Computer Methods and Programs in Biomedicine 253 (2024) 108238

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

On the evaluation of deep learning interpretability methods for medical images under the scope of faithfulness

Vangelis Lamprou, Athanasios Kallipolitis *, Ilias Maglogiannis

Department of Digital Systems, University of Piraeus, 80, M. Karaoli & A. Dimitriou St, Piraeus 18534, Greece

ARTICLE INFO

Keywords:
Deep learning
Medical images
Attribution maps
Evaluation metrics
Faithfulness

ABSTRACT

Background and Objective: Evaluating the interpretability of Deep Learning models is crucial for building trust and gaining insights into their decision-making processes. In this work, we employ class activation map based attribution methods in a setting where only High-Resolution Class Activation Mapping (HiResCAM) is known to produce faithful explanations. The objective is to evaluate the quality of the attribution maps using quantitative metrics and investigate whether faithfulness aligns with the metrics results.

Methods: We fine-tune pre-trained deep learning architectures over four medical image datasets in order to calculate attribution maps. The maps are evaluated on a threefold metrics basis utilizing well-established evaluation scores.

Results: Our experimental findings suggest that the Area Over Perturbation Curve (AOPC) and Max-Sensitivity scores favor the HiResCAM maps. On the other hand, the Heatmap Assisted Accuracy Score (HAAS) does not provide insights to our comparison as it evaluates almost all maps as inaccurate. To this purpose we further compare our calculated values against values obtained over a diverse group of models which are trained on non-medical benchmark datasets, to eventually achieve more responsive results.

Conclusion: This study develops a series of experiments to discuss the connection between faithfulness and quantitative metrics over medical attribution maps. HiResCAM preserves the gradient effect on a pixel level ultimately producing high-resolution, informative and resilient mappings. In turn, this is depicted in the results of AOPC and Max-Sensitivity metrics, successfully identifying the faithful algorithm. In regards to HAAS, our experiments yield that it is sensitive over complex medical patterns, commonly characterized by strong color dependency and multiple attention areas.

1. Introduction

Convolutional Neural Networks (CNNs) have become increasingly popular in recent years, emerging as a powerful tool that has revolutionized the field of Deep Learning (DL). Their impressive performance results in various domains is attributed to their ability to handle vast amounts of image data and extract meaningful feature representations through a process called *convolution*. This eliminates the need for laborious manual feature engineering, which is typically required in traditional Machine Learning (ML) techniques. However, the success of CNNs often comes at the expense of increased complexity, making it more challenging to comprehend the reasoning behind their decisions. This phenomenon is often referred to as a *trade-off* [1] between the performance of the models and their ability to produce interpretable predictions. Striking a balance between the two is crucial from a human perspective to ensure that the models remain practical and usable.

The ability of a model to generalize effectively is a critical attribute, indicating that the model has learned the underlying patterns and relationships within the data, rather than merely memorizing the training examples. But even if a model produces accurate predictions, is it acceptable for these predictions to be uninterpreted? The lack of transparency in Artificial Intelligence (AI) systems can undermine *trust* and fuel public dispute, particularly in situations where the models have a direct impact on human lives and require meticulous handling. Additionally, it fosters uncertainty and hinders the production of robust and State-of-the-Art (SoA) systems. All this has led to the development of the *eXplainable Artificial Intelligence* (XAI) field during the last decade, which aims at assisting ML engineers create trustworthy, fair and high-performing AI models for real-world applications. To this day, the field has valuable contribution to various domains such as auditing [2], autonomous driving [3], cybersecurity [4], material science and chemistry [5], climate science [6] and, most importantly for our study, medical and healthcare [7–10], with comprehensive literature surveys in each area highlighting its significance.

* Corresponding author.
E-mail address: nasskall@unipi.gr (A. Kallipolitis).

https://doi.org/10.1016/j.cmpb.2024.108238
Received 5 October 2023; Received in revised form 1 May 2024; Accepted 21 May 2024
Available online 28 May 2024
0169-2607/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

In the context of this study, we consider a category of XAI algorithms known as *attribution map* methods and address the task of distinguishing those which can accurately describe the model's viewpoint when making a prediction, formally known as *faithful to the model*, from those which cannot. Taking inspiration from the fact that a human would show preference to the faithful ones, it is natural to investigate whether attribution map metrics will assign them with better evaluation scores as well. We note that establishing a universal mathematical statement for a metric seems rather unlikely and as a result we conduct experimental evaluations instead.

In our experiments, we develop a scheme which quantitatively evaluates Grad-CAM [11] and HiResCAM [12] medical attribution maps, for architectures where only HiResCAM is faithful to. We aim to determine whether metrics such as Area Over Perturbation Curve (AOPC) [13], Max-Sensitivity [14] and Heatmap Assisted Accuracy Score (HAAS) [15] favor the HiResCAM attribution maps over the ones produced by Grad-CAM. Our findings suggest that AOPC and Max-Sensitivity effectively identify the faithful maps, while HAAS does not reflect the ground truth that HiResCAM produces faithful maps and Grad-CAM does not. Overall, this work includes the following contributions:

- We focus on the task of identifying faithful maps in the context of quantitative evaluation metrics. This approach introduces a new dimension to the behavior of a metric.
- We implement the Max-Sensitivity and HAAS metrics in the context of medical attribution maps, previously unused for this domain-specific mappings.
- As a side result of the main task, we detect a sensitive behavior of HAAS towards medical data compared to benchmark non-medical ones.

The rest of the paper is organized as follows: Section 2 refers to the theoretical background related to this study, Section 3 introduces the proposed methodology, Section 4 presents the experimental setup and results, Section 5 focuses on the analysis and discussion of our findings and finally Section 6 includes a summary and final thoughts for future investigation.

2. Background

2.1. The family of CAM-based algorithms

Attribution map methods take as input a test image and a trained CNN classifier and generate class-specific explanations in the form of 2-dimensional attribution maps. Each pixel in the attribution map indicates the model's perception about the importance of the image pixel to the respective class.

We consider a CNN model $f \in \mathcal{F}$ to be a function $f : \mathbb{R}^D \to \mathbb{R}^M$, with input $x \in \mathbb{R}^D$, $D = W * H * 3$, W the width and H the height of an RGB image, and output $f(x) = (f_1(x), f_2(x), \dots, f_M(x)) \in \mathbb{R}^M$ the scores of an M-class problem. Furthermore, an attribution method is a map $\mathcal{A} : \{1, 2, \dots, M\} \times \mathcal{F} \times \mathbb{R}^D \to \mathbb{R}^{W*H}$, $\mathcal{A} = \mathcal{A}(m, f, x) \equiv \mathcal{A}_m(f, x)$, which calculates importance scores of the prediction $f_m(x)$ for features (pixels) of x.

The concept of producing Class Activation Maps (CAM) was initially introduced in [16]. It is applicable to CNNs of the *CAM architecture*, also referred to as *Conv - GAP - Class Scores* CNNs, which comprise networks with a convolutional part followed by Global Average Pooling (GAP) and raw class scores layers. We denote by $\{A^f\}_{f=1,\dots,F}$ the feature maps of the last convolutional layer, where $A^f \in \mathbb{R}^{D_1 * D_2}$, $D_1, D_2 \in \mathbb{N}$, and by $\{w_m^f\}_{f=1,\dots,F}$ the weights that connect the GAP layer with the prediction of class $m \in \{1, 2, \dots, M\}$. In turn, the CAM attribution map of class m is given by:

$$\mathcal{A}_m^{\text{CAM}} = \sum_{f=1}^{F} w_m^f A^f \tag{1}$$

In this context, the weight w_m^f represents an importance score which captures the connection between A^f and the raw score of target class m. Some examples of well-known CNNs of the CAM architecture include ResNet [17], DenseNet [18] and GoogLeNet [19].

