Influence of training and expertise on deep neural network attention and human attention during a medical image classification task

Rémi Vallée Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

Tristan Gomez Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

Arnaud Bourreille* CHU Nantes, Institut des Maladies de l'Appareil Digestif, CIC Inserm 1413, Université de Nantes, Nantes, France

Nicolas Normand* Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

Harold Mouchère* Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

Antoine Coutrot* Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
Univ Lyon, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, Lyon, France

In many different domains, experts can make complex decisions after glancing very briefly at an image. However, the perceptual mechanisms underlying expert performance are still largely unknown. Recently, several machine learning algorithms have been shown to outperform human experts in specific tasks. But these algorithms often behave as black boxes and their information processing pipeline remains unknown. This lack of transparency and interpretability is highly problematic in applications involving human lives, such as health care. One way to "open the black box" is to compute an artificial attention map from the model, which highlights the pixels of the input image that contributed the most to the model decision. In this work, we directly compare human visual attention to machine visual attention when performing the same visual task. We have designed a medical diagnosis task involving the detection of lesions in small bowel endoscopic images. We collected eye movements from novices and gastroenterologist experts while they classified medical images according to their relevance for Crohn's disease diagnosis. We trained three state-of-the-art deep learning models on our carefully labeled dataset. Both humans and machine performed the same task. We extracted artificial attention with six different post hoc methods. We show that the model attention maps are significantly closer to human expert attention maps than to novices', especially for pathological images. As the model gets trained and its performance gets closer to the human experts, the similarity between model and human attention increases. Through the understanding of the similarities between the visual decision-making process of human experts and deep neural networks, we hope to inform both the training of new doctors and the architecture of new algorithms.

Introduction

Made outstanding progress

During the past 20 years, machine learning algorithms have made outstanding progress in key domains such as computer vision, language processing, and decision-making. In particular, deep neural network architecture outperformed humans in well-defined tasks and environments such as Atari video games

Citation: Vallée, R., Gomez, T., Bourreille, A., Normand, N., Mouchère, H., & Coutrot, A. (2024). Influence of training and expertise on deep neural network attention and human attention during a medical image classification task. *Journal of Vision*, *24*(4):6, 1–27, https://doi.org/10.1167/jov.24.4.6.

https://doi.org/10.1167/jov.24.4.6 Received September 23, 2022; published April 8, 2024 ISSN 1534-7362

(Badia et al., 2020), various board games (Silver et al., 2018), the characterization of precise bacteria (Buetti-Dinh et al., 2019), and medical diagnosis (Abràmoff, Lavin, Birch, Shah, & Folk, 2018; Brinker et al., 2019; Brown et al., 2018; Esteva et al., 2017; Rajpurkar et al., 2018). However, these great performances were mostly produced by opaque models with an increasing number of parameters. This lack of transparency and interpretability is highly problematic in applications involving human lives, such as health care (Cabitza, Rasoini, & Gensini, 2017; Dave, Naik, Singhal, & Patel, 2020), criminal risk assessment (Angwin, Larson, Kirchner, & Mattu, 2013), or autonomous driving (Codevilla, Santana, Lopez, & Gaidon, 2019).

Here, we propose to investigate the interpretability of deep learning algorithms in a computer-aided medical diagnosis task. Medicine is a field with a growing demand for machine-learning approaches to ease the burden on often overworked doctors. However, doctors need algorithms that not only are performing well but are also trustworthy, transparent, interpretable, and explainable for human experts (Holzinger, Langs, Denk, Zatloukal, & Müller, 2019).

More precisely, we compare the parts of the images that humans and machines use to make their predictions in a medical image classification task. To do so, we use the attention maps produced by deep neural networks (artificial attention) and the human eye position maps recorded in an eye-tracking experiment (human attention).

Human attention

Our brain constantly receives a tremendous amount of information. Despite its substantial capacity, it cannot simultaneously process all the incoming stimuli. To select the most pertinent ones, the brain uses a filter, called attention. Our eye movements are a dynamic manifestation of where we direct our attention. We use them to sample our visual field, guided by both bottom-up and top-down mechanisms.

Bottom-up attention: Also called exogenous attention, it is a process driven by the stimuli, where salient features are automatically selected by the visual system. Since this process is only image based, it is much simpler to model, and many bottom-up visual saliency models have been proposed in the literature (for reviews, see Borji, 2021; Borji and Itti, 2013; Kümmerer and Bethge, 2021).

Top-down attention: Also called endogenous attention, this process is driven by the observer and influenced by their prior knowledge, their center of interest, the task at hand, or their cognitive state (de Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019). Since human factors are harder to model, this process has received less attention from the computer vision community (although see Schutt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2019; Tanner and Itti, 2019).

