specifically for the medical imaging domain. This is particularly relevant for complex tasks such as wound assessment, where subtle visual patterns and contextual information often influence clinical judgments.

The Explainability Agreement Index introduced in this work offers a practical approach to quantitatively assess the alignment between AI explanations and clinical reasoning. However, the EAI itself may be sensitive to the granularity of the clinician's annotations and the thresholding applied to the saliency maps. In this sense, it is important to note that explainability methods, particularly gradient-based approaches such as saliency maps, are known to exhibit instability under small input perturbations. In this study, a consistent and deterministic pipeline was applied across all images and models, including a fixed thresholding strategy to identify pixels with positive influence, ensuring internal consistency and comparability of the Explainability Agreement Index (EAI) results. However, no additional stability-enhancing techniques (e.g., smoothing) were incorporated. As a result, potential variability in the generated explanations may influence the EAI values. This limitation is inherent to the explainability methods employed and is indirectly captured by the proposed metric. Future work should investigate the robustness of EAI under stability-controlled explainability techniques to further strengthen the reliability of the evaluation framework.

Although this study focuses on chronic wound imaging, the proposed Explainability Agreement Index (EAI) is inherently modality-agnostic and can be extended to other medical imaging domains such as radiology or digital pathology, where visual explanations can be compared against expert-derived annotations. However, its applicability depends on the availability of spatially resolved reference standards, such as pixel-level or region-level annotations provided by clinicians or derived from established diagnostic criteria. In settings where such annotations are not available, alternative forms of agreement (e.g., region-based or feature-level comparisons) may be required. Therefore, while the EAI provides a flexible framework, its implementation may need to be adapted depending on the imaging modality and the nature of the clinical ground truth.

In conclusion, our findings emphasize that explainability must be incorporated into the validation framework of AI models destined for clinical use. Metrics such as the EAI can play a critical role in bridging the gap between algorithmic performance and clinical trustworthiness, helping to ensure that AI systems are not only accurate but also transparent and clinically meaningful. Continued research is essential to improve explainability techniques and to integrate them as standard components in the development and evaluation of AI tools for medical imaging.

5. Conclusion

This study underscores the critical importance of integrating explainability into the evaluation and validation of AI models intended for medical image analysis. While convolutional neural networks demonstrated promising diagnostic performance in classifying wound images, our findings reveal that high classification accuracy alone does not ensure clinical reliability or acceptance. The substantial discrepancy observed between model predictions and expert clinical reasoning, as quantified by the Explainability Agreement Index (EAI), highlights a significant gap that must be addressed to build trustworthy AI systems.

The introduction of the EAI provides a quantitative framework for assessing how closely AI-generated visual explanations align with the interpretive patterns of clinicians. Our results show that current explainability techniques, such as LIME and Saliency maps, still fall short in consistently identifying clinically relevant regions, emphasizing the need for further refinement of these methods or the development of new approaches specifically adapted to medical imaging contexts.

Ultimately, bridging the gap between AI performance and explainability is essential to foster clinician trust and to ensure the safe and effective integration of AI into clinical workflows. Future research should focus not only on improving the predictive capabilities of AI models but also on enhancing the interpretability of their outputs, thereby moving closer to AI systems that are both accurate and clinically meaningful.

Finally, while this study focuses on explainability as a key component of clinical trustworthiness, it is important to recognize that trust in AI systems is inherently multifactorial. Aspects such as model calibration, uncertainty estimation, and failure mode analysis also play a critical role in ensuring safe and reliable deployment in clinical settings. Therefore, explainability should be understood as one necessary dimension within a broader framework for evaluating trustworthy AI. Future work should integrate these complementary aspects to provide a more comprehensive assessment of clinical reliability and adoption.

Declarations

Funding:

This research received no external funding.

Data availability: The dataset used in this study is publicly accessible and can be obtained from the original source as cited in reference [22] of the manuscript. Researchers interested in reproducing or extending this work can access the same dataset under the corresponding usage terms. Any additional data generated during the study (e.g., processed outputs or annotations) are available from the corresponding author upon reasonable request.

Author Contribution

Author Contributions: Conceptualization, A.S.; methodology, A.S.; software, A.S.; validation, A.S. and R.R.; clinical study, C.M. and A.F.; writing—original draft, A.S.; writing—review and editing, A.S., R.R. and S.G.; supervision, R.R. and S.G. All authors have read and agreed to the published version of the manuscript.

Data Availability

The dataset used in this study is publicly accessible and can be obtained from the original source as cited in reference [22] of the manuscript. Researchers interested in reproducing or extending this work can access the same dataset under the corresponding usage terms. Any additional data generated during the study (e.g., processed outputs or annotations) are available from the corresponding author upon reasonable request.