However, CAM cannot generate explanations for CNNs containing multi-layer dense classifier or Flatten layer, nor can it provide insights for a convolutional layer other than the last one. In order to overcome these limitations and provide explanations for a wider range of models, the Grad-CAM [11] algorithm was introduced, as an extension of CAM. It is applicable to CNNs with a differentiable network between the final convolutional layer and the prediction layer and enables explanations with respect to any convolutional layer within the network. This is achieved by using the gradient as a way to back-propagate from the calculated score to the pixel of interest. The GAP calculation of the CAM architecture is incorporated into the Gradient Averaging step of Grad-CAM, which is used to weight the feature maps. For this group of networks, we denote by $\{A^f\}_{f=1,\dots,F}$ a convolutional layer (not necessarily the last one), where $A^f \in \mathbb{R}^{D_1 * D_2}$, $D_1, D_2 \in \mathbb{N}$. For class $m \in \{1, 2, \dots, M\}$, the Grad-CAM attribution map with respect to $\{A^f\}_{f=1,\dots,F}$ is given by:

$$\mathcal{A}_m^{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{f=1}^{F} a_m^f A^f\right) \tag{2}$$

where

$$a_m^f = \frac{1}{D_1 D_2} \sum_{i=1}^{D1} \sum_{j=1}^{D_2} \frac{\partial s^m}{\partial A_{ij}^f} \qquad \text{(Gradient Averaging step)} \tag{3}$$

At the same time, although its success is amplified with the remarkable results in the sanity checks of [20], it is also associated with concerns [12,21] about the influence of the Gradient Averaging step on the attribution maps. For this reason, HiResCAM [12] was introduced and, unlike Grad-CAM, it replaces the Gradient Averaging step with the Hadamard product, denoted by $\odot$. This modification introduces the advantage of preserving the gradient effect on a pixel level by utilizing both its value and sign. For class $m \in \{1, 2, \dots, M\}$, the HiResCAM attribution map is given by:

$$\mathcal{A}_m^{\text{HiResCAM}} = \text{ReLU}\left(\sum_{f=1}^{F} \frac{\partial s^m}{\partial A^f} \odot A^f\right) \tag{4}$$

From an application viewpoint, both algorithms enjoy extensive usage in medical-related studies. In regards to Grad-CAM, it has been employed for interpreting predictions of models trained over a variety of anatomical locations and modalities, including chest X-rays [22], chest CT images [23], cardiovascular CT images [24], brain histology images [25] and eye fundus images [26] among others. Similarly, HiResCAM has been used in the context of 3D chest CT volumes [21] and endoscopic images [27]. Notably, in both cases, the implemented models have a structure which is optimal for producing faithful HiResCAM attribution maps, as discussed below.

2.2. Faithfulness of the CAM-based algorithms

We consider the concept of *faithfulness* of a XAI algorithm to a model as a property which *'.. refers to how accurately it (i.e. the algorithm) reflects the true reasoning process of the model..'* (see [28]). In particular to CNN architectures, adopting the definition of [12], an attribution map will be *faithful* to the model if the sum of its pixel values reflects the class score calculation, up to a class-dependent bias term deviation.

For CNNs of the *CAM architecture* one can prove that the CAM algorithm yields faithful attribution maps ([16], section 2) and that Grad-CAM ([11], section 3.1) and HiResCAM ([12], section 3.6.2) collapse to CAM when we calculate the gradients for the last convolutional layer. As a result, Grad-CAM and HiResCAM are equivalent in terms of values and faithfulness. On the other hand, for CNN architectures that utilize a Flatten layer followed by one fully connected layer, denoted

Table 1
Grad-CAM vs HiResCAM - Theory summary.

CNN structure	Gradients w.r.t. last Conv layer	Grad-CAM vs HiResCAM
Conv - GAP - Class Scores	yes	Equivalent & Faithful
	no	Not equivalent & Not faithful
Conv - Flatten - Class Scores	yes	Not equivalent. Only HiResCAM is faithful
	no	Not equivalent & Not faithful
Conv - GAP/Flatten - Dense - Class Scores	yes	Not equivalent & Not faithful
	no	Not equivalent & Not faithful

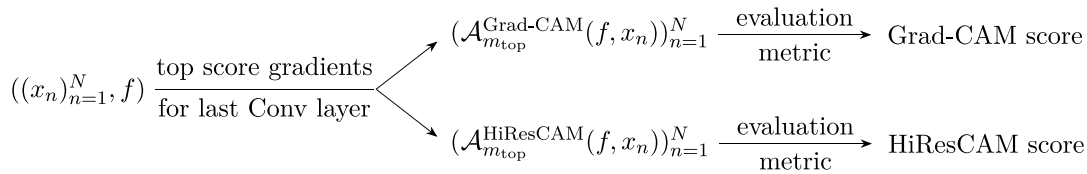

Fig. 1. Workflow of proposed methodology.

by *Conv - Flatten - Class Scores*, Grad-CAM and HiResCAM are no longer extensions of CAM and they always have different values. On the faithfulness aspect though, in [12] section 3.5.1 the authors prove that HiResCAM maps become faithful when the gradients are calculated for the last convolutional layer. Finally, in any other scenario we note that the algorithms are neither equivalent nor can produce faithful maps. The conclusions of section 2.2 are summarized in Table 1, providing essential insights on the role of faithfulness in our study.

3. Methodology

3.1. Workflow

Upon reviewing Table 1 it becomes evident that the combination of non-equivalent values and the faithfulness of HiResCAM turns the setting of *Conv - Flatten - Class Scores* structures with gradients calculated at the last convolutional layer into a *reference setting* where the algorithms are distinguishable. In the remaining ones, both algorithms are either equivalent or non-faithful. Consequently, this particular setting serves as a compelling ground for hosting a meaningful comparison between Grad-CAM and HiResCAM.

This is realized upon the methodology proposed in Fig. 1. Given a test dataset $(x_n)_{n=1}^N$ and a trained CNN model f adjusted to the reference setting, we first compute Grad-CAM and HiResCAM attribution maps for each image for the top predicted class m_{top}, eventually yielding the maps sequences $(\mathcal{A}_{m_{top}}^{Grad-CAM}(f, x_n))_{n=1}^N$ and $(\mathcal{A}_{m_{top}}^{HiResCAM}(f, x_n))_{n=1}^N$. In turn, for each method we quantitatively evaluate the quality of its maps on a threefold metrics scheme. This allows us to investigate faithfulness and collect insights with respect to each metric's separate perspective. At the same time, it collectively forms a well-rounded view of the topic and potentially increases the consistency and objectivity of our conclusions in case the calculated results of two or more metrics align over the series of experiments.

3.2. Evaluation framework

Explanations serve as a powerful post-test analysis tool because they offer a tangible representation of what a model has learned during training and establish a connection between the inner workings of the model and the real world. However, the challenge of choosing between attribution maps generated either from the same model via different XAI algorithms or from different models via the same XAI algorithm is common among practical applications and suggests the need to consider evaluation tools as well.

In existing literature, there are various studies which fall within the former case and although there is no direct reference to faithfulness, their results could potentially be analyzed in the context of our framework as well. Notable examples, such as [29–33], employ evaluation metrics, different from those chosen for this work, on attribution maps where Grad-CAM could be considered faithful. In view of this study metrics, one may initially turn to [34] for an AOPC implementation on satellite images. However, the absence of an open-source implementation prevents us from drawing definitive conclusions about the maps' faithfulness and incorporate the experiments to our perspective.

On top of that, we make special reference to [35], as its results can be analyzed for our purposes, introducing a meaningful link with our work. The authors train DenseNet models of the *CAM architecture* on earth image data coming from the BigEarthNet [36] and SEN12MS [37] datasets and, among others, evaluate Grad-CAM attributions with respect to the last convolutional layer by employing Max Sensitivity and Area Under the Most Revelant First (MoRF) curve, a self-explanatory variation of AOPC. In this setting Grad-CAM produces faithful attribution maps (recall Table 1) while the rest algorithms do not. Upon reviewing the results of Table 2 in [35], we first observe that Grad-CAM exhibits optimal performance across both datasets concerning Max-Sensitivity. Furthermore, it achieves optimal performance for the SEN12MS dataset and ranks third for the BigEarthNet dataset with regard to the Area Under MoRF curve.

Our evaluation utilizes the metrics AOPC, Max Sensitivity and HAAS, to objectively measure the quality of medical attribution maps. For this kind of mappings, AOPC has been used in various applications [38–40], while Max Sensitivity and HAAS remain unexploited.