Machine attention

Attention in neural networks can be viewed as a top-down mechanism. It can also be split into two main categories: learned attention and post hoc attention.

Learned attention: This type of attention refers to a precise type of model where choosing the relevant part of an input is established as a goal by the objective function. We can distinguish two subtypes of learned attention: hard attention, which consists of the binary selection of the parts of the input that will be used for the final decision (Mnih, Heess, Graves, & Kavukcuoglu, 2014; Xu et al., 2015), and soft attention, which consists of the weighted parts of the original input used to make the decision (Bahdanau, Cho, & Bengio, 2015; Sharma, Kiros, & Salakhutdinov, 2015; Woo, Park, Lee, & Kweon, 2018). The latest is the most widespread form of learned attention due to its crucial role in the architecture of transformer networks.

Post hoc attention: Unlike learned attention, post hoc attention is computed once the training of the neural network is over. Its goal is to extract the parts of the input that lead to the final decision, by focusing on the information hidden in the "black-box" model. Post hoc attention can be quantified in different ways. For example, the gradient can be back-propagated to generate class spatial attention maps (Simonyan, Vedaldi, & Zisserman, 2013; Springenberg, Dosovitskiy, Brox, & Riedmiller, 2015), or the global average pooling of the gradient of a specific class can be used to weight the features activation map of a target layer (Selvaraju et al., 2017).

Computer vision and the science of human visual processing have been informing each other for many years (Gerhard, Wichmann, & Bethge, 2013; Hyvärinen, Hurri, & Hoyer, 2009; Olshausen and Field, 1996). With the advent of artificial neural networks, the question of whether deep networks and neurobiological systems use similar representations in support of similar tasks has gained even more traction (Barrett, Morcos, & Macke, 2019; Jacob, Rt, Katti, & Arun, 2021; Serre, 2019; Sinz, Pitkow, Reimer, Bethge, & Tolias, 2019). In particular, a few studies directly compared how humans and neural networks deploy their visual attention while performing the same task (Lai et al., 2020; Qi, Zheng, Yang, Cao, & Hsiao, 2023; Rong, Xu, Akata, & Kasneci, 2021). We can find some examples of this kind of parallel in reinforcement learning with Atari games (Guo et al., 2021), meal preparation (Li, Liu, & Rehg, 2020), driving (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017), and visual question answering (Sood, Kögel, Strohm, Dhar, & Bulling, 2021). Das et al. (Das, Agrawal, Zitnick, Parikh, & Batra, 2017) propose a comparison on a click-based

saliency dataset and conclude that attention models and humans do not observe the same areas during a visual question-answering task. This annotation method has been shown to produce noisy data that do not reflect the reality of human fixations (Tavakoli, Ahmed, Borji, & Laaksonen, 2017). Lai et al. (2020) compare the results obtained with soft attention architectures and real eye-tracking data from a publicly available eye-tracking dataset. They show that when both human visual attention and artificial attention are task driven, the higher the performance of the networks, the closer the artificial attention is to human attention. Even if this work is limited by a strong heterogeneity between the experimental conditions of humans and machine and by the limited number of participants, their results encourage us to continue working on this comparison.

As stated in the beginning of this introduction, the aid in medical diagnosis is arguably one of the fields where explainable artificial intelligence could be the most useful to society. However, only a few preliminary studies compared human and machine visual attention on a medical image classification task, and these preliminary studies only featured up to three medical experts, which makes any generalization difficult (Muddamsetty, Jahromi, & Moeslund, 2021). Explaining the decisions of deep neural networks has gained interest in recent years for the computer-aided medical diagnosis community. Although the comparison of human and machine visual attention in a medical image classification task has been explored (Thakoor, Koorathota, Hood, & Sajda, 2020), even in the context of wireless capsule endoscopy (WCE) images (Gatoula, Dimas, Iakovidis, & Koulaouzidis, 2021), these studies often include a limited number of participants, making any generalization difficult. This may be due to the novelty of the approach and the difficulty of recruiting medical experts for an eye-tracking experiment, as they are already swamped. In this article, we aimed to go beyond these limitations by recording the eye-tracking data of 22 participants, including 10 medical experts, while they performed a medical image classification task. We also trained state-of-the-art deep convolutional neural networks on the same task and computed its post hoc visual attention maps with different methods. This allowed us to compare the visual exploration strategies between human novices and human experts, as well as human versus machine visual attention maps.

