

# XAIMed: A Diagnostic Support Tool for Explaining AI Decisions on Medical Images

Mattia Daole<sup>1</sup><sup>a</sup>, Pietro Ducange<sup>1</sup><sup>b</sup> Francesco Marcelloni<sup>1</sup><sup>c</sup>, Giustino Claudio Miglionico<sup>1</sup><sup>d</sup>,  
Alessandro Renda<sup>1</sup><sup>e</sup> and Alessio Schiavo<sup>1,2</sup><sup>f</sup>

<sup>1</sup>*Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, Pisa 56122, Italy*

<sup>2</sup>*LogObject AG, Thurgauerstrasse 101a, Opfikon, 8152, Switzerland*  
{mattia.daole, giustino.miglionico, alessio.schiavo}@phd.unipi.it,  
{pietro.ducange, francesco.marcelloni, alessandro.renda}@unipi.it

**Keywords:** Explainable AI, Deep Learning, Medical Image Analysis, Convolutional Neural Networks, Saliency Maps, Diagnostic Support Tool


**Abstract:** Convolutional Neural Networks have demonstrated high accuracy in medical image analysis, but the opaque nature of such deep learning models hinders their widespread acceptance and clinical adoption. To address this issue, we present XAIMed, a diagnostic support tool specifically designed to be easy to use for physicians. XAIMed supports diagnostic processes involving the analysis of medical images through Convolutional Neural Networks. Besides the model prediction, XAIMed also provides visual explanations using four state-of-art eXplainable AI methods: LIME, RISE, Grad-CAM, and Grad-CAM++. These methods produce saliency maps which highlight image regions that are most influential for a model decision. We also introduce a simple strategy for aggregating the different saliency maps into a unified view which reveals a coarse-grained level of agreement among the explanations. The application features an intuitive graphical user interface and is designed in a modular fashion thus facilitating the integration of new tasks, new models, and new explanation methods.


## 1 INTRODUCTION


In recent years, Deep Learning (DL) models have achieved remarkable success across various fields, including Computer Vision, Natural Language Processing, and Cybersecurity (Shinde and Shah, 2018; Raghu and Schmidt, 2020). In healthcare, the capability of DL models to analyze and recognize patterns can significantly enhance the interpretation of large volumes of medical images, such as those obtained from Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and X-ray scans. Applications include the detection of anatomical structures, segmentation, classification, prediction, and computer-aided diagnosis (Shen et al., 2017). In this context, Convolutional


Neural Networks (CNNs) are particularly suited for analyzing medical images due to their ability to learn spatial hierarchies of features.


While advances in DL led to the development of AI-empowered clinical diagnoses support systems with performance comparable with that of clinicians (Sokolovsky et al., 2018; Tschandl et al., 2019; Hannun et al., 2019), the widespread adoption of such systems is still hindered by several challenges, mainly regarding the reliability of model diagnoses (Shortliffe and Sepúlveda, 2018), the clinical soundness of model behaviors (Magrabi et al., 2019) and the lack of trustworthiness and transparency in the decision-making process (Shortliffe and Sepúlveda, 2018; Solomonides et al., 2021). The concept of *trustworthy AI* has recently been considered also by government entities, as witnessed, for example, by the adoption of the AI Act<sup>1</sup> and of the GDPR<sup>2</sup>: AI systems should be accountable and transparent, and their decisions should be understood and trusted by human


<sup>a</sup> <https://orcid.org/0009-0005-2708-2805>

<sup>b</sup> <https://orcid.org/0000-0003-4510-1350>

<sup>c</sup> <https://orcid.org/0000-0002-5895-876X>

<sup>d</sup> <https://orcid.org/0009-0003-0093-4735>

<sup>e</sup> <https://orcid.org/0000-0002-0482-5048>

<sup>f</sup> <https://orcid.org/0009-0005-2147-2853>

<sup>1</sup><https://artificialintelligenceact.eu/the-act/>

<sup>2</sup><https://gdpr-info.eu/>

users. According to GDPR all individuals have the right to obtain “meaningful explanations of the logic involved” (Guidotti et al., 2018).

The need for trustworthiness in AI systems contributes to the rise of eXplainable Artificial Intelligence (XAI) (Doshi-Velez and Kim, 2017): XAI aims to explain AI algorithms and their decision-making processes, enhancing trust and facilitating the integration of AI into critical domains such as healthcare, (Samek et al., 2021; Arnaout et al., 2021; DeGrave et al., 2021). In medical imaging, one of the most common XAI approach relates to the generation of Saliency Maps (SMs) (Borys et al., 2023). SMs are visual representations that highlight regions of an image that are most influential in a model decision, thus enabling healthcare professionals to assess the clinical relevance of these regions (Simonyan et al., 2014).