4. Experimental setup

4.1. Datasets

For the purposes of our experiments we employ four well-known medical image datasets. Firstly, the *CRC* dataset [41], which consists of 107,180 images and has been widely used to study the patterns of colorectal cancer (CRC). Then we employ the *Covid-19 Radiography Database* [42,43], consisting of 21,165 X-ray images, created to familiarize the ML community with the Covid-19 disease and distinguish it from other common chest-related diseases. The third dataset is the *HAM10000* dataset [44], comprising a collection of 10,015 dermatoscopic images extracted from 7,470 distinct pigmented skin lesions. Finally, we utilize *BreakHis* dataset [45], divided into 2,480 benign and 5,429 malignant microscopic images of breast tumor tissues. They are captured at different magnification levels, including 40X, 100X, 200X and 400X, and organized in image groups referenced as slides.

4.2. Evaluation metrics

For the rest of this study, in the interest of notation simplicity, the index m in an attribution map $\mathcal{A}_m = \mathcal{A}_m(f, x)$ shall be removed and we will always consider the top predicted class.

Firstly, we discuss the AOPC metric [13]. Given an attribution map $\mathcal{A}(f, x)$ we consider a heatmap H to be a region-level transformation of $\mathcal{A}(f, x)$ via Average Pooling. The metric was introduced in the context of heatmap evaluation and developed upon the notion of the MoRF curve, which keeps track of the decrease in the class probability when we iteratively perturb the most important image regions, as dictated by H. A large area over the MoRF curve corresponds to early and steep decreases in the curve, indicating a heatmap that effectively captures the class information within a few regions.

From a mathematical perspective, the class probability at the k-th perturbation step is $f(x^{(k)}_{MoRF})$, where $x^{(k)}_{MoRF}$ is a version of the image x perturbed at the top k important regions. After L perturbation steps the AOPC score of the image x can be estimated by the quantity:

$$\frac{1}{L+1}\sum_{k=1}^{L}[f(x^{(0)}_{MoRF}) - f(x^{(k)}_{MoRF})] \tag{5}$$

Secondly, the Max-Sensitivity metric [14] was developed in order to quantify the extent to which the class explanation is influenced by minor perturbations in the test point. For an image x, the metric score is calculated over a variety of perturbed images y, which are sampled from a ball with center x and radius r, as the maximum distance between the class explanations of the original image $\mathcal{A}(f, x)$ and the perturbed image $\mathcal{A}(f, y)$.

$$SENS_{MAX}(\mathcal{A}, f, x, r) = \max_{y:\|x-y\|_\infty \le r} \|\mathcal{A}(f, x) - \mathcal{A}(f, y)\|_F \tag{6}$$

We denote by $\|\cdot\|_\infty$ the maximum norm and by $\|\cdot\|_F$ the Euclidean (or Frobenius) norm. In practise, formula (6) can be estimated by Monte-Carlo sampling, as instructed by the authors in the source code [46] and, similar to AOPC, the calculation is extended on a dataset level by taking the average over all image scores. Notably, it is natural to prefer explanations with low Max-Sensitivity, as this suggests that the explanability algorithm is resilient to small variations in the input point and can generate similar explanations for similar inputs.

Lastly, we consider the HAAS score [15]. As the name suggests, it relies on the idea that if an attribution map $\mathcal{A}(f, x)$ (referenced as Heatmap in the original publication) provides an accurate explanation of the model's viewpoint, then it should be able to adjust the image pixels in such a way that the model's overall predictive capability is improved. This is realized, by the HA (Heatmap Assisted) image, given by:

$$HA(x, \mathcal{A}(f, x)) = \max\{-1, \min\{1, x_{norm}(1 + \mathcal{A}(f, x)_{norm})\}\} \tag{7}$$

for image x_{norm} and attribution pixels $\mathcal{A}(f, x)_{norm}$ normalized to the $[-1, 1]$ range. In turn, over a test dataset, the total change in the model performance is captured in the HAAS score, given by:

$$HAAS = \frac{\text{Accuracy over HA images}}{\text{Accuracy over original images}} \tag{8}$$

We observe that if the HAAS value is greater than 1 then the attribution maps effectively explain the importance of the features, while if the value is below 1 they do not convey this information effectively.

4.3. Training strategy

In order to build models we utilized a 10-fold cross validation scheme. Each fold was sampled based on a stratified approach with respect to the dataset class population variable, ensuring that the class fold percentage is the same as the class total population percentage. In addition, for HAM10000 and BreakHis, we adopted a skin lesions and breast tissue train–validation–test set *independent* approach respectively, so that the models' performance is not influenced by variations of the training images and the final models are not subject to additional bias. Finally, we resized all images to the pre-trained models recommended input size, $224 * 224$ pixels, normalized their values to $[-1, 1]$ and augmented the training data via vertical and horizontal flips and random rotations.

The pre-trained architectures were imported from PyTorch's *torchvision* class and transformed to the *Conv - Flatten - Class Scores* structure (see Fig. 2) first by removing the following layers from the end of the convolutional part of each architecture: in VGG19 the ReLU and MaxPool2d layers, in ResNet34 the BatchNorm2d layer and in ResNet50 the BatchNorm2d and ReLU layers, and secondly by replacing the GAP layer with a Flatten in all networks. We trained our models with a *weighted* Cross Entropy loss function, to tackle class imbalance issues, and applied an Adam optimizer. At the same time, as described in Table 2, we treated multiple training hyper-parameters in a model-specific way. The Early Stopping technique is custom written and monitors the mean validation recall of the classes of interest via the *patience* parameter (last row of Table 2). In our experiments it was adapted to the needs of medical-related models by monitoring the disease related classes in all four datasets. Last but not least, the convolutional *trainable layers* for each model are described in the first row of the table via their *torchvision* layer name.

4.4. Experimental results

The predictive power of the cross-validation models is described in Table 3, in terms of the Balanced Accuracy and Area Under Curve (AUC) metrics. For each dataset, model type and metric, we report the sample Mean, Standard Deviation (Std) and 95% Confidence Interval for the Mean (CI), derived under the assumption that the calculated values follow Gaussian distribution. We notice that CRC and Covid-19 Database produce robust models. In terms of Balanced Accuracy, HAM10000's severe class imbalance has a visible impact while BreakHis stands in the middle between HAM10000 and the other two datasets.

In accordance to 3.1, the subsequent steps introduce the interpretability parts of the workflow. Due to time-demanding calculations, we have considered the cross validation model whose Balanced Accuracy is closest to the Mean Balanced Accuracy, as per Table 3. Notably the Grad-CAM and HiResCAM attribution maps are implemented via the *pytorch-grad-cam* class [47], which is compatible with PyTorch models, and their gradients are calculated with respect to the following layers: *layer4[2].conv2* for ResNet34, *layer4[2].conv3* for ResNet50 and *features[-1]* for VGG19.

Employing AOPC score as a post-test metric can prove to be a valuable tool for gaining insights into how well a trained model has learned the classes of interest. Yet, its calculation is coupled with several critical decisions, as the creation of heatmap H can vary in terms of size and the manipulation of image pixels can be accomplished in many ways. This includes applying different functions to introduce random or constant noise and employing diverse approaches, such as replacing entire pixels or making subtle modifications to them. The findings of our experiments are summarized in Table 4. For each dataset and model, aiming to cover a variety of cases, we considered AOPC scores for Grad-CAM and HiResCAM heatmaps of decreasing sizes $4 * 4$, $8 * 8$, $11 * 11$ and $14 * 14$, as shown in the table rows. The calculated scores are based on the correctly classified images, with pixels being replaced with re-sampled uniform noise at each perturbation step. In order to compare the attribution methods, we recall that larger AOPC values (highlighted in the table) indicate heatmaps of higher quality. For instance, considering the scores for Covid-19 dataset, ResNet34 model and $4 * 4$ heatmap size, we observe that HiResCAM produces *better* maps from the AOPC objective as it scores 0.73, while Grad-CAM scores 0.69.

In regards to Max-Sensitivity score, its implementation is also subject to a set of hyper-parameters, including the level r of uniform noise

Table 2
Training configurations.