Method

Medical image classification task: Detection of Crohn's disease lesions

In early 2000, the development of the video capsule endoscopy (VCE) allowed the complete examination of the small intestine (Iddan, Meron, Glukhovsky, & Swain, 2000). This led to an improvement in the diagnosis of Crohn's disease (CD) and its early treatment (Eliakim, 2017) by enabling the direct assessment of the small bowel lesions (Gal, Geller, Fraser, Levi, & Niv, 2008). Although the importance of VCE is established, it is not yet widespread for the diagnosis of Crohn's disease, and x-ray imaging with contrast, MRI, sonography, and traditional endoscopy are often preferred (Chen, Zhou, & Weltman, 2018). One of the main reasons is the extensive time required to review the VCE images. Each endoscopic video examination generates between 50,000 and 60,000 images and lasts several hours (from two to six frames per second). On average, the review time of a video by a gastroenterologist is estimated between 30 and 60 minutes (McAlindon, Ching, Yung, Sidhu, & Koulaouzidis, 2016). The review time is much shorter than the actual video duration because gastroenterologists know where to look and can skip the least relevant segments. The time-consuming nature of this task increases the demand for algorithms that would allow gastroenterologist experts to save time on their diagnosis and focus on the treatment of the disease.

Machine training and attention extraction

Dataset for training

We created a publicly available dataset CrohnIPI, which contains 3,498 labeled images from VCE (de Maissin et al., 2021). The image resolution was 640 × 640 pixels. These images have been carefully annotated in three different phases. The first phase was realized by a first expert who selected VCE images from 63 different patients with Crohn's disease. In the second phase, each image was independently labeled by three different experts. In the third phase, the three same experts reached a consensus on the images that they classified differently in the second phase (41% of the original dataset with at least one discordant observer).

Deep neural network for VCE image classification

Based on our publicly available dataset, we trained three state-of-the-art deep convolutional neural networks, VGG16, VGG19 (Simonyan & Zisserman, 2015), and ResNet34 (He, Zhang, Ren, & Sun, 2016), to classify images between two different classes: pathologic and nonpathologic. Pathologic images contained at least one of these seven Crohn-related lesions: erythema, edema, ulceration between 3 and 10 mm, ulceration over 10 mm, aphthoid ulceration, and stenosis. Nonpathologic images did not contain any lesions. To train each neural network, we used the entire dataset, split into two subsets: 80% for the training phase and 20% for the validation one. The networks

have been trained on ImageNet, except for the output classification layer, which has been initialized randomly. Then we fine-tuned the networks on the 3,500 images of the CrohnIPI dataset. No learning rate decay has been used, as it did not show any improvement compared to a fixed learning rate. The networks have been trained for a maximum of 60 epochs, with an early stop in case of 20 epochs without improvement on the validation set.

The network was then tested on the 250 images that compose the eye-tracking dataset. The ground truth was defined as the majority vote of both junior and senior doctors between the pathologic and the nonpathologic classes. We grouped all the Crohn-related lesions under the same pathologic label for both humans and machine. The accuracy of the eye-tracking image dataset, with 10 times cross-validation, is 81.4% for VGG16, 82.2% for VGG19, and 84.1% for ResNet34.

Post hoc attention extraction

To create artificial attention maps, we used six different post hoc methods: GradCAM (Selvaraju et al., 2017), guided GradCAM (Selvaraju et al., 2017), guided back-propagation (Springenberg et al., 2015), vanilla gradients (Simonyan et al., 2013), Score-CAM (Wang et al., 2020), and Randomized Input Sampling for Explanation (RISE) (Petsiuk, Das, & Saenko, 2018).

The GradCAM method uses activation maps weighted by importance coefficient. The vanilla gradients and the guided back-propagation method are based on the same idea: computing the gradient of the output with respect to the input. The difference between those two is that for the guided back-propagation method, we only compute the positive gradients. Thus, the gradient method shows the pixel that contributes the most to the output, whereas the guided back-propagation method only shows pixels that contribute positively to the output. The guided GradCAM method is the guided back-propagation method weighted by the GradCAM map.

Contrary to the first four methods, Score-CAM and RISE are not gradient based. They address some of the problems present in gradient methods like the presence of noise in saliency maps caused by gradient saturation/evanescence or the potential overemphasis on specific feature maps. Score-CAM assigns weights to individual feature maps based on the class score observed when masking the areas activating the maps. RISE involves dividing the image into a rectangular grid and calculating random binary masks. This procedure of masking the image and measuring the score variation is repeated several thousand times in practice to estimate which areas of the image, when not masked, lead to the largest class score and thus are the most important for the decision.

We applied a Gaussian blur on all the back-propagation methods. The only method that can produce negative values is the vanilla gradient. In this case, we took the absolute value of the maps. For a more complete description of these methods, see Appendix D.

Human attention dataset

To quantify the human attentional behavior on a medical classification task, we realized an eye-tracking experiment as described in the following section. To avoid the pitfalls when comparing human and machine attention, we specifically designed choices as explained in the discussion section.