Evaluating the effectiveness of XAI methods, including those based on SM, is inherently complex due to the subjective nature of interpretation. Quantitative evaluations include metrics that measure the alignment between areas identified through SM and those identified by expert clinicians. A misalignment may arise when DL models learn complex functions to map input images to classes without necessarily capturing clinically relevant features. It has been shown that sometimes they rely on shortcuts, such as artifacts or specific markings in images (e.g., logos or text labels), which fictitiously improve classification accuracy but do not correspond to true modeling of diagnostic features (Lapuschkin et al., 2019; Geirhos et al., 2020). Consequently, while a model might classify an image accurately, the regions it focuses on may not align with those clinicians consider important for diagnosis, potentially compromising the model trustworthiness and hindering computer-aided diagnosis (Cerekci et al., 2024).

In a recent study (Barda et al., 2020), authors highlight three key components for designing explanations in clinical diagnosis support systems: *why* the system provided a certain diagnosis, *what* should be included in the explanation, and *how* explanations should be presented to users. Authors in (Hwang et al., 2022) proposed a user-centered clinical decision support system designed to assist clinical technicians in reviewing AI-predicted sleep staging results. The aim of the authors was to address the lack of clinical interpretability and user-friendly interfaces in existing AI systems. Their findings suggest that integrating clinically meaningful explanations into AI systems through a user-centered design process is an effective strategy for developing a clinical diagnosis support system for sleep staging. The study highlights the importance of providing explanations that

align with clinical knowledge and workflows, which can enhance the adoption and utility of AI in clinical practice. The user interface of the AI-based clinical diagnoses support systems should be practical in clinical environments, where the time and resources of clinicians are constrained (Holzinger et al., 2017; Shortliffe and Sepúlveda, 2018). The development of these tools could alleviate time-consuming and costly clinical tasks and also enhance the performance of clinical practitioners (Younes and Hanly, 2016).

To practically assess the usefulness of DL models for medical image classification, a combination of both specialized medical expertise and DL knowledge is necessary. Medical professionals can provide insights into the clinical significance of the identified regions, while DL experts can contribute on unfolding the model’s functioning by designing and leveraging explainability techniques. Evidently, the development of software and tools to facilitate this synergy is crucial to foster trustworthiness in AI-empowered diagnostics.

In this paper we present XAIMed, short for eXplaining AI decisions on Medical images, a user-friendly application designed to contribute in bridging the gap between the opaque nature of CNNs and the need for explainability in clinical settings: our primary goal is to provide clinicians with a practical tool to assess models behavior through intuitive visualizations. The application allows clinicians to choose between several classification tasks for medical images and to add new tasks as needed, supported by a technical operator. For each task, clinicians can select specific images for analysis, view the confidence level of predicted classes, and obtain contextual explanations through four SM-based explainability methods: LIME (Ribeiro et al., 2016), RISE (Pet-siuk et al., 2018), Grad-CAM (Selvaraju et al., 2016), and Grad-CAM++ (Chattopadhyay et al., 2018). These methods are widely used in healthcare and extensively studied in literature (Borys et al., 2023). Additionally, the proposed application provides a visualization that aggregates the four SMs into a comprehensive view, by locally quantifying the level of agreement between the explanations. This tool aims to facilitate collaboration between DL experts and medical professionals by providing a practical means to explain the model behaviour on specific images of interest. This approach allows medical professionals to assess the model predictions based on their expertise, understand how often they agree with the model decisions, and determine whether the features identified by the model align with their diagnostic criteria.

The rest of this paper is organized as follows. In Section 2 we provide some preliminaries on CNNs

and SM-based explanation methods. In Section 3 we describe our application. In Section 4 we illustrate an example of use case on Invasive Ductal Carcinoma detection. Finally, in Section 5 we draw some concluding remarks.

## 2 BACKGROUND

The adoption of CNNs for classification of medical images is revolutionizing the healthcare sector, enabling faster and more accurate diagnoses (Litjens et al., 2017; Yamashita et al., 2018). These models consist of multiple layers, typically including convolutional layers, pooling layers, and fully connected layers. Convolutional layers are designed to automatically and adaptively learn spatial hierarchies of features by applying a set of filters on the input images. The parameters of these filters are learned during the training process, allowing the network to extract relevant features from the input data (LeCun et al., 1998). The layers closer to the input learn low-level features such as edges, textures, and simple shapes, while deeper layers learn more complex, high-level features such as parts of objects and entire objects (Krizhevsky et al., 2012). This hierarchical learning makes CNNs highly effective for image classification tasks. Pooling layers are used to reduce the dimensionality of the feature maps, thus decreasing the computational load and reducing the risk of overfitting. Moreover, these layers help to make the network invariant to small translations of the input image, improving its robustness (Scherer et al., 2010). The downstream classification task is typically accomplished through a fully connected network: the first layer takes as input the high-level features extracted from the convolutional backbone, whereas the last layer returns the probability associated with each class by exploiting appropriate activation functions, namely sigmoid and softmax in the binary and multi-class case, respectively. Notably, such probabilities can be considered as a proxy for model confidence in the decision making process.