	CRC		Covid-19		HAM10000		BreakHis	
	ResNet34	VGG19	ResNet34	VGG19	ResNet50	VGG19	ResNet50	VGG19
Trainable layers	layer3, layer4	feature layers 19 to 34	layer1 to layer4	feature layers 21 to 34	layer4	feature layers 12 to 34	None	None
Learning rate	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-5}	10^{-4}	10^{-4}	10^{-4}
Weight decay (L2)	None	None	None	None	10^{-1}	10^{-1}	None	None
Batch size	256	256	64	128	256	128	16	16
Patience	7	7	20	20	20	20	7	7

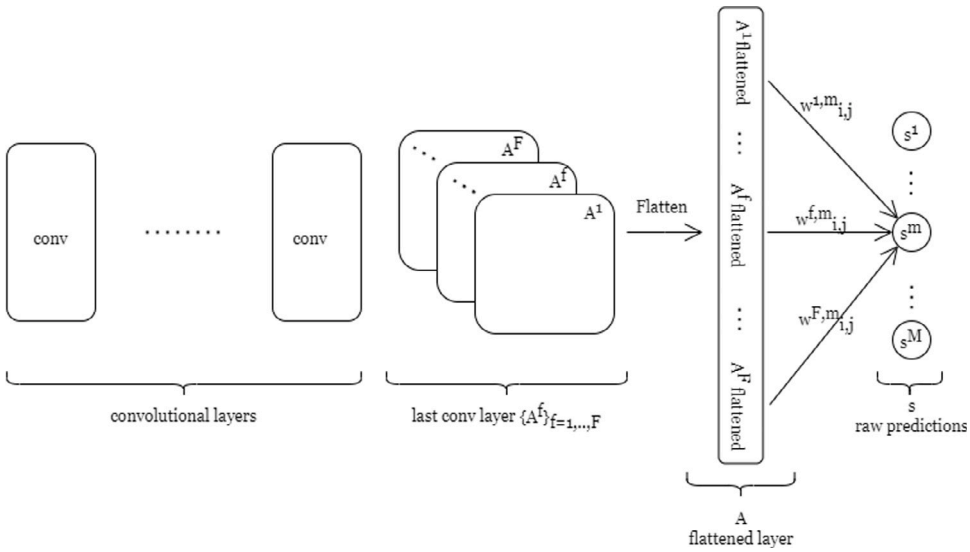

Fig. 2. The *Conv - Flatten - Class Scores* architecture.

Table 3
Testing results.

		Bal accuracy			AUC		
		Mean	Std	95% CI	Mean	Std	95% CI
CRC	ResNet34	0.889	0.009	[0.883, 0.895]	0.993	0.003	[0.991, 0.995]
	VGG19	0.94	0.01	[0.934, 0.946]	0.992	0.005	[0.989, 0.995]
Covid-19	ResNet34	0.956	0.006	[0.952, 0.96]	0.987	0.002	[0.984, 0.986]
	VGG19	0.947	0.008	[0.942, 0.952]	0.99	0.003	[0.988, 0.992]
HAM10000	ResNet50	0.677	0.025	[0.662, 0.692]	0.922	0.013	[0.914, 0.93]
	VGG19	0.72	0.023	[0.706, 0.734]	0.928	0.014	[0.919, 0.937]
BreakHis	ResNet50	0.859	0.017	[0.849, 0.869]	0.943	0.013	[0.935, 0.951]
	VGG19	0.84	0.015	[0.831, 0.849]	0.919	0.019	[0.907, 0.931]

Table 4
AOPC scores.

			CRC		Covid-19		HAM10000		BreakHis	
			ResNet34	VGG19	ResNet34	VGG19	ResNet50	VGG19	ResNet50	VGG19
Heatmap size	4 × 4	Grad-CAM	0.57	0.51	0.69	0.63	0.58	0.33	0.31	0.34
		HiResCAM	0.57	**0.52**	**0.73**	**0.65**	**0.6**	**0.36**	**0.36**	**0.4**
	8 × 8	Grad-CAM	0.58	0.55	0.72	**0.71**	0.64	0.36	0.27	0.35
		HiResCAM	**0.59**	**0.59**	0.72	0.69	**0.65**	**0.38**	**0.29**	**0.45**
	11 × 11	Grad-CAM	0.6	0.6	0.73	0.71	0.65	0.37	0.26	0.35
		HiResCAM	0.6	**0.63**	**0.75**	**0.72**	**0.69**	0.37	**0.29**	**0.43**
	14 × 14	Grad-CAM	0.59	0.59	0.72	**0.7**	0.68	0.36	0.27	0.37
		HiResCAM	**0.61**	**0.62**	**0.74**	0.68	0.68	0.36	**0.3**	**0.48**

perturbation and the number of perturbed instances y. As the radius value increases, it is natural to consider increased number of samples as well. Consequently, in our experiments, based on available hardware and time complexity criteria, for each correctly classified image, we drew 20 samples for radius values 0.05 and 0.1, 30 samples for radius values 0.2 and 0.3 and 40 samples for radius values 0.4 and 0.5. Our

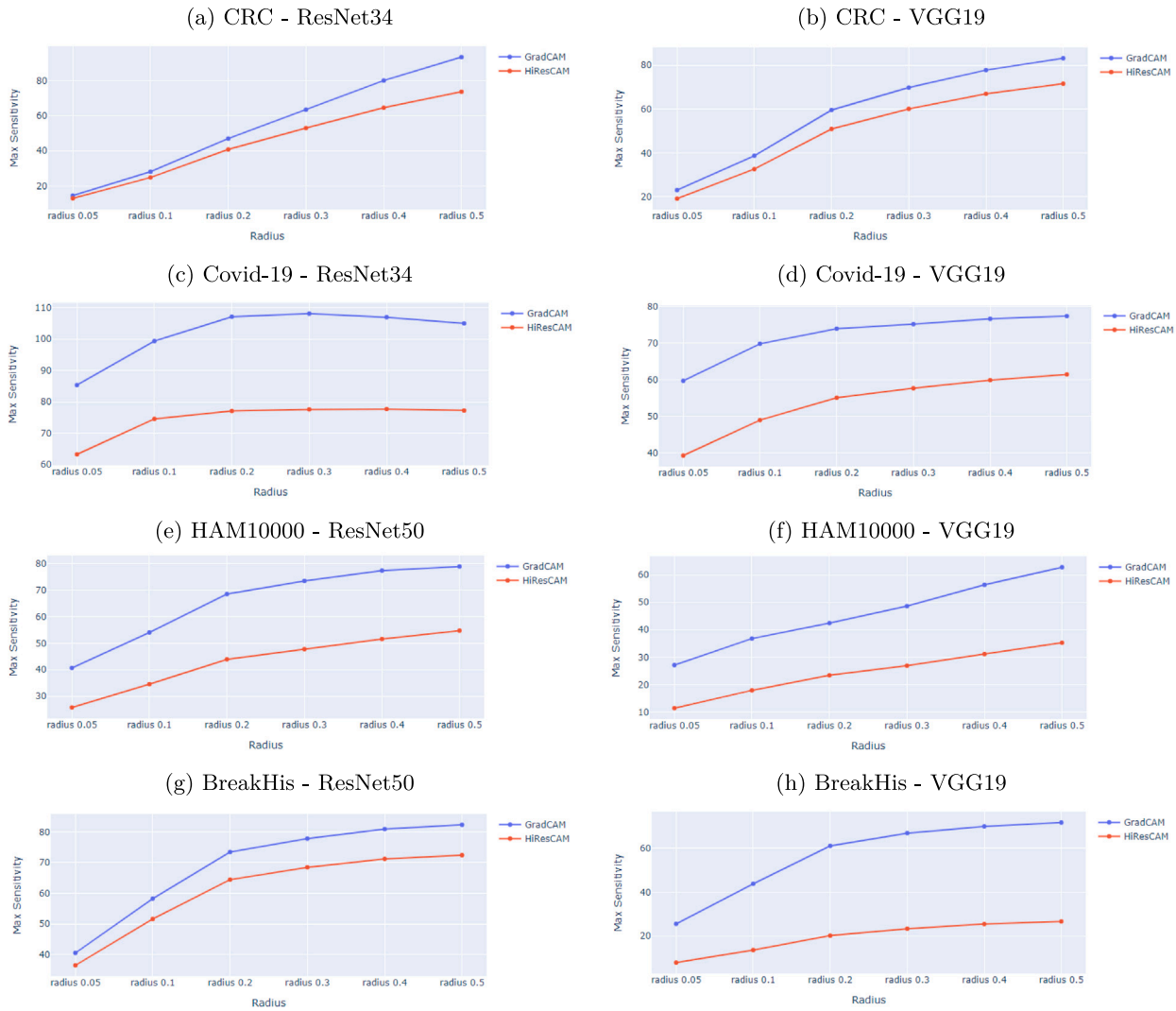

Fig. 3. Max-Sensitivity plots.