Participants

We recorded the gaze of 23 participants. Eleven were senior gastroenterologists with over 5 years of experience, each having reviewed more than 100 video-capsules (VCE) in their careers. Five were junior doctors with fewer than 100 VCEs reviewed and between 1 and 5 years of experience. The remaining seven were novices who had never seen any bowel images. With around 30,000 images per VCE, 100 VCEs represent approximately 3 million images, a significant number that is typically only reached by senior doctors. We removed one of the senior gastroenterologists due to repeated calibration failures. The analyses were performed on a final group of 22 participants (13 males, 9 females). We are aware that 11 experts is considered a low sample size to conduct inferential statistical analyses. We justify this number by the considerable challenge of finding medical experts able to spare time for a behavioral experiment. To gather this dataset, we had to set up an itinerant eye-tracking system and visited several university hospitals in France over 1 year.

To assess how this sample size affects interobserver consistency, we quantified the similarity between experts' attention maps in a leave-M-out setting, where all but M subjects are used to compute an average attention map, which is then compared to the attention map of the left-out expert. We used Pearson's correlation coefficient (CC) to compare the maps; see Metrics. By varying the M, we obtained a curve of scores that shows a saturation effect for small M, indicating that adding more experts to the dataset would not drastically impact interobserver consistency; see Figure B2. We bootstrapped this analysis 10 times, randomly permuting the order of the left-out experts, and took the average over these 10 permutations.

Stimuli

The stimuli consisted of 250 VCE images from two patients diagnosed with Crohn's disease. From these two patients, we created a representative and

balanced panel of images encountered in patients with Crohn's disease. Crohn's lesions are quite stereotypical; they do not significantly vary between patients. The image resolution was 640 × 640 pixels. To limit the imbalance between the number of pathological and nonpathological images, they were selected by an expert who did not participate in the experiment. To label the images as pathological or not pathological, we used the 15 senior and junior doctors votes on each image. When a Crohn-related lesion was detected the corresponding image was labeled as pathological; otherwise, it was labeled as not pathological. The "I don't know" option corresponded to an abstention. In total, 58.0% of the images were classified as nonpathological, 40.8% as pathological, and 1.2% as indeterminate (three images). The three indeterminate corresponded to the images where the number of votes for pathological and nonpathological classes was equal. They were removed from the next analyses.

Experts and novices

Based on this ground truth, we ranked all the participants based on their correct classification score. Participants with an accuracy score above 80% formed the expert group (11 participants). The other 11 participants formed the novice group. As expected, all novices were classified in the novice group, but also one senior and three junior doctors (see in Table A1 in Appendix A). The novice group has a mean accuracy of 65.0% ($SD = 9.4\%$), mean specificity of 55.7% ($SD = 16.8\%$), and mean sensitivity of 77.9% ($SD = 18.8\%$). The expert group has a mean accuracy of 91.1% ($SD = 3.9\%$), mean specificity of 90.9% ($SD = 7.8\%$), and mean sensitivity of 91.5% ($SD = 4.7\%$).

Note that the task at hand in our study was not to diagnose Crohn's disease but to tell whether the image contained a Crohn-related lesion. As a comparison, in various clinical contexts, VCE diagnostic performance for Crohn's disease has been shown to vary from 49% to 77% (average = 61%) (Yang, Keum, & Jeen, 2016).

Apparatus

The stimuli were displayed on a DELL P2417H monitor at 60 Hz (1,920 × 1,080 pixels). The screen size was 52.7 cm by 29.6 cm. Eye-tracking data were recorded at 60 Hz with an eye tracker, "The EyeTribe."[1] Participants sat 60 cm away from the screen.

Procedure

The experiment workflow is described in Figure 1. The experiment consisted of two phases of approximately 20 minutes, with a mandatory break in between. Each phase consisted of reviewing 125 images. Images were presented in a different random order for every participant and independently of their clinical context, in the center of the screen. The participants could take a break after each image. No feedback was given to participants, preventing them from learning during the experiment. A 9-point calibration procedure was realized at the beginning of the experiment, after 20 successive images, and after a break. Between each image, a drift correction was also performed, to prevent a possible drift of the measurement device. The drift correction phase consisted of fixating a central fixation cross. If the gaze of the participant landed on the cross for at least half a second, the stimulus was displayed; otherwise, a new calibration was initiated. The stimulus was then displayed for 2 seconds during which eye-tracking data were recorded. When the image disappeared, a checkbox was displayed, asking the participant whether the image contained a pathological Crohn-related lesion.

Eye-tracking data processing

The eye-tracker raw data consist of timestamp t and position (x, y) for each sample. As the sequential nature of human gaze behavior does not have an equivalent in neural network post hoc attention, in our study, we focused on the spatial information. We parsed the eye positions into binary fixation maps (pixel = 1 if fixated, 0 otherwise) and continuous fixation maps (attention maps). The attention maps were obtained by convolving a two-dimensional (2D) Gaussian kernel across the fixation locations of each observer. Examples can be found in Figure 2.