Saliency maps (SMs) are a prominent technique in XAI, providing visual explanations by highlighting regions of an image that significantly influence a model predictions. Saliency-based approaches designed for CNNs leverage the spatial information preserved through convolutional layers to identify parts of an image that contribute most to the resulting decision (Van der Velden et al., 2022). SMs explain why a trained opaque model takes a certain decision for any single input instance: as such they are considered a *local* post-hoc explanation method. The salient parts of an image, that have the highest attribution to the

prediction, are highlighted in attribution maps. These maps are typically represented as heatmaps where a suitable color code indicate the contribution to the model output (Ancona et al., 2017). Visual explanations are particularly relevant in medical image analysis due to their ease of understanding, which helps ascertain whether a model decision-making aligns with that of a clinician.

The generation of attribution maps can be categorized into perturbation-based and backpropagation-based methods (Singh et al., 2020). *Perturbation-based* methods analyze the effect of altering input features on the model output. This is typically achieved by removing, masking, or modifying parts of the input image, performing the forward pass to compute the model’s prediction, and measuring the deviation from the initial prediction. *Backpropagation-based* methods use gradients and activations during the backpropagation stage of DL models to estimate the impact of each input feature. Specifically, derivatives of the model output with respect to each input dimension, such as every pixel in an input image to the model, are computed. If the gradient is large, it implies that even a tiny change in that dimension may drastically change the model’s output, testifying the importance of that dimension.

In the following, we focus on four popular saliency-based methods, which are integrated into our application: LIME, RISE, Grad-CAM, and Grad-CAM++.

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) is a perturbation-based method designed to explain individual predictions of an opaque model by approximating it locally with a simpler, interpretable model. First, a dataset of perturbed samples is created by slightly altering the original input. For image data, this involves grouping pixels into superpixels, which are contiguous regions with similar pixel intensities, and then randomly switching off these superpixels by setting their values to a baseline, such as zero or the median value. Then, the primary, opaque, model is used to predict the outcomes for each perturbed sample, and these predictions, along with the perturbed samples, form a new dataset. LIME assigns weights to the perturbed samples based on their proximity to the original input, ensuring that samples more similar to the original input have a greater influence on the surrogate model. An interpretable surrogate model is then trained on this new dataset, where the inputs are the perturbed samples and the outputs are the predictions of the primary model. Examples of commonly used interpretable surrogate models are Linear Regression, which fits a linear model to the data and pro-

vides straightforward coefficients indicating the importance of each feature, or Decision Trees, which create a tree-like model of decisions where the importance of features can be easily visualized and understood, offering a balance between simplicity and predictive power. Once trained, the surrogate model is used to explain the primary model’s prediction for the original input. As the surrogate model is inherently interpretable, its parameters, such as coefficients in linear regression or splits in a decision tree, can be used to understand which parts of the input were most influential in the prediction. LIME strengths lie in its model-agnostic nature, making it applicable to any type of model, and its ability to provide local explanations that are specific to individual predictions. However, it can be computationally intensive due to the need for multiple perturbations and it may produce coarse attribution maps due to the superpixel approach.

RISE (Randomized Input Sampling for Explanation of Black-box Models) (Petsiuk et al., 2018) generates saliency maps by randomly masking parts of the input image and observing the impact on the model’s predictions. In a nutshell, it creates a large number of binary masks, where each mask randomly occludes different parts of the image. These masks are then applied to the input image to generate perturbed versions of the image. The model prediction scores (class probabilities) for each perturbed image are recorded. The importance of each region in the original image is determined by aggregating the prediction scores and weighting them according to the presence of each pixel in the binary masks. Essentially, areas that consistently affect the model’s prediction when occluded are identified as important. This method provides robust and comprehensive SMs, but it requires numerous forward passes through the model, thus being computationally expensive. Furthermore, the results depend on the predefined parameters used for generating the masks (Cooper et al., 2022), e.g., the number of masks, their size, and the fraction of pixels occluded in each mask.

Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2016) calculates the gradients of a target class flowing into the final convolutional layer of a CNN to produce a localization map highlighting important regions. Grad-CAM is specifically designed for CNNs and is computationally efficient. It effectively highlights regions with high-level semantics and detailed spatial information, although it tends to produce coarse maps that may lack fine-grained details (Chattopadhyay et al., 2018). Despite this limitation, Grad-CAM efficiency and ability to highlight important regions in CNNs make it practi-

cal for real-time interpretability.

Grad-CAM++ builds upon Grad-CAM by providing localization through a weighted combination of gradients. While Grad-CAM is simpler and faster, making it suitable for analysis of images where a single class is dominant, Grad-CAM++ offers greater precision in images involving multiple objects or classes.

The choice of LIME, RISE, Grad-CAM, and Grad-CAM++ is based on their complementary strengths and widespread use in the literature. Specifically, Grad-CAM and Grad-CAM++ are among the most frequently used methods in the medical field due to their effectiveness in highlighting relevant features in CNNs (Borys et al., 2023). Although these methods are often applied to the feature maps of the last convolutional layer, they can be configured to focus on different layers depending on the desired trade-off between high-level abstract information and detailed spatial resolution. LIME and RISE, on the other hand, utilize perturbation techniques that consider the entire model, offering an overall perspective that considers the entire neural network. It is important to note that LIME and RISE are computationally more intensive compared to Grad-CAM and Grad-CAM++. The perturbation-based approach of LIME and RISE requires generating numerous variations of the input data and analyzing the model’s responses, which can result in a delay, making the output available only after a few seconds. In contrast, Grad-CAM and Grad-CAM++ are more efficient as they directly utilize the gradients flowing through the network, allowing the SMs to be produced almost instantaneously. In our application, the SMs produced by these four methods are aggregated into a cumulative explanation view within our application, combining the immediate, detailed insights from Grad-CAM and Grad-CAM++ with the comprehensive, albeit slower, perspectives from LIME and RISE.

### 3 THE PROPOSED XAIMed APPLICATION

XAIMed, the acronym for “eXplaining AI decisions on Medical images”, is a desktop application designed to support doctors during the diagnostic process of medical examinations based on medical images. XAIMed enables image classification by leveraging CNN models and employs well-established XAI methods to complement diagnoses with visual saliency-based explanations. SMs highlight the regions of the image that contributed most to the model decision. Our application features a user-friendly

Graphical User Interface (GUI), which allows doctors to easily navigate and utilize its diagnostic support capabilities. It is worth noticing that XAIMed is designed in a modular fashion which makes it very straightforward to include new tasks, new models, and new explanation methods. In the following, we first discuss the use cases of XAIMed and then present the proposed approach for obtaining saliency-based explanations.

### 3.1 XAIMed Use Cases

The Use Case diagram of XAIMed is reported in Fig. 1.

XAIMed envisions two types of users: a technical user, whose use cases are highlighted in yellow, and a physician user, whose activities are highlighted in green. The use cases shared by both actors are depicted in blue.

The technical user configures the application so that the physician can exploit its functionalities. For simplicity, we will refer to a medical imaging diagnostic use case as a “task” throughout the rest of the paper.

Technical users can configure tasks, add task descriptions, delete existing tasks, add DL models along with their metadata, and delete DL models. As for the task configuration, a dedicated folder must be created and named to identify the task: such name will be used within the application for displaying purposes. The folder will contain a text file with a description of the task: contextual and domain-specific information can be included to help physician users understand task details. A task can be associated with one or more DL models, i.e., CNNs. For each model, a dedicated subfolder must be created within the task folder. Each model directory must contain the files needed for its deployment, such as the files with the weights and the specifications of the architecture. The model directory will also include a brief textual description of the dataset and the CNN model, detailing relevant information including the number of images used for training and testing, their resolution, their distribution across target classes, and the inference time of the CNN model. Furthermore, a descriptive image that helps users understand the model performance can be included. For instance, such visual aid may consist of a confusion matrix.

Beside configuring tasks, the technical user can add new CNN models, along with the related information, to existing tasks. The technical user can also remove tasks and models as needed.

Once the tasks are configured according to these specifications, the application is readily available for

use by a physician user, who interacts with the system through the GUI.

Both physician and technical users can browse the available tasks and explore their descriptions. For each task, the GUI lists one or more CNN models, along with their respective textual and visual descriptions.