Table 5
HAAS scores (Medical datasets).

	CRC		Covid-19		HAM10000		BreakHis	
	ResNet34	VGG19	ResNet34	VGG19	ResNet50	VGG19	ResNet50	VGG19
Grad-CAM	0.5	0.76	**0.84**	**0.85**	0.8	0.78	0.94	0.91
HiResCAM	**0.54**	**0.79**	0.72	0.83	**0.82**	**0.88**	**0.95**	**1.07**

results are plotted in Fig. 3, where each plot refers to a different dataset-model pair and contains the Grad-CAM (blue color) and HiResCAM (red color) Max-Sensitivity scores for increasing radius values along the x-axis. We recall that lower Max-Sensitivity values indicate heatmaps of *better* quality.

Finally, the results of the HAAS score are summarized in Table 5, recalling that a method which produces accurate maps should be able to reach a score above 1. Notably, the metric is free of hyper-parameters, making it an appealing choice from an applied perspective as the calculation of the HA image is fast to implement. Given a medical dataset and a trained model, the HAAS score for Grad-CAM (first row) yields values below 1 in all experiments and for HiResCAM (second row) values below 1 in almost all experiments; the only exception being the maps produced by the VGG19 model over the BreakHis dataset, managing to improve the overall accuracy by 7%. Upon comparing the methods' values column-wise, the higher score is highlighted in the table.

This unexpected behavior suggests that the metric does not value the quality of the explanations and eliminates the contribution of HAAS in the comparison between Grad-CAM and HiResCAM. Consequently, it leads us to investigate its cause and if it is specifically related to the nature of the medical datasets. To this purpose, we employed the benchmark datasets of the HAAS publication: Cifar-10 [48], STL-10 [49] and Imagenette [50] (small version of ImageNet [51] with 10 classes). Per dataset, we fit a loop of 16 VGG19 architectures, over a variety of hyper-parameter configurations, in order to track the range of HAAS values and compare with the medical HAAS scores. Our findings on the HiResCAM maps are documented in Table 6, containing the maximum and minimum HAAS values obtained per loop, accompanied with the respective mean AUC score of the model. In our experiments, we note that we customized the Cifar-10 and STL-10 architectures to match the image input size, as suggested in Fig. 4 of [15], in order to prevent the image from collapsing into a single number.

Table 6
HAAS scores (Non-Medical datasets).

		Cifar-10 VGG19 loop	STL-10 VGG19 loop	Imagenette VGG19 loop
HiResCAM	Max HAAS Score	1.009	1.034	1.002
	Mean AUC	0.981	0.966	0.995
	Min HAAS Score	0.970	0.978	0.986
	Mean AUC	0.969	0.899	0.889

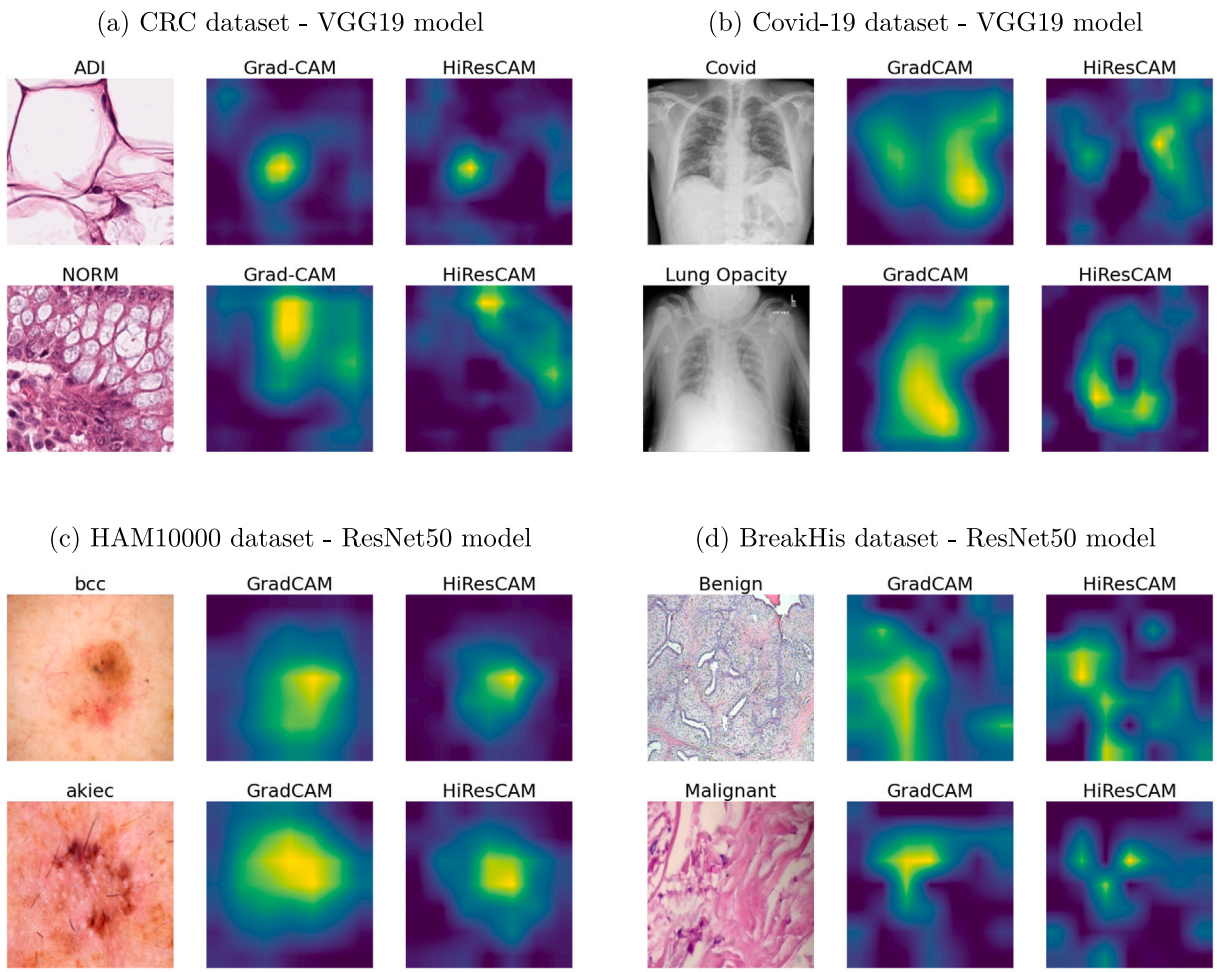

Fig. 4. Examples of attribution maps. *Grad-CAM produces maps with large and smooth-boundary attention areas while HiResCAM high-resolution maps. The images come from the following dataset classes: (a) ADI:Adipose and NORM:Normal colon mucosa, (b) Covid and Lung opacity, (c) bcc:Basal cell carcinoma and akiec:Actinic keratoses, (d) Benign and Malignant.*

As far as the Cifar-10 dataset is concerned, the implemented model loop consisted of all possible combinations of the following hyper-parameters, resulting in 16 different model configurations in total: batch size $32, 64, 128$ or 256, learning rate 10^{-4} or 10^{-6} and weight decay 10^{-2} or *None*. Similarly, for the STL-10 dataset: batch size 32 or 128, learning rate 10^{-4} or 10^{-6}, scheduler with step $= 2$ and $\gamma = 0.5$ or *None* and weight decay 10^{-1} or *None*. Finally, for Imagenette, batch size $32, 64, 128$ or 256, learning rate 10^{-4} or 10^{-6} and scheduler with step $= 2$ and $\gamma = 0.5$ or *None*.