References

1. Rudin, C.: Stop Explaining Black-box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat. Mach. Intell. [Online]. **1**(5), 206–215 (2019)
2. Loh, Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., Acharya, U.R.: Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 226, 107161–107161. (2022). https://doi.org/10.1016/j.cmpb.2022.107161
3. Lee, J.G., et al.: Deep learning in medical imaging: general overview. Korean J. Radiol. **18**(4), 570 (2017). 10.3348/kjr.2017.18.4.570
4. Cinà, G., Röber, T.E., Goedhart, R., İlker Birbil, Ş.: Semantic Match: Debugging Feature Attribution Methods in xAI for Healthcare. ArXiv [Cs.IA]. arXiv. (2023). http://arxiv.org/abs/2301.02080
5. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable Artificial Intelligence (xAI): Concepts, taxonomies, opportunities and challenges toward responsible IA. Inf. Fusion. **58**, 82–115 (2020)
6. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., et al.: Explainable IA (xAI): Core Ideas, Techniques, and Solutions. ACM Comput. Surveys. **55**(9), 1–33 (2023)

7. Johnson, W.L.: Agents that learn to explain themselves. In Proceedings of the Twelfth National Conference on Artificial Intelligence, pgs. 1257–1263. (1994)
8. Gunning, Aha, D.W.: DARPA's Explainable Artificial Intelligence (xAI) Program. IA Magazine. **40**(2), 44–58 (2019)
9. Du, M., Liu, N., Hu, X.: Techniques for Interpretable Machine Learning. Commun. ACM. **63**, 68–77 (2019). [CrossRef]
10. Lipton, Z.C., Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. **16**, 31–57 (2018). [CrossRef]
11. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, New York, USA pp. 1135–1144, (2016). 10.1145/2939672.2939778
12. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv arXiv:1312.6034. (2013)
13. Selvaraju, Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int. J. Comput. Vision. **128**(2), 336–359 (2020). https://doi.org/10.1007/s11263-019-01228-7
14. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. Adv. Neural. Inf. Process. Syst., 4765–4774. (2017)
15. Chaddad, A., Hu, Y., Wu, Y., Wen, B., Kateb, R.: Generalizable and Explainable Deep Learning for Medical Image Computing: An Overview. Curr. Opin. Biomedical Eng., 100567. (2024)
16. Zhang, Weng, Y., Lund, J.: Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. Diagnostics (Basel). **12**(2), 237 (2022). https://doi.org/10.3390/diagnostics12020237
17. Dhamnetiya, Jha, R.P., Shalini, S., Bhattacharyya, K.: How to Analyze the Diagnostic Performance of a New Test? Explained with Illustrations. J. Lab. Physicians. **14**(1), 090–098 (2022). https://doi.org/10.1055/s-0041-1734019
18. Venugopal, Takhar, R., Gupta, S., Mahajan, V.: Clinical Explainability Failure (CEF) & Explainability Failure Ratio (EFR) – Changing the Way We Validate Classification Algorithms. J. Med. Syst. **46**(4), 20–20 (2022). https://doi.org/10.1007/s10916-022-01806-2
19. Leaper, D., Assadian, O., Edmiston, C.E.: Approach to chronic wound infections. Br. J. Dermatol. **173**(2), 351–358 (2015)
20. O'Shea, K., Nash, R.: An Introduction to Convolutional Neural Networks. CoRR, abs/1511.08458. (2015)
21. Weiss, Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data. **3**(1) (2016). https://doi.org/10.1186/s40537-016-0043-6
22. Yap, M.H., Cassidy, B., Pappachan, J.M., O'Shea, C., Gillespie, D., Reeves, N.D.: Analysis towards classification of infection and ischaemia of diabetic foot ulcers. In 2021 IEEE EMBS International

Conference on Biomedical and Health Informatics (BHI) (pp. 1–4). IEEE. (2021), July
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778). (2016)
24. Tan, M., Le, Q., Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June ; pp. 6105–6114. (2019)
25. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700–4708). (2017)
26. Langa, A.S., Bolaño, R.R., Carrión, S.G., Elorrieta, I.U.: Color Normalization Through a Simulated Color Checker Using Generative Adversarial Networks. Electron. (Basel). **14**(9), 1746 (2025). https://doi.org/10.3390/electronics14091746