The physician users can select a CNN model and the folder containing the medical images they want to analyze within XAIMed. Notably, the physician can switch between different tasks, models, image folders and can also add or remove images within the specified folder at any time. The images compliant with a selected model are displayed in a list. The physician can browse the list of images: when an image is selected, the diagnosis, i.e., the model output, and the associated confidence values are automatically displayed. For interpretability purpose, we also discretized the confidence values, i.e. the class probabilities, through equal-width discretization with three bins. The probability range  $[0, 1]$  of each class is divided into *Low*  $[0, 0.33]$ , *Medium*  $(0.33, 0.66]$ , and *High*  $(0.66, 1]$  in order to provide the physician users also with a coarse-grained information regarding the confidence level.

Furthermore, for any selected image, the physician can generate SMs using the XAI methods described in Section 2, namely GradCAM, GradCAM++, RISE and LIME. The generated maps are displayed collectively within the same window, providing an overall perspective of the results. A detailed view can be accessed for each SM and such explanations can be saved as image files. In the following subsection, we discuss how detailed saliency-based explanations are obtained.

### 3.2 Saliency-based explanations

In their detailed view, an input image and the associated SMs are partitioned into a grid of nine equally sized square cells. For each cell of the grid and for each XAI method, the following descriptive statistics are calculated based on the saliency value attributed to the pixels within the cell: mean, median, minimum, maximum, and standard deviation. The three cells with the highest mean value of saliency attribution for each SM are adequately highlighted to indicate the regions with the greatest overall impact on the diagnosis provided by the CNN model according to the associated XAI method.

The proposed application also provides a cumulative visual explanation obtained by aggregating the four SMs into a single one. The rationale for this operation is depicted in Fig. 2.



162 whole-mount slide images of breast cancer specimens. These specimens have been scanned at a magnification of 40x to facilitate the identification of IDC, the most prevalent subtype among all breast cancers. The image patches, each measuring  $50 \times 50$  pixels and featuring 3 channels (RGB), are categorized into two groups: 78,786 patches labeled as IDC-positive and 198,738 patches identified as IDC-negative. Note that the case study considers a binary classification task but the application also supports multi-class classification problems.

To address the IDC classification task, a simple CNN model is adopted: its architecture is detailed in Table 1.

Table 1: CNN architecture adopted for the IDC classification case study.

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Name	Type					
Input	Image	1	50x50x3	-	-	-
1	Convolutional	16	50x50x16	3x3	1	ReLU
	BatchNorm	16	50x50x16	-	-	-
	Max Pooling	16	25x25x16	2x2	2	-
4	Convolutional	32	25x25x32	3x3	1	ReLU
	BatchNorm	32	25x25x32	-	-	-
	Max Pooling	32	12x12x32	2x2	2	-
7	Fully connected	-	256	-	-	ReLU
	Dropout	-	256	-	-	-
9	Fully connected	-	1	-	-	Sigmoid

The adopted CNN comprises two sets of convolutional layers, each succeeded by a batch normalization and a max pooling layer. A fully connected layer of 256 neurons precedes the output layer composed by a single unit with sigmoid activation function. Dropout is added after the fully connected hidden layer, with dropout rate of 0.5. Adam optimizer is employed to minimize the binary cross-entropy loss throughout the training process.

A hold-out validation strategy was considered for assessing the generalization capability of the model. The 20% of the dataset was exploited as test set. The remaining 80% was divided into 80% training and 20% validation set. Notably, patches from the same patient were consistently assigned to the same set throughout the process.

Results for IDC Detection are presented in Table 2, in terms of precision, recall, and f1-score for both classes.

Table 2: Results for IDC detection on the test set.

Class	Precision	Recall	F1-score	Support
NO-IDC	0.90	0.91	0.90	42935
IDC	0.77	0.74	0.75	16216

As expected, the unbalanced dataset makes the identification of the IDC minority class quite challenging. However, the results can overall be con-

sidered reasonable. It must be emphasized that this work does not aim to advance the state-of-art with respect to specific tasks or neural architectures, but rather to show how such elements can be integrated within a user-friendly application to support diagnosis by physician users. The sole purpose of the performance evaluation is therefore to verify that the resulting models are reasonably accurate and thus to ensure that the explainability analysis is valid and meaningful.

## 4.2 XAIMed usage: step-by-step demonstration

Within the XAIMed application, the configuration step must be performed as discussed in Section 3. Figure 3 shows the folder structure for the task considered in this case study.

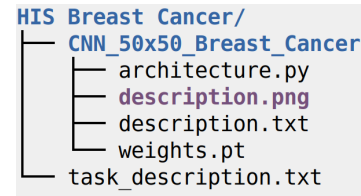


Figure 3: Overview of the task folder configured for the HIS Breast Cancer classification case study.