5. Discussion

First, we discuss the results of AOPC and Max-Sensitivity presented in Table 4 and Fig. 3 of section 4.4 respectively. We address them together since their analysis is based on the same reasoning. Regarding AOPC, we notice that large sized heatmaps $4 * 4$ tend to favor HiResCAM in 7 out of 8 model cases. As the size decreases, the effect gradually weakens. Specifically, heatmaps of size $8 * 8$ and $11 * 11$ side with HiResCAM in 6 out of 8 cases, while heatmaps of size $14 * 14$ side with HiResCAM in 5 out of 8 cases. In addition, considering all heatmap sizes for each tested model, we see that the metric favors HiResCAM in 7 out of 8 model experiments, excluding only VGG19 for Covid-19 Database. At the same time, when it comes to Max-Sensitivity, it is apparent that HiResCAM scores lower in all 8 experiments across all tested models and radii.

In order to gain insights into the factors behind the favorable results of HiResCAM, we revisit the definitions of Grad-CAM and HiResCAM, as outlined in section 2.1. A notable distinction between the algorithms lies in the treatment of gradients. Grad-CAM calculates Gradient Averages, as described in formula (3), to assign weights to the feature maps. Conversely, HiResCAM employs the Hadamard product to weigh each pixel of a feature map with its corresponding gradient. This distinctive approach of HiResCAM enables direct contributions to the calculated class score while preserving the gradient's influence on a pixel level, encompassing both its value and sign. As a result, HiResCAM is able to generate fine-grained high resolution attribution

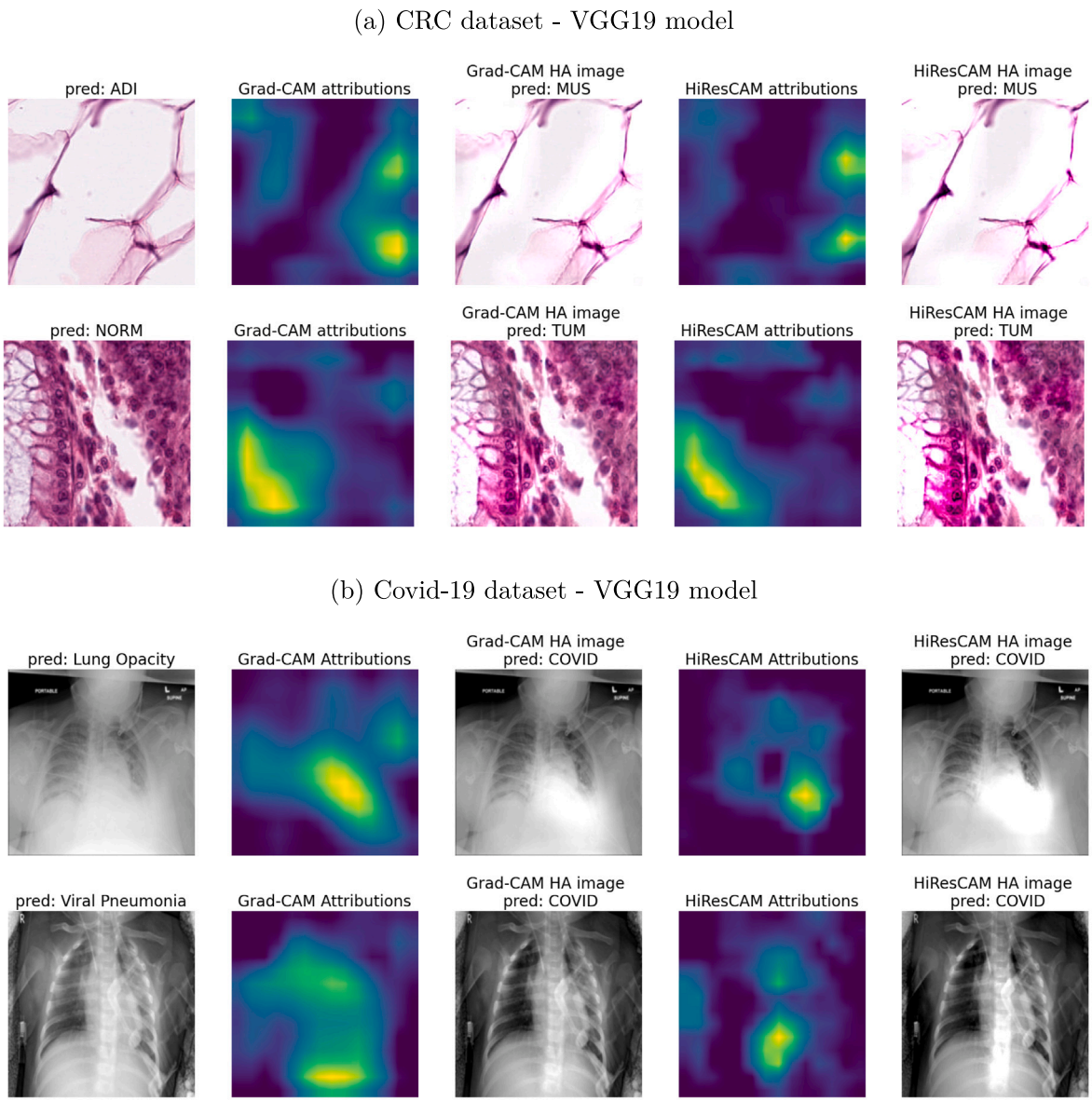

Fig. 5A. Examples of misclassified HA images for CRC and Covid-19 datasets. *We refer to the following classes: (a) ADI:Adipose, MUS:Smooth muscle, NORM:Normal colon mucosa and TUM:Colorectal adenocarcinoma epithelium (b) Lung opacity, Covid and Viral pneumonia.*

maps, while Grad-CAM produces maps characterized by larger and smoother areas of interest because of the Gradient Averaging step effect. This visual dissimilarity becomes explicit in Fig. 4. While both techniques identify similar attention areas in the medical images, the Grad-CAM explanations take up a larger region of the image surrounded by smoother boundaries. Consequently, the high-resolution HiResCAM attribution maps have the potential to capture finer details of the image and achieve more accurate localization of the salient regions. This is successfully depicted in the calculated AOPC scores. Likewise, they demonstrate greater resilience to minor perturbations in the input image and yield lower Max-Sensitivity results.

In regards to HAAS score, the results over the medical datasets are summarized in Table 5 of section 4.4 and in almost all cases, 15 out of 16 experiments, we calculated HAAS value below 1. This raises concerns about the quality of the attribution maps but also inspires further analysis to determine the reasons behind these results.

We initially observe that HAAS was tested in [15] on datasets, such as Cifar-10, STL-10, and Imagenette, whose classes are determined by single objects. They include images labeled as airplanes, cars, birds, cats, horses, garbage trucks, golf balls, cassette players etc. that can primarily be recognized based on their distinctive shapes, with less emphasis on subtle color variations. On the other hand, medical images are more complex. The classes could have many attention areas and a stronger color dependency. Consequently, in the context of the positive attributions generated by Grad-CAM and HiResCAM, emphasizing the image pixels intensity might not necessarily assist the model to locate a learned pattern; it could potentially confuse it instead. In Figs. 5A and 5B we refer to representative cases of this issue. For each dataset, we consider correctly classified images whose HA image variation presents visible color tampering in the class attention area and is misclassified. Among others, the most typical examples include the NORM and Lung Opacity images, in (a) and (b) subplots of Fig. 5A, where the loss of information between the original and the HA image in the region of interest is rather noticeable.