Figures

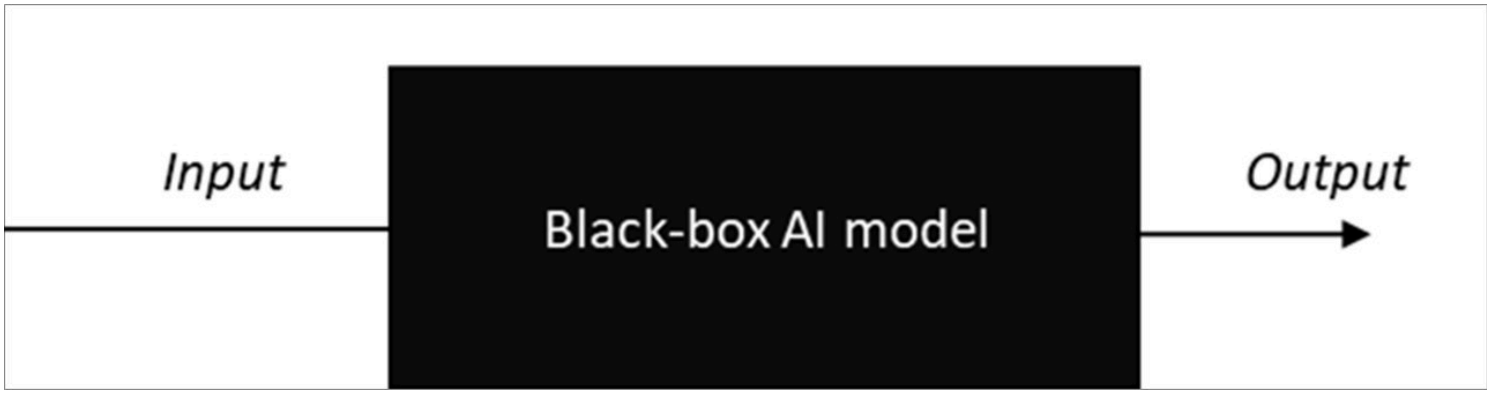


Figure 1

Conceptual framework of a black-box AI model: a) Typically reliant on statistical techniques to derive outputs from inputs b) Opaque in identifying which input features drive specific outputs c) This lack of transparency impacts both end-users and system developers.

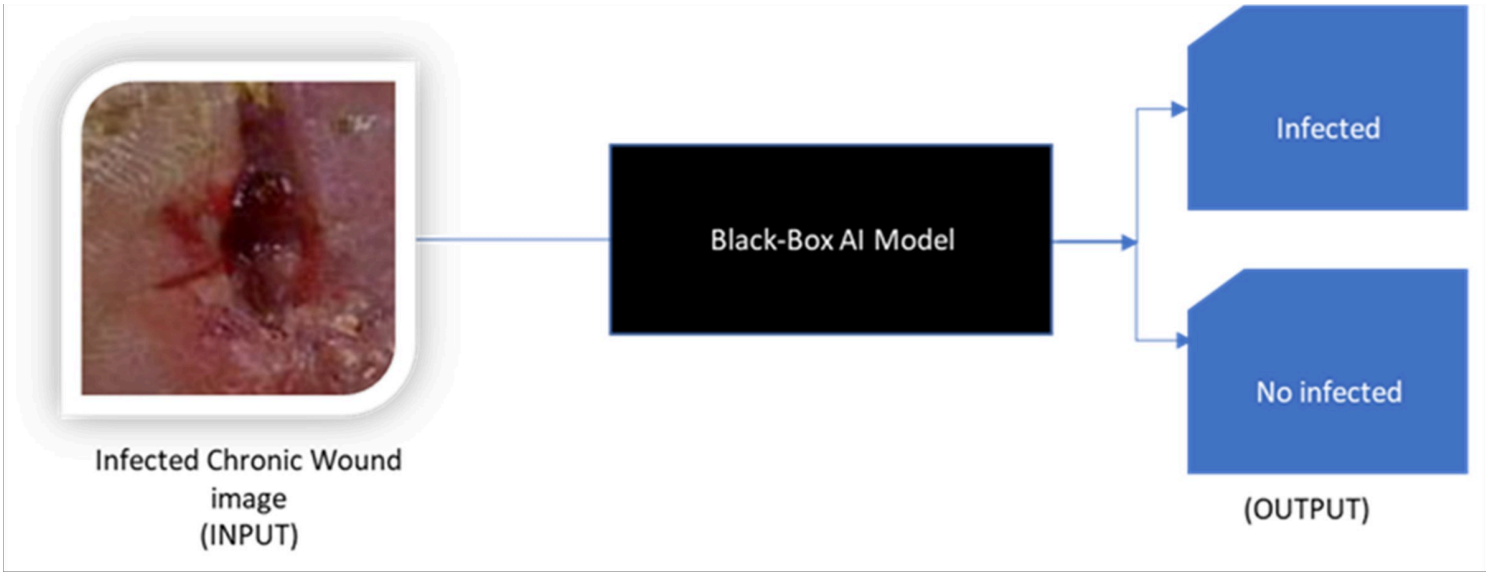


Figure 2

Illustration of a black-box algorithm used for chronic wound infection detection. The algorithm produces an output (infected or non-infected wound) without providing any explanation for the reason behind its classification.