The top-level folder identifies the task and contains the `task_description.txt` file with a textual description of the task. Furthermore, it also contains a dedicated sub-folder for the model adopted in our case study. The following files are stored therein: the pytorch model weights file (`weights.pt`), the script with the architecture specification for loading the CNN (`architecture.py`), a textual description (`description.txt`) and a visual description (`description.png`) to provide the user with a comprehensive overview of the model specification and performance. When a task is appropriately configured, it becomes available to users of the application through a dedicated button in the GUI.

When the application is started users are greeted with the *Home* screen depicted in Fig. 4.

The interface provides a navigation bar on the left-hand side and a brief description of the application. Furthermore, it includes a “help” feature which offers a guide to the main functionalities of XAIMed. In the following we describe in detail the steps for selecting a task, selecting a model, and analyzing the model predictions and explanations for a given input image.

First, by navigating to the *Task & Model Selection* tab, the user can select a task and an associated model. Figure 5 shows the *Task & Model Selection*



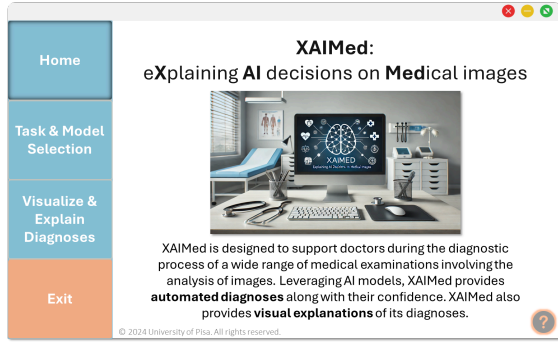


Figure 4: XAIMed *Home* screen. It provides a brief description of the application and access to a user guide. The navigation bar on the left-hand side shows the functionalities offered by the application.

screen.

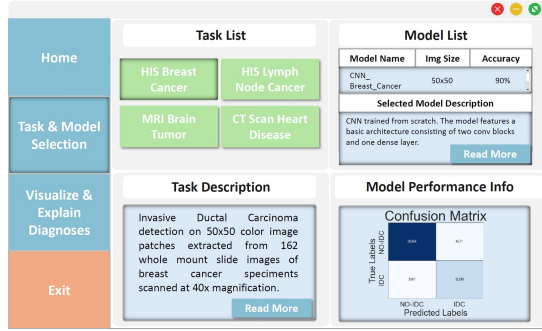


Figure 5: The *Task & Model Selection* tab. It displays the tasks and associated models available in XAIMed.

As depicted in Fig. 5, the screen is divided into four parts. The top-left part shows the available tasks: we select the HIS Breast Cancer as the current case study. The bottom-left part displays the content of the `task_description.txt` file. The top-right part allows selecting a trained model among those available for the task under investigation. The model described in Section 4.1 is selected: the description of the model and its performance measured on a dedicated test set can be found at the bottom-right of the screen. Model performance is shown by means of an image which in this case shows the confusion matrix on the test set. Evidently, if a custom alternative visual representation of the model performance is required, it will be sufficient to replace the `description.png` file with the desired one.

Once a model is selected, the user can access the *Visualize & Explain Diagnoses* tab (Fig. 6).

The user can select the image folder and choose among the images compliant with the selected model. In this case study, we employed a set of images extracted from the test set of the HIS Breast Cancer dataset. The inference process carried out on a se-

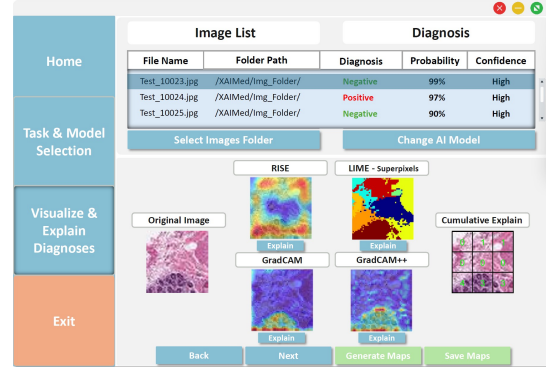


Figure 6: The *Visualize & Explain Diagnoses* tab. It displays saliency-based explanations for a selected image.

lected image provides information about the predicted class along with the confidence value and the discretized confidence level. Most importantly, the SMs can be generated for any selected image, using the four XAI methods described in Section 2. Furthermore, the cumulative saliency-based explanation is obtained as described in Section 3.2: an integer in  $[0, 4]$  indicating the level of agreement among the four methods is superimposed to each of the nine cells in which the input image is partitioned. The original image is displayed alongside the SMs to facilitate comparison.