In an attempt to investigate the relationship between HAAS and medical data we utilized the benchmark datasets of the HAAS publication for comparison purposes. As explained in section 4.3, per dataset

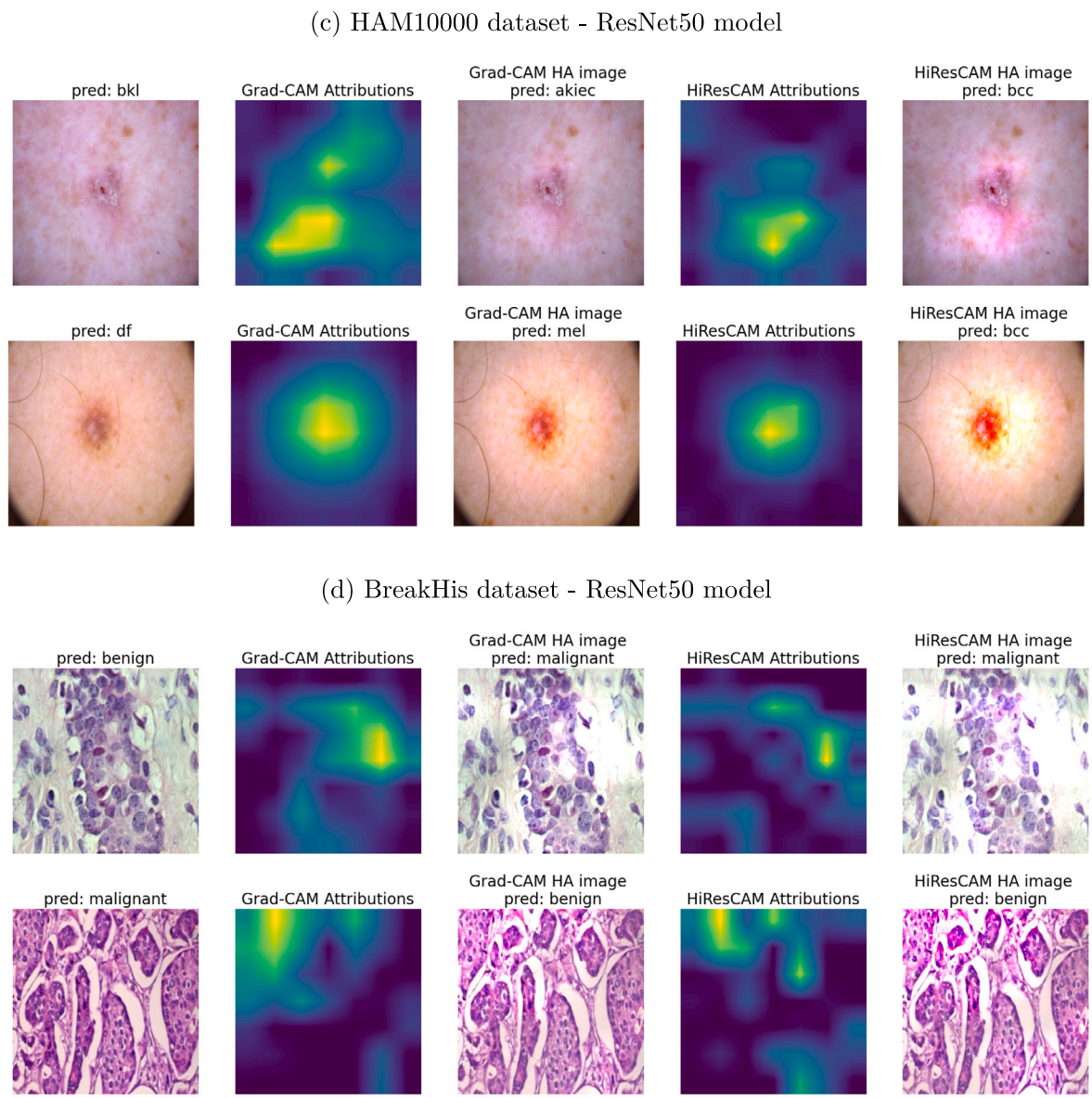

Fig. 5B. Examples of misclassified HA images for HAM10000 and BreakHis datasets. *We refer to the following classes: (c) bkl:Benign keratosis-like lesions, akiec:Actinic keratoses, bcc:Basal cell carcinoma, df:Dermatofibroma and mel:Melanoma (d) Benign and Malignant.*

we built a loop of 16 VGG19 models, configured over a variety of hyper-parameter values, in order to track the HiResCAM HAAS scores range. Following the results of Table 6, each loop yields maximum HAAS score above 1 for a highly performing model and minimum HAAS score slightly below 1 for a well performing model. In other words, it was possible to derive meaningful HAAS scores when utilizing the non-medical datasets and models that were not optimally trained. This stands in contrast to our study involving medical datasets and models that underwent more meticulous and extensive training and suggests evidence that HAAS may exhibit greater sensitivity to medical data.

Further to the above, one might also consider the difference in pixel values between the original and the HA image in general, as described in formula (7). For instance, under a Grad-CAM or HiResCAM attribution value of $1/2$, an 1-dimensional pixel coordinate with value $1/2$ in the original image is transformed into $3/4$ in the HA image and similarly a coordinate with value $-1/2$ into $-3/4$. As a result, when evaluating the model on a HA image, it is possible that its values *deviate considerably* from the distribution on which the model was trained. While this observation is not thoroughly investigated in our study, it could serve as a motivation for future research and clarifications. It may affect any dataset and model, as even the calculations over non-medical data revealed models, such as the ones with minimum HAAS score in Table 6, with very good performance but a HAAS value strictly smaller than 1.

Last but not least, inherent to any scientific study, we address the limitations that one needs to take into account upon reading our work. Most significantly, it is important to consider that the nature of our posed question makes it unlikely to establish a general mathematical formulation that encapsulates the behavior of a specific model architecture or faithful algorithm with respect to some metric. This constraint necessitates case-by-case experimentation. Indeed, from a calculation perspective, denoting by f a CNN model, $\mathcal{A}_m^{\text{faithful}}$ a faithful algorithm (*i.e.* for class m, there exists $b^m \in \mathbb{R}$ such that $s^m = \sum_{i,j}(\mathcal{A}_m^{\text{faithful}})_{i,j} + b^m$), $\mathcal{A}_m^{\text{non-faithful}}$ a non-faithful algorithm (*i.e.* for class m, there does not exist $b^m \in \mathbb{R}$ such that $s^m = \sum_{i,j}(\mathcal{A}_m^{\text{non-faithful}})_{i,j} + b^m$) and μ a metric such as AOPC of formula (5), Max-Sensitivity of formula (6) or HAAS of formula (7), then it does not necessarily follow that $\mu(f, \mathcal{A}_m^{\text{faithful}})$ will get *better* value than $\mu(f, \mathcal{A}_m^{\text{non-faithful}})$, either '$\leq$' or '$\geq$' depending on the metric. Furthermore, as introduced in section 3.2, the availability of implementations that can be tailored to our specific objectives and

chosen metrics is notably limited, making it challenging to compare our work with others and draw more robust conclusions. In the absence of suitable counterparts, our study serves as a pioneering exploration that lays the groundwork for future investigations in this area. Finally, from an experimental perspective, it is crucial to emphasize the issue of time-consuming calculations, all conducted using a V100 GPU machine. This limitation impedes our ability to repeat the entire workflow multiple times and draw conclusions based on statistical analysis tests, for AOPC scores (Table 4), Max-Sensitivity plots (Fig. 3) and HAAS scores (Table 5). To compensate for this limitation, a repetitive process for the calculation of the respective evaluation metrics (AOPC, Max-Sensitivity) with different heatmap and radius sizes is deployed. The demonstrated consistency throughout the experiments increase the validity of the results, calculated for the most representative model in terms of Balanced Accuracy as discussed in section 4.4.

6. Conclusions

In this study we investigate the relationship between faithfulness and quantitative evaluation metrics in the context of medical attribution maps. In order to produce faithful HiResCAM maps that can decompose the predicted class score we consider the setting proposed in [12], employing *Conv - Flatten - Class Scores* models and HiResCAM score gradients with respect to the last convolutional layer of the model. Notably, in this setting Grad-CAM cannot produce faithful maps. Our implementation fine-tunes ResNet and VGG pre-trained architectures and evaluates the generated maps via the AOPC, Max Sensitivity and HAAS metrics with the objective to investigate if the metrics can successfully distinguish between faithful and non-faithful attribution maps.