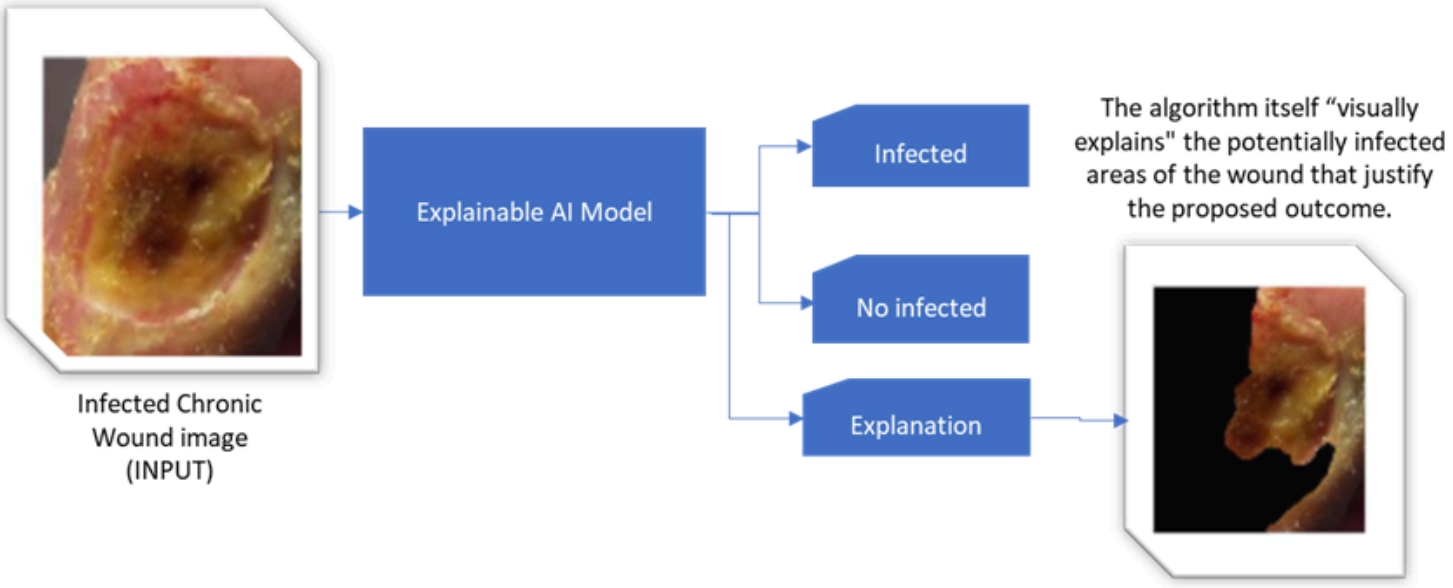


Figure 3

Explainable AI model for chronic wound diagnosis.