Metrics

To compare the spatial distributions of attention, we used two metrics widely used to compare gaze and attention spatial distributions (Bylinskii, Judd, Oliva, Torralba, & Durand, 2019).

- Normalized scanpath saliency (NSS): The NSS is the average of the values of the z-scored attention map at the fixation locations (Peters, Iyer, Itti, & Koch, 2005). It is defined by the following equation, with a given attention map P, a binary fixation map Q^B, and i the index of the ith pixel:

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \overline{P_i} \times Q_i^B$$

where

$$N = \sum_i Q_i^B$$

and

$$\overline{P} = \frac{P - \mu(P)}{\sigma(P)}$$

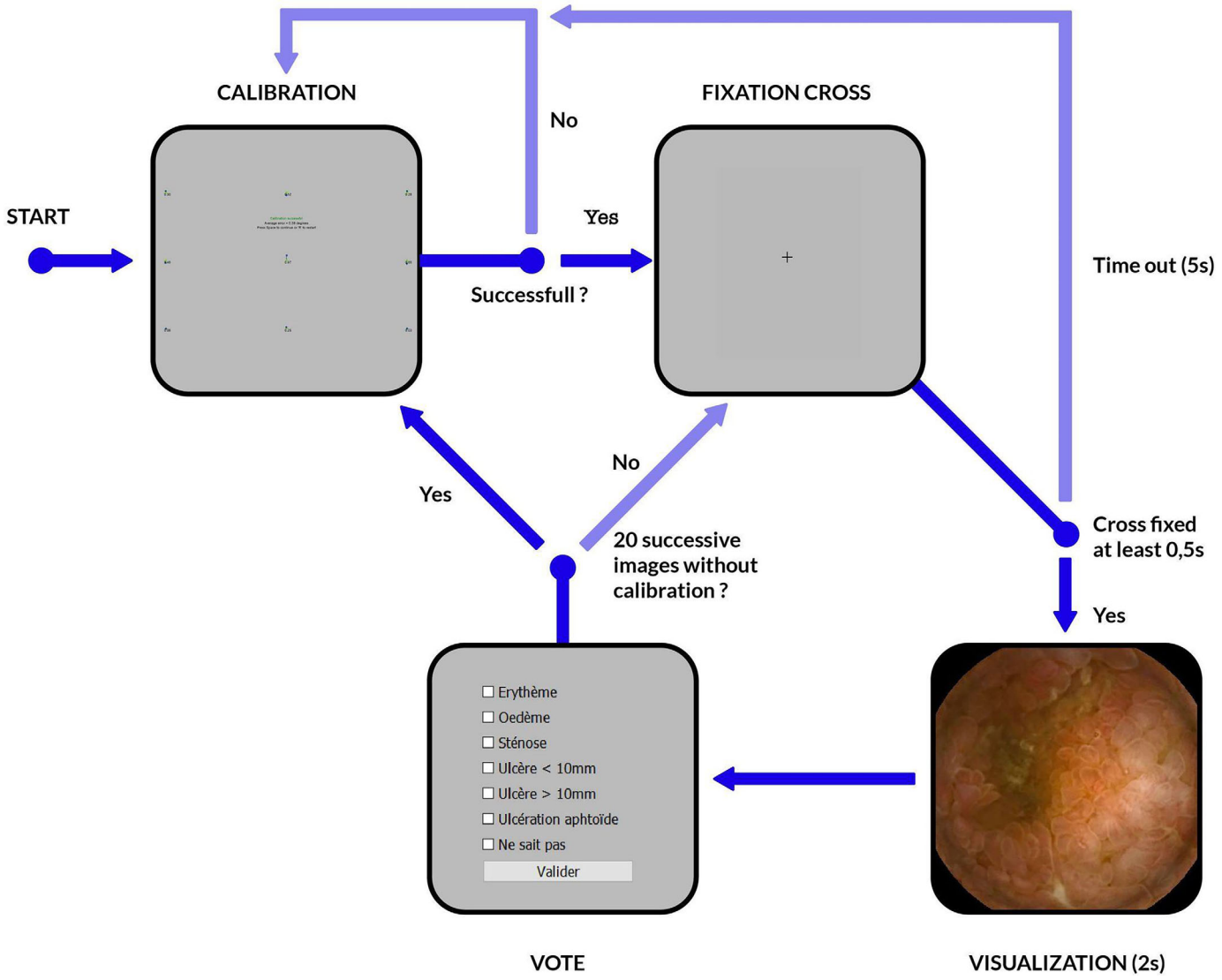


Figure 1. Design of the eye-tracking experiment.