The *Visualize & Explain Diagnoses* tab allows obtaining additional information about the SMs provided by the four XAI methods. Figure 7 shows the detailed view regarding the GradCAM++ method.

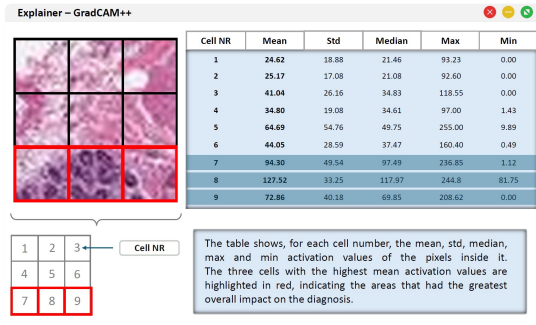


Figure 7: Detailed view of one of the saliency-based explanation methods, namely GradCAM++.

Descriptive statistics are reported for each cell of the grid partitioning and the three ones with the highest mean values are suitably marked on the image to highlight the most important regions for the diagnosis, according to the chosen method.

For the sake of an adequate visualization, we report in Fig. 8 the example original image along with the visual saliency-based explanations provided within XAIMed.



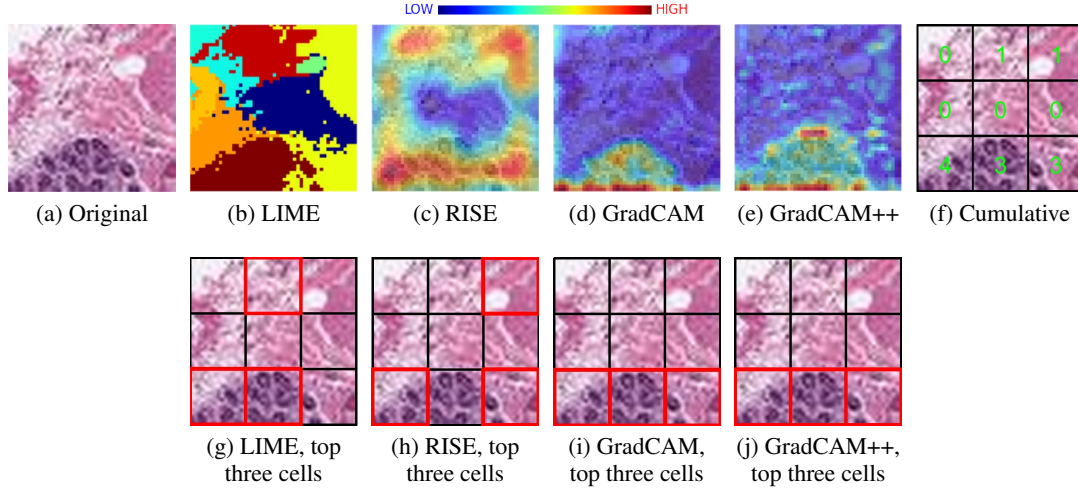


Figure 8: Saliency-based explanations: (a) original image; (b-e) SMs extracted with the four XAI methods; (f) Cumulative SM obtained by aggregating the four explanations; (g-j) grid partitioning of the original image in which the three cells with highest mean saliency value according to the respective XAI method are highlighted.

For each XAI method, the resulting SM is shown in the top row (8b-8e). The bottom row (8g-8j) shows the grid partitioning of the original image in which the three cells with the highest mean saliency value according to the respective XAI method are highlighted. The cumulative SM provides an aggregated view as the count of how many times cells have been highlighted as the most influential for the diagnosis.

In the example, the model classifies the image as negative with high confidence. We observe how the various XAI methods give high importance to the lower region of the image: in the aggregate view, in fact, all lower cells have a count value of 3 or 4.

Finally, the user can opt for persistently storing the generated maps.

The user can obviously switch between different sets of images, different models and different tasks at any time. This allows for a dynamic and thorough exploration of the data according to the user’s needs.

## 5 CONCLUSIONS

In this paper, we introduce XAIMed as a decision support tool designed for computer-aided diagnosis based on medical imaging, within the framework of Explainable AI. The proposed tool provides physician users with diagnoses from a Convolutional Neural Network (CNN) model, suitably trained for a given medical image classification task. Furthermore, four local post-hoc visual explanation methods are implemented within XAIMed, namely RISE, LIME, GradCAM, GradCAM++. Each of the four state-of-art methods produces a Saliency Map (SM) enabling

the identification of the most influential regions for the diagnosis of an input image. A cumulative aggregated SM is computed as the level of agreement among the four methods in order to catalyze physician users’ attention towards macro-regions of the images. XAIMed functionality not only provides diagnostic outcomes but also enhances understanding of the model decision-making process, thereby giving users additional insights to evaluate the accuracy and trustworthiness of the model diagnoses. The application has been implemented in a modular fashion: in the future, we aim to exploit its flexibility to include further visual explanation methods and to refine the explanation aggregation strategy. Another interesting development of the present work is the involvement of domain experts, i.e., physicians, to evaluate the usability and usefulness of XAIMed in clinical practice.