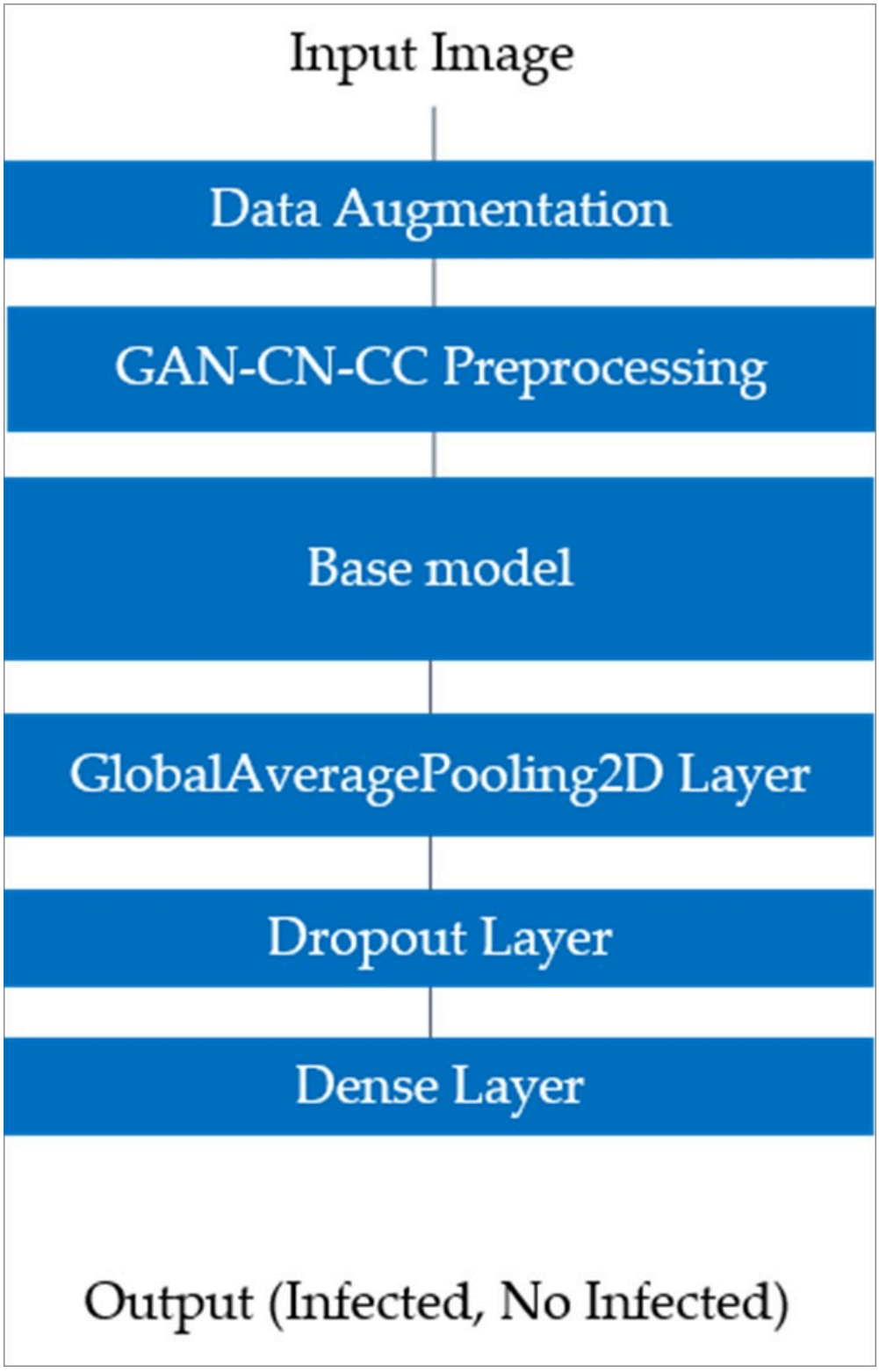


Figure 4

After the base CNN, a global average pooling layer was applied. To mitigate overfitting, a dropout layer was subsequently added. The final classification was performed using a fully connected dense layer, suitable for binary classification tasks.