NSS = 0 means that the fixations and the attention map are unrelated, NSS > 0 means that they are positively related, and NSS < 0 means that they are negatively related.

- Pearson's correlation coefficient (CC): This metric quantifies the linear relationship between two variables. It is defined by the following equation, with two attention maps P and Q.

$$CC(P, Q) = \frac{\sigma(P, Q)}{\sigma(P) \times \sigma(Q)}$$

Results

Figure 2 shows examples of human and artificial attention on different images of the eye-tracking dataset. All these artificial attention maps are obtained on ResNet34 trained on the CrohnIPI and tested on our eye-tracking image dataset. Even if these methods rely on quite different approaches (e.g., gradient-based vs. not gradient-based), we observe a certain consistency between the artificial attention maps.

To analyze the associations between metric scores, image labels, and level of expertise, we will use linear mixed models with metric scores as dependent variables; label, expertise, and their interaction as fixed effects; and images and observers as random effects. This section is divided into two parts. First, we focus on the differences in attention distribution between the expert and nonexpert human groups. We show that the image label (pathological or not) and the expertise level of the participant have a significant influence on the spatial distribution of human visual attention. Second, we compare artificial and human attention maps.

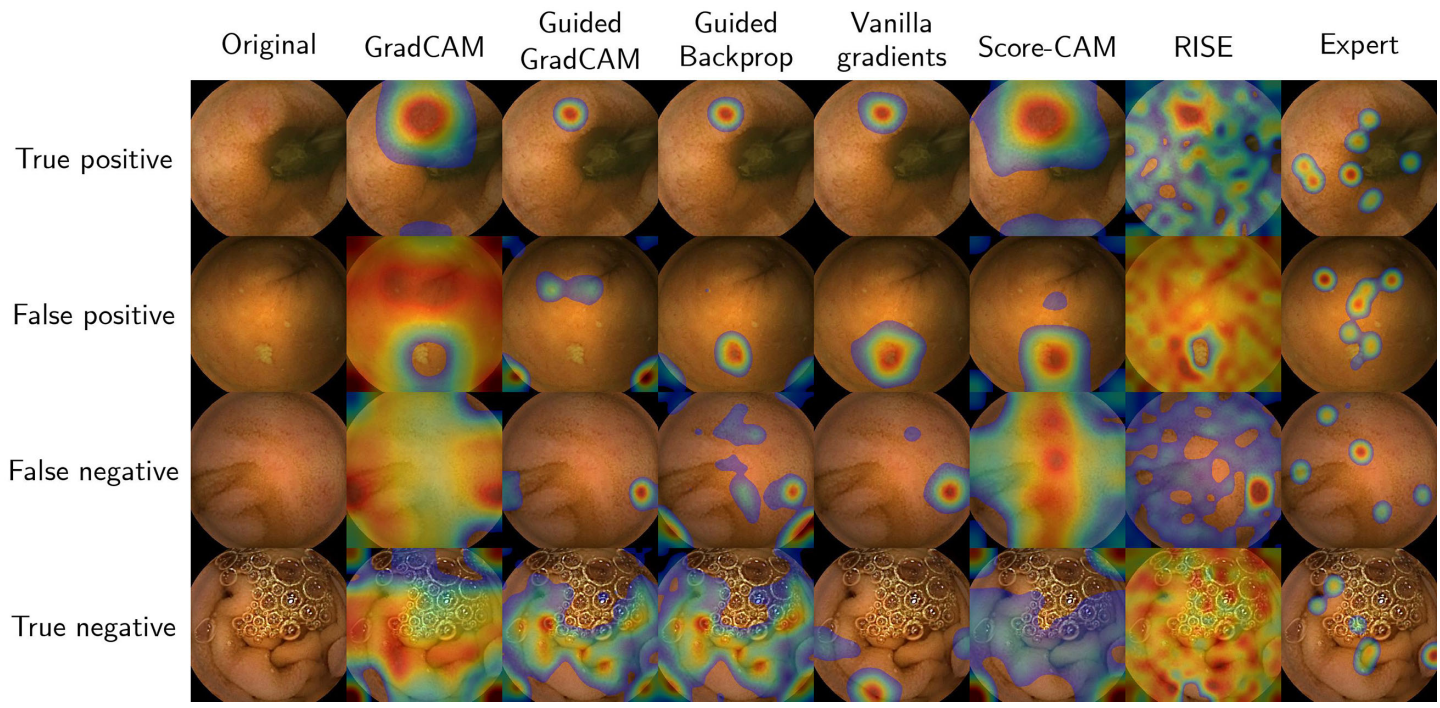


Figure 2. Examples of machine and human attention maps on different images of the eye-tracking dataset. The predictions (pathological or not) are computed with ResNet34 trained on the CrohnIPI dataset. The artificial attention maps are computed with six different methods from the same deep neural network training. The last column corresponds to the eye positions of a human expert.