## ACKNOWLEDGEMENTS

This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI” and the PNRR “Tuscany Health Ecosystem” (THE) (Ecosistemi dell’Innovazione) - Spoke 6 - Precision Medicine & Personalized Healthcare (CUP I53C22000780001) under the NextGeneration EU programme, and by the Italian Ministry of University and Research (MUR) in the framework of the FoReLab and CrossLab projects (Departments of Excellence).

## REFERENCES

- Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.
- Arnaout, R., Curran, L., Zhao, Y., Levine, J. C., Chinn, E., and Moon-Grady, A. J. (2021). An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nature Medicine*.
- Barda, A., Horvat, C., and Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making*, 20:257.
- 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 – saliency-based XAI approaches. *European Journal of Radiology*, 162:110787.
- Cerekci, E., Alis, D., Denizoglu, N., Camurdan, O., Ege Seker, M., Ozer, C., Hansu, M. Y., Tanyel, T., Ok-suz, I., and Karaarslan, E. (2024). Quantitative evaluation of saliency-based explainable artificial intelligence (XAI) methods in deep learning-based mammogram analysis. *European Journal of Radiology*, 173:111356.
- Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE.
- Cooper, J., Arandjelović, O., and Harrison, D. J. (2022). Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping. *Pattern Recognition*, 129:108743.
- Cruz-Roa, A., Basavanahally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., and Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, 9041.
- DeGrave, A. J., Janizek, J. D., and Su-In, L. (2021). Ai for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*.
- Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Gian-notti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).
- Hannun, A. Y., Rajpurkar, P., Haghpanshi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69.
- Holzinger, A., Biemann, C., Pattichis, C., and Kell, D. (2017). What do we need to build explainable ai systems for the medical domain?
- Hwang, J., Lee, T., Lee, H., and Byun, S. (2022). A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: User-centered design and evaluation study. *J Med Internet Res*, 24(1):e28659.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.
- Magrabi, F., Ammenwerth, E., McNair, J. B., Keizer, N. F. D., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P. J., Vehko, T., Wong, Z. S., and Georgiou, A. (2019). Artificial intelligence in clinical decision support: Challenges for evaluating ai and practical implications. *Yearbook of Medical Informatics*, 28(1):128–134. Epub 2019 Apr 25.
- Petsiuk, V., Das, A., and Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.
- Raghu, M. and Schmidt, E. (2020). A survey of deep learning for scientific discovery.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.
- Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.
- Scherer, D., Müller, A., and Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer.
- Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. (2016). Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*.
- Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248. PMID: 28301734.
- Shinde, P. P. and Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (IC-CUBE)*, pages 1–6.

- Shortliffe, E. H. and Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, 320(21):2199–2200.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.
- Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of imaging*, 6(6):52.
- Sokolovsky, M., Guerrero, F., Paisarnrisomsuk, S., Ruiz, C., and Alvarez, S. (2018). Human expert-level automated sleep stage prediction and feature discovery by deep convolutional neural networks. In *Proceedings of the 17th International Workshop on Data Mining in Bioinformatics (BIOKDD2018), in Conjunction with the ACM SIGKDD Conference on Knowledge Discovery and Data Mining KDD2018*.
- Solomonides, A., Koski, E., Atabaki, S., Weinberg, S., McGreevey, J., Kannry, J., Petersen, C., and Lehmann, C. (2021). Defining amia’s artificial intelligence principles. *Journal of the American Medical Informatics Association : JAMIA*, 29.
- Tschandl, P., Rosendahl, C., Akay, B. N., Argenziano, G., Blum, A., Braun, R. P., Cabo, H., Gourhant, J.-Y., Kreusch, J., Lallas, A., Lapins, J., Marghoob, A., Menzies, S., Neuber, N. M., Paoli, J., Rabinovitz, H. S., Rinner, C., Scope, A., Soyer, H. P., Sinz, C., Thomas, L., Zalaudek, I., and Kittler, H. (2019). Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA Dermatology*, 155(1):58–65.
- Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470.
- Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629.
- Younes, M. and Hanly, P. (2016). Minimizing inter-rater variability in staging sleep by use of computer-derived features. *Journal of Clinical Sleep Medicine*, 12:1347–1356.