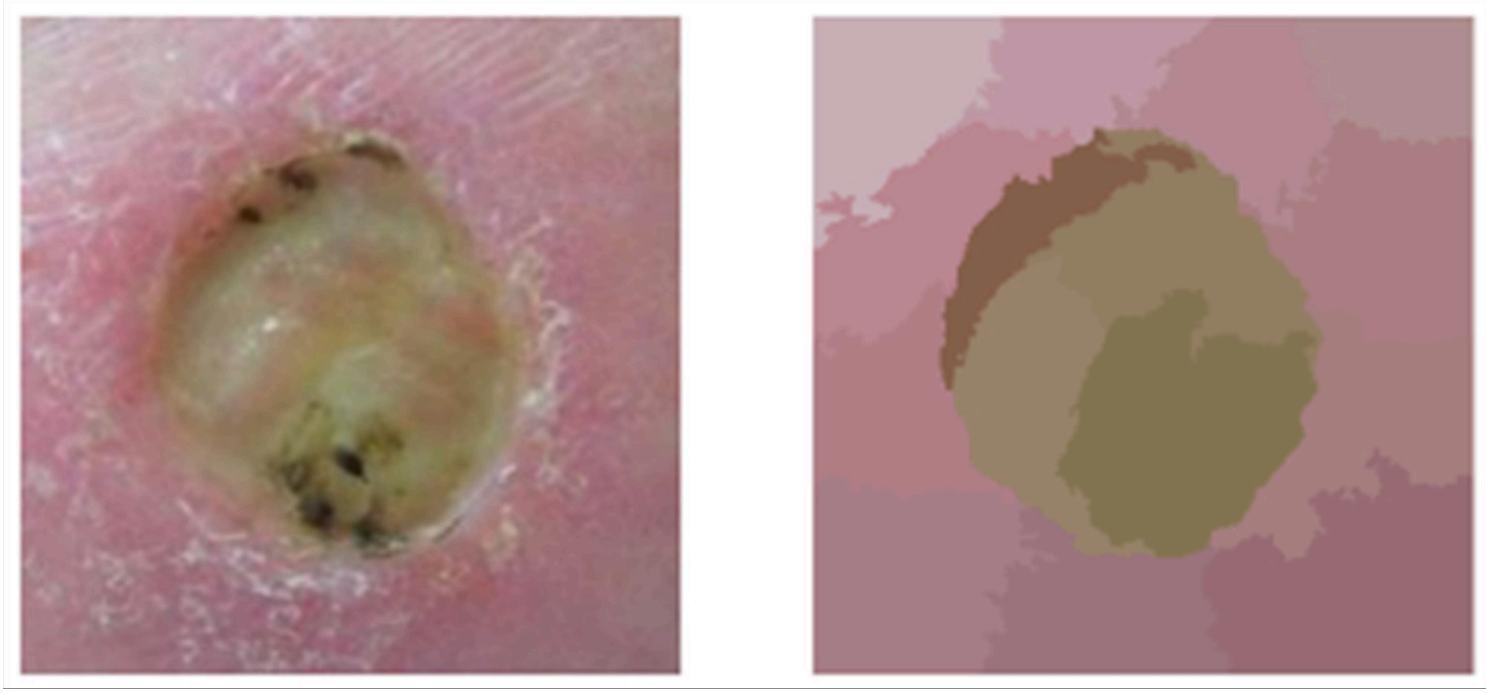


Figure 5

a) Original image b) Superpixels segmentation

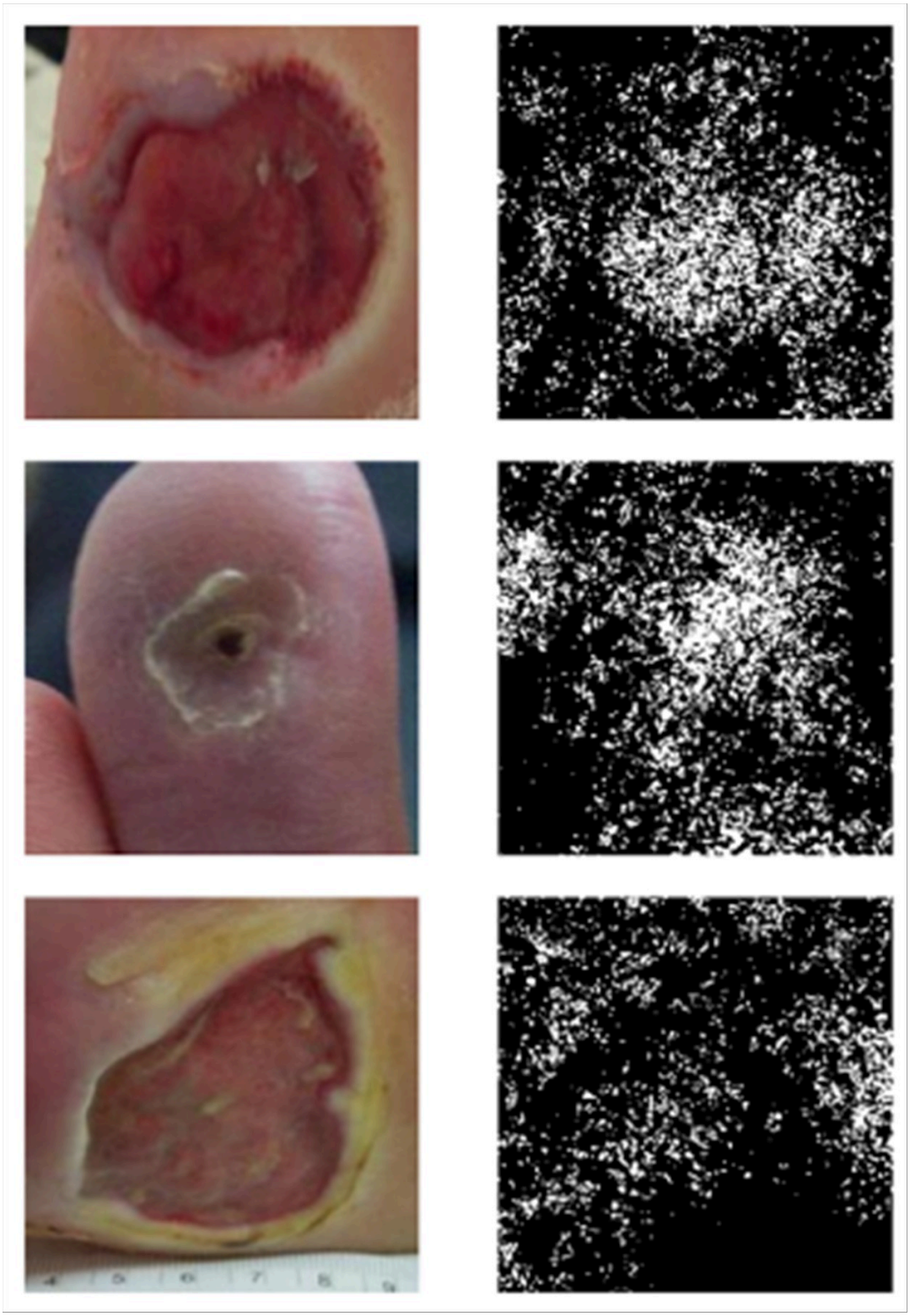


Figure 6

Example of Saliency maps-generated results. a) Original image b) Pixels with positive influence in the diagnosis.