Effect of expertise and image label on human gaze behavior

In this section, we compare how experts and novices move their eyes across pathological and nonpathological images.

Dispersion and distance to center

The average distance to the screen center and the average gaze dispersion are two simple and highly interpretable metrics to quantify a participant's gaze behavior.

The distance to the center can represent how active a participant has been during the exploration of a stimulus. If the participant is passive, they will likely stay near the image center waiting for the experiment to move on, while a more proactive participant will have a more dynamic gaze behavior and explore areas further away from the screen center. The dispersion of the eye positions quantifies how the participant spread their attention over the stimuli. If the participant scattered their eye positions all over the image, the dispersion will be high. If the participant focuses on a specific area, the dispersion will be low. For each metric, we computed a linear mixed model with expertise (expert, not expert), label (pathological, not pathological), and their interaction as fixed effects, with random intercepts for each participant and image: score $\sim$ expertise*label + (1 | participant) + (1 | image).

For the distance to center, the label effect was not significant $t(1, 5{,}994) = 1.2387$, $p = 0.21$, Cohen's $d = 0.04$, 95% CI [−0.02, 0.09], nor was the effect of expertise, $t(1, 5{,}994) = -0.896$, $p = 0.37$, Cohen's $d = 0.21$, 95% CI [0.16, 0.27]. However, we found a significant effect of the interaction, $t(2, 5{,}994) = -2.1866$, $p = 0.03$. For the dispersion, the effect of the label was significant $t(1, 5{,}994) = -6.1922$, $p < 0.001$, Cohen's $d = 0.27$, 95% CI [0.22, 0.32], but not the effect of the expertise, $t(1, 5{,}994) = -1.183$, $p = 0.23$,Cohen's $d = 0.16$, 95% CI [0.11, 0.21]. We found a significant effect of the interaction, $t(2, 5{,}994) = 5.33$, $p < 0.001$. As shown in Figure 3, the dispersion is higher in nonpathological images than in pathological images for both experts and nonexperts, and this effect is stronger in experts. This can be interpreted as the lesions present in pathological images attracting attention, hence decreasing the dispersion. The experts are more likely to spot them, and this effect is logically strengthened in this group.

The distance to the center and the gaze dispersion coarsely quantify gaze behavior but cannot tell whether observers specifically attended the same locations. In the next section, we will compare the spatial distribution of attention with the metrics previously introduced.

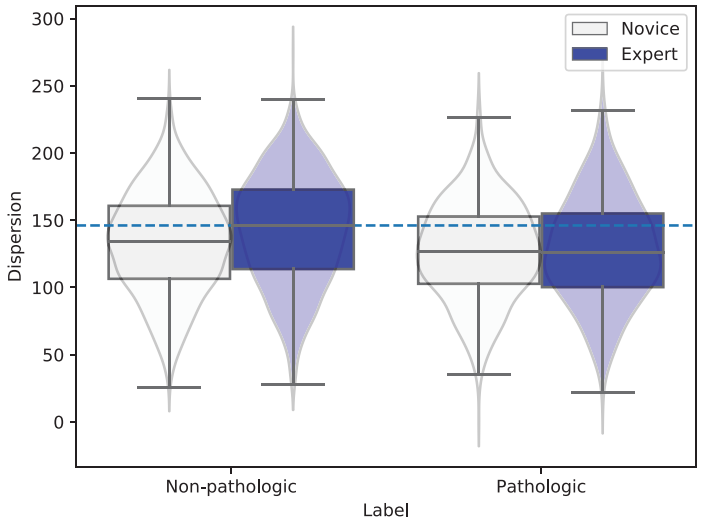


Figure 3. Association between gaze dispersion, level of expertise, and image label. Gaze dispersion corresponds to the variance of eye positions. The boxes represent the upper and lower quartiles around the median. The dashed line corresponds to the highest median value. The transparent curves represent the distributions of the data points.