Our findings indicate that Max-Sensitivity and AOPC exhibit a strong preference for HiResCAM attribution maps, as they effectively differentiate between salient and non-salient regions of the image, ultimately yielding more informative and resilient explanations. This is greatly attributed to the existence of the Hadamard product, which allows HiResCAM to exploit the gradient effect on a pixel level and produce high resolution mappings. In contrast, Grad-CAM's Gradient Averaging step camouflages the gradient information into feature map averages, yielding larger and less informative attention areas. On the other hand, the HAAS score does not contribute to our objective since it evaluates almost all attribution maps as inaccurate. In turn, we explore the dependency of HA images on class features that could potentially vary between the medical datasets of our study and the non-medical ones originally tested in [15]. To this purpose, for each non-medical dataset we trained a *diverse* group of models and compared the range of HAAS values to the medical ones. We conclude that HAAS is more sensitive when it comes to complex X-ray and digital pathology patterns, characterized by stronger color dependency and many attention areas, in comparison to the single object classes with weak color dependency present in the benchmark datasets.

To sum up, this work is an early step towards developing a more comprehensive understanding of the interplay between quantitative evaluation metrics and faithfulness. Our conclusion aligns well with the calculations of [35], further reinforcing the observation that AOPC and Max Sensitivity can identify a faithful algorithm. Future research may expand the experimental setting by considering alternative pairs of faithful and non-faithful algorithms, introducing different metrics, new types of datasets and more pre-trained architectures. Ultimately, this could lead to a more robust and universal view of the dependencies and limitations of this area.

Supplementary material

The source code of our experiments can be found at: https://github.com/vggls/medical_xai

CRediT authorship contribution statement

Vangelis Lamprou: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Athanasios Kallipolitis:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Ilias Maglogiannis:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

References

[1] Plamen P Angelov, et al., Explainable artificial intelligence: an analytical review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 11 (5) (2021) e1424.
[2] Chanyuan Abigail Zhang, Soohyun Cho, Miklos Vasarhelyi, Explainable artificial intelligence (xai) in auditing, Int. J. Account. Inf. Syst. 46 (2022) 100572.
[3] Daniel Omeiza, et al., Explanations in autonomous driving: A survey, IEEE Trans. Intell. Transp. Syst. 23 (8) (2021) 10142–10162.
[4] Nicola Capuano, et al., Explainable artificial intelligence in cybersecurity: A survey, IEEE Access 10 (2022) 93575–93600.
[5] Felipe Oviedo, et al., Interpretable and explainable machine learning for materials science and chemistry, Acc. Mater. Res. 3 (6) (2022) 597–607.
[6] Hakan Başağaoğlu, et al., A review on interpretable and explainable artificial intelligence in hydroclimatic applications, Water 14 (8) (2022) 1230.
[7] Bas HM Van der Velden, et al., Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.
[8] Johannes Allgaier, et al., How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare, Artif. Intell. Med. 143 (2023) 102616.
[9] Amitojdeep Singh, Sourya Sengupta, Vasudevan Lakshminarayanan, Explainable deep learning models in medical image analysis, J. Imaging 6 (6) (2020) 52.
[10] Zohaib Salahuddin, et al., Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Comput. Biol. Med. 140 (2022) 105111.
[11] Ramprasaath R Selvaraju, et al., Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[12] Rachel Lea Draelos, Lawrence Carin, Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks, 2020, arXiv e-prints, arXiv–2011.
[13] Wojciech Samek, et al., Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2016) 2660–2673.
[14] Chih-Kuan Yeh, et al., On the (in) fidelity and sensitivity of explanations, Adv. Neural Inf. Process. Syst. 32 (2019).
[15] Junhee Lee, et al., Heatmap assisted accuracy score evaluation method for machine-centric explainable deep neural networks, IEEE Access 10 (2022) 64832–64849.
[16] Bolei Zhou, et al., Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[17] Kaiming He, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[18] Gao Huang, et al., Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[19] Christian Szegedy, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[20] Julius Adebayo, et al., Sanity checks for saliency maps, Adv. Neural Inf. Process. Syst. 31 (2018).
[21] Rachel Lea Draelos, Lawrence Carin, Explainable multiple abnormality classification of chest CT volumes, Artif. Intell. Med. 132 (2022) 102372.
[22] Luca Brunese, et al., Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays, Comput. Methods Programs Biomed. 196 (2020) 105608.
[23] Stephen M. Humphries, et al., Deep learning enables automatic classification of emphysema pattern at CT, Radiology 294 (2) (2020) 434–444.
[24] Sema Candemir, et al., Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network, Comput. Med. Imaging Graph. 83 (2020) 101721.

[25] Ziqi Tang, et al., Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline, Nat. Commun. 10 (1) (2019) 2173.
[26] Qier Meng, Yohei Hashimoto, Shin'ichi Satoh, How to extract more information with less burden: Fundus image classification and retinal disease localization with ophthalmologist intervention, IEEE J. Biomed. Health Inf. 24 (12) (2020) 3351–3361.
[27] Doniyorjon Mukhtorov, et al., Endoscopic image classification based on explainable deep learning, Sensors 23 (6) (2023) 3176.
[28] Alon Jacovi, Yoav Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? 2020, arXiv preprint arXiv:2004.03685.
[29] Zifan Wang, et al., Interpreting interpretations: Organizing attribution methods by criteria, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 10–11.
[30] Minh N. Vu, et al., Evaluating explainers via perturbation, Naval Postgraduate School, Monterey, California, 2019.
[31] Xiao-Hui Li, et al., Quantitative evaluations on saliency methods: An experimental study, 2020, arXiv preprint arXiv:2012.15616.
[32] Samuele Poppi, et al., Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2299–2304.
[33] Benjamin Vandersmissen, Jose Oramas, On the coherence of quantitative evaluation of visual expalantion, 2023, arXiv preprint arXiv:2302.10764.
[34] Jan-Peter Kucklick, Oliver Müller, Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal, ACM Trans. Manag. Inf. Syst. 14 (1) (2023) 1–24.
[35] Ioannis Kakogeorgiou, Konstantinos Karantzalos, Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing, Int. J. Appl. Earth Obs. Geoinf. 103 (2021) 102520.
[36] Gencer Sumbul, et al., Bigearthnet: A large-scale benchmark archive for remote sensing image understanding, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2019, pp. 5901–5904.
[37] Michael Schmitt, et al., SEN12MS–A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion, 2019, arXiv preprint arXiv:1906.07789.
[38] Alina Lopatina, et al., Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis, Front. Neurosci. 14 (2020) 609468.
[39] Athanasios Kallipolitis, Panayiotis Yfantis, Ilias Maglogiannis, Improving explainability results of convolutional neural networks in microscopy images, Neural Comput. Appl. (2023) 1–19.
[40] Jing Zhang, et al., Explainability for regression CNN in fetal head circumference estimation from ultrasound images, in: Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, IMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3, Springer, 2020, pp. 73–82.
[41] Jakob Nikolas Kather, Niels Halama, Alexander Marx, 100000 Histological images of human colorectal cancer and healthy tissue (v0.1), 2018, http://dx.doi.org/10.5281/zenodo.1214456.
[42] Muhammad E.H. Chowdhury, et al., Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8 (2020) 132665–132676.
[43] Tawsifur Rahman, et al., Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, Comput. Biol. Med. 132 (2021) 104319.
[44] Philipp Tschandl, Cliff Rosendahl, Harald Kittler, The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scient. Data 5 (2018) 180161.
[45] Fabio A. Spanhol, et al., A dataset for breast cancer histopathological image classification, IEEE Trans. Biomed. Eng. 63 (7) (2015) 1455–1462.
[46] Max sensitivity GitHub, 2019, URL https://github.com/chihkuanyeh/saliency_evaluation.
[47] Jacob Gildenblat, contributors, Pytorch library for CAM methods, 2021, GitHub https://github.com/jacobgil/pytorch-grad-cam.
[48] Alex Krizhevsky, Geoffrey Hinton, et al., Learning Multiple Layers of Features from Tiny Images, Toronto, ON, Canada, 2009.
[49] Adam Coates, Andrew Ng, Honglak Lee, An analysis of single-layer networks in unsupervised feature learning, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.
[50] Imagenette dataset, 2019, URL https://github.com/fastai/imagenette.
[51] Olga Russakovsky, et al., ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, http://dx.doi.org/10.1007/s11263-015-0816-y.