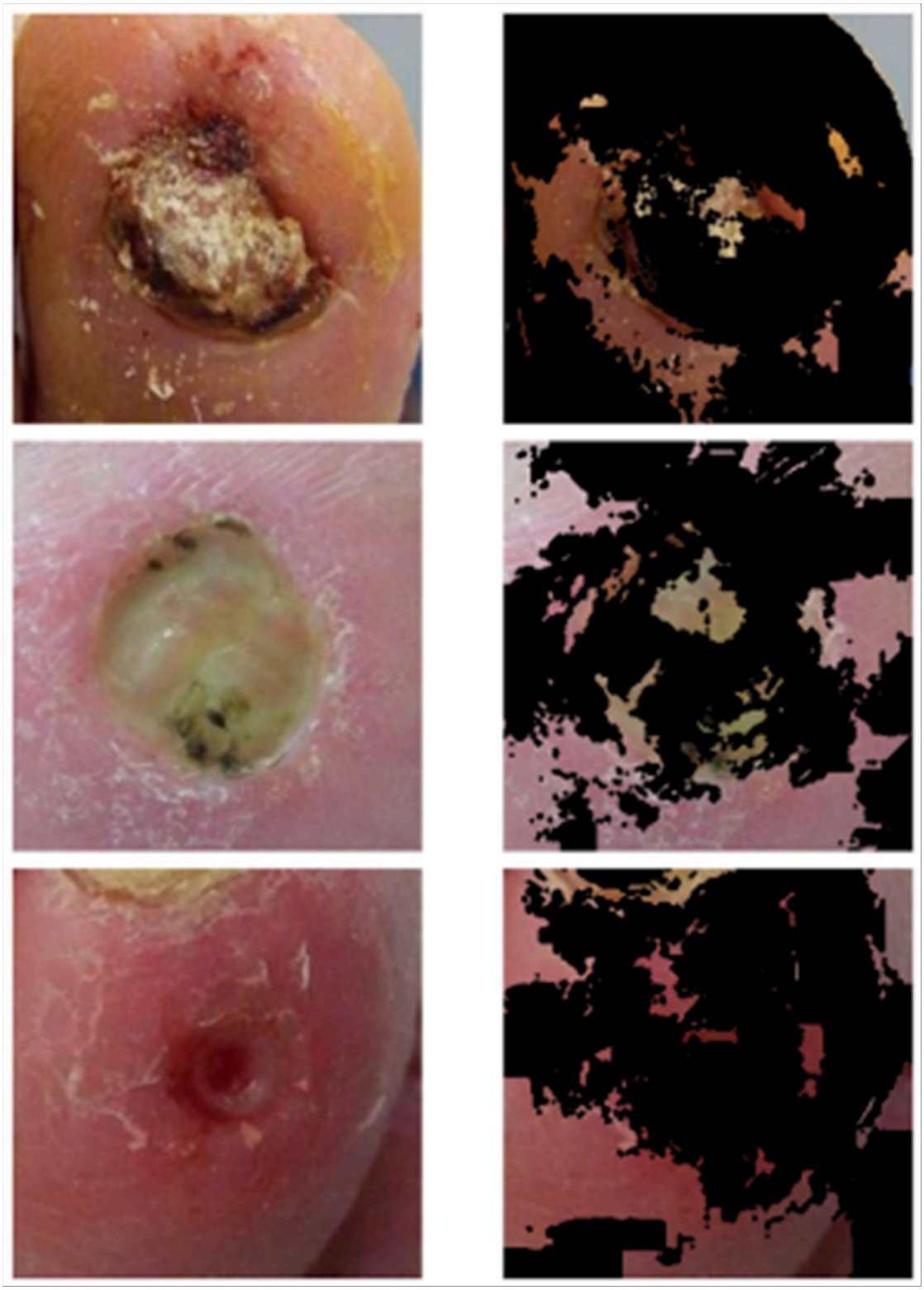


Figure 7

Example of LIME-generated results. a) Original image b) Superpixels with positive influence in the diagnosis.

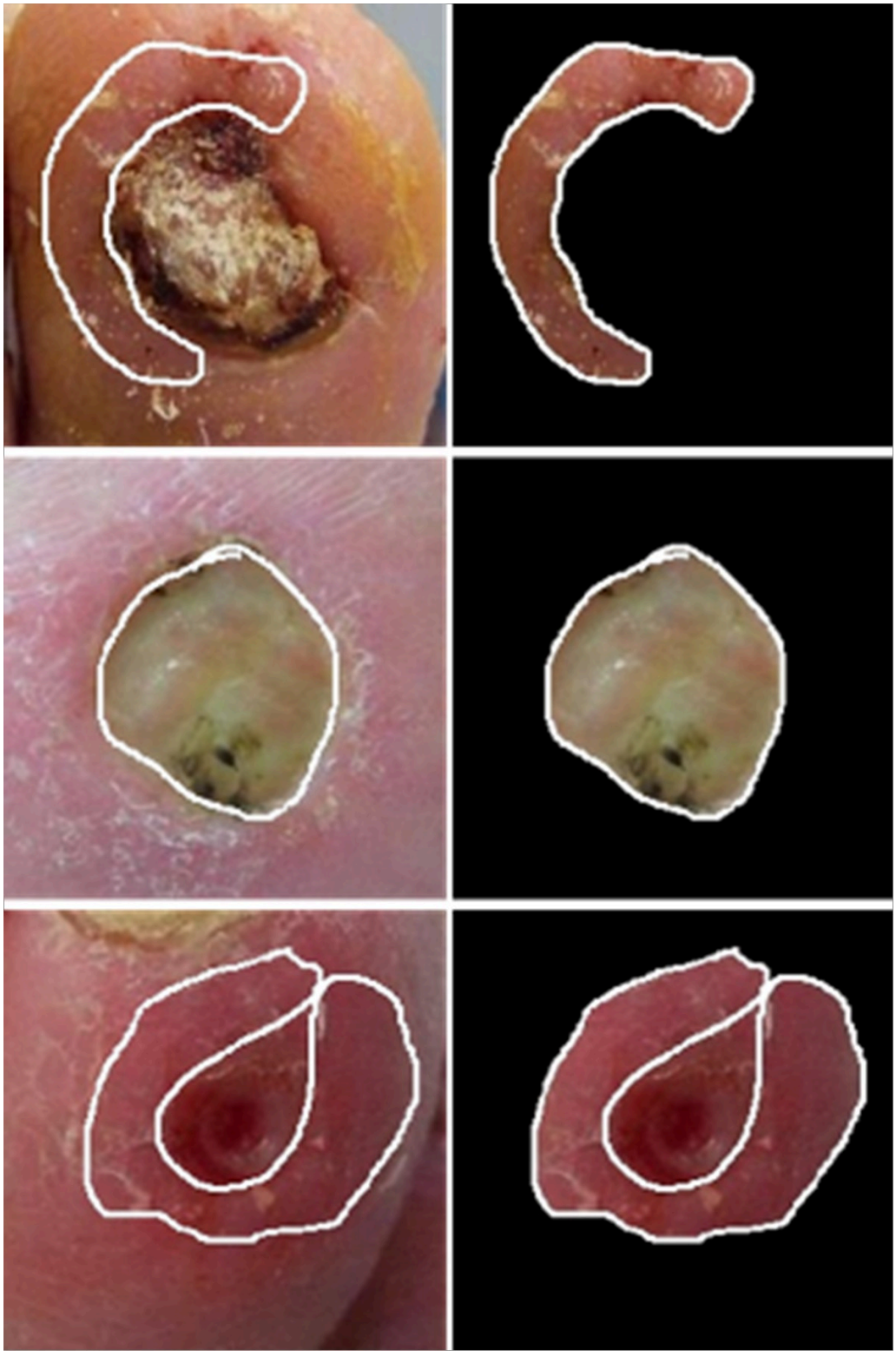


Figure 8

Example of clinician ’s explanations. Within the white boundaries are the areas with a positive influence on the diagnosis.

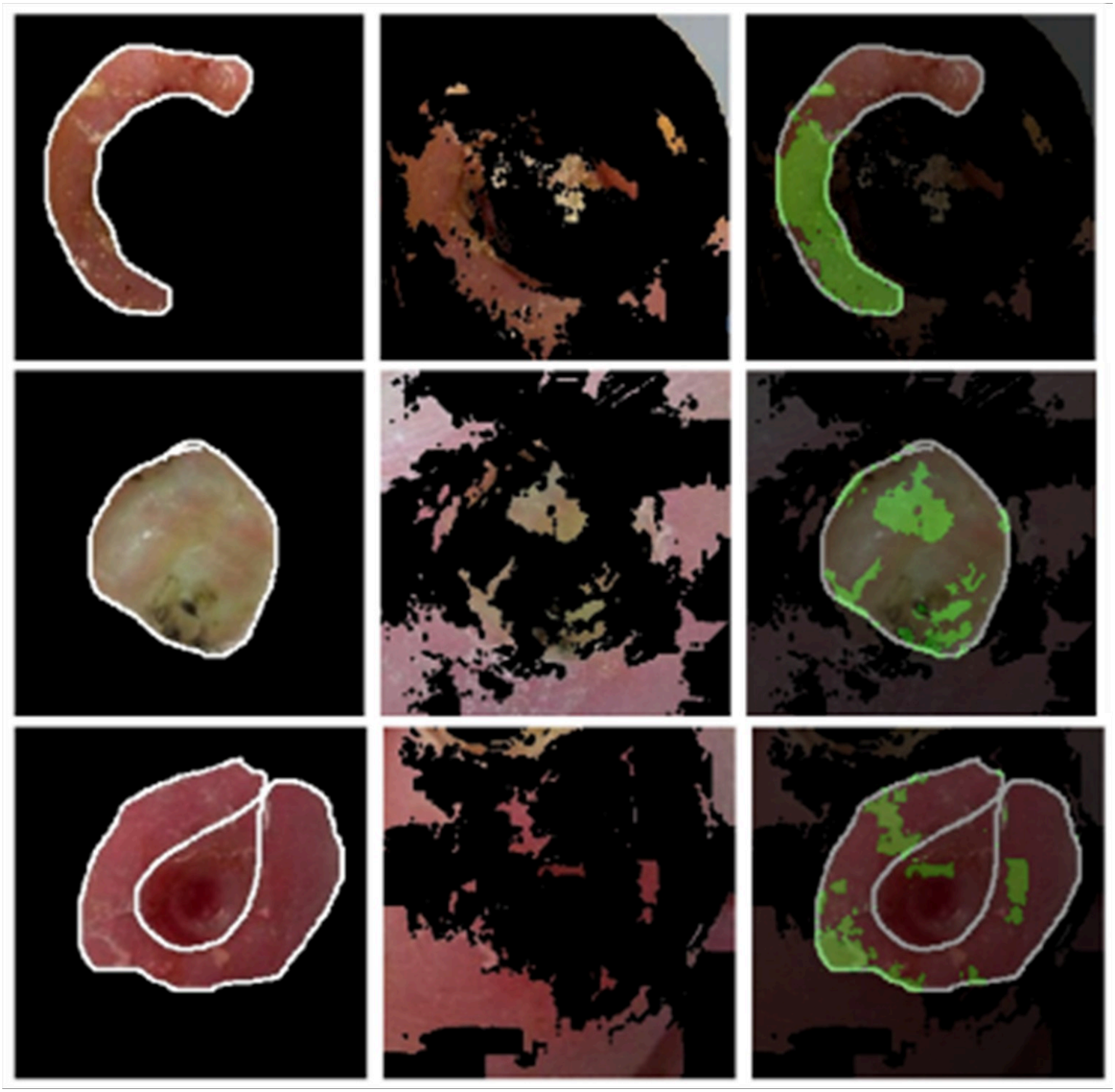


Figure 9

(a) Clinician explanations (b) LIME explanations (c) In green, pixels that match between the clinician's explanations and the explanations provided by LIME.