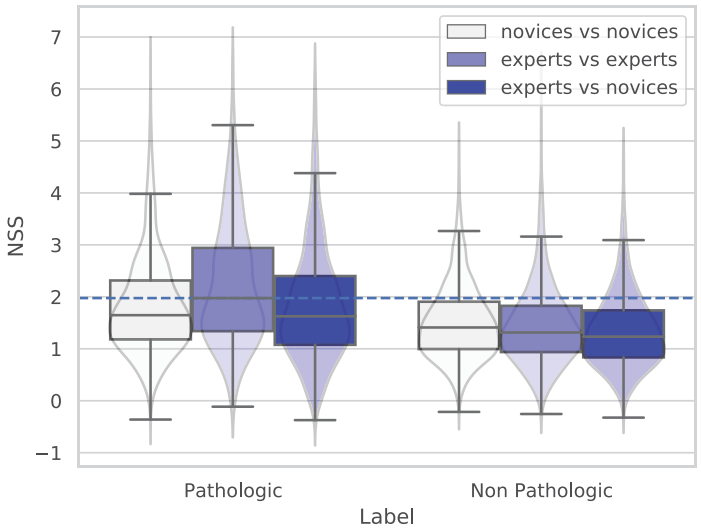


Figure 4. Comparison of human attention maps. The similarity between attention maps is assessed by the normalized scanpath saliency score (NSS; higher means closer distributions of attention). For pathological and nonpathological images. The "novices vs. novices" and "experts vs. experts" labels correspond to the intragroup comparisons. The "experts vs. novices" label corresponds to the intergroup comparison.

Comparison of the attention maps between humans

For pathological and nonpathological images, we assessed the effect of expertise on the spatial distribution of attention through two types of comparison: intragroup and intergroup; see Figure 4.

For the intragroup comparison (see boxplots "novices vs. novices" and "experts vs. experts"), we created a reference attention map by averaging the attention maps of every member of a group excluding the one being processed (leave-one-out procedure). Thanks to the saliency metrics, we obtained two scores (NSS, CC) for each subject, on each image, indicating whether the subject on a given image was watching the same areas as the members of their group. We also realized an intergroup comparison, by averaging the maps of one group and comparing this map with all the maps of the other group. In the following, we give the results with the NSS score, but the results with CC were similar and are available in see Figure B1 in Appendix B.

As for the dispersion and distance to the center, we computed for the NSS score a linear mixed model with label, expertise, and their interaction as fixed effects and images and observers as random effects. We found a significant effect of image label, $t(1, 10{,}776) = 11.62$, $p < 0.001$, Cohen's $d = 0.57$, 95% CI [0.53, 0.61], and of its interaction with the expertise level of observers, $t(2, 10{,}776) = -14.01$, $p < 0.001$. The effect of expertise was not significant, $t(1, 10{,}776) = -0.17$, $p = 0.86$. As shown in Figure 4, the NSS score was higher for pathological images than for nonpathological ones. This shows that on pathological images, the spatial distribution of visual attention is more similar across observers than on nonpathological images. This could be due to the presence of lesions guiding the attention to the same areas.

To further investigate the interaction between label and expertise, we computed two independent linear mixed models with expertise as a fixed effect and random intercepts for participants and images. The first one only uses pathological images, and the second one only uses nonpathological images.

We found a significant effect of expertise with pathological images ($t(1, 4{,}442) = -2.36$, $p = 0.02$), but not with nonpathological images ($t(1, 6{,}334) = -0.21$, $p = 0.83$). In terms of effect sizes, on pathological images, "experts/experts" vs. "novices/novices," Cohen's $d = 0.37$, 95% CI [0.28, 0.45]; "experts/experts" vs. "experts/novices," Cohen's $d = 0.51$, 95% CI [0.43, 0.60]; "novices/novices" vs. "experts/novices," Cohen's $d = 0.24$, 95% CI [0.15, 0.32]. On nonpathological images, "experts/experts" vs. "novices/novices," Cohen's $d = 0.09$, 95% CI [0.02, 0.16]; "experts/experts" vs. "experts/novices," Cohen's $d = 0.05$, 95% CI [0.02, 0.12]; "novices/novices" vs. "experts/novices," Cohen's $d = 0.15$, 95% CI [0.08, 0.22]. This could be interpreted as an effect of the pathological lesions drawing the attention of the experts to the same areas, while the nonexperts are less guided by them. On nonpathological images, there is nothing to guide the spatial visual attention of either experts or nonexperts, hence no differences between groups.

	Pathologic			Non Pathologic		
	novices	experts	machine	novices	experts	machine
novices	0.26	0.27	0.07	0.20	0.15	0.01
experts		0.35	0.13		0.18	0.02

Table 1. Proportion of shared variance between the attention maps of human novices, human experts, and the ResNet34 model extracted with the vanilla gradient algorithm. The higher the value, the more similar the attention maps.

Machine versus human attention maps

In this section, we compare the human and machine attention maps. As described previously, we used six different post hoc methods to compute for each image the map of the pixels involved in the decision-making process of three different deep neural networks. We detail the results with the vanilla gradient method and ResNet34, but other machine attention extraction methods (guided back-propagation, GradCAM, guided GradCAM, RISE, and Score-CAM) applied on other networks (VGG16 and VGG19) lead to similar results.

With the vanilla gradient method

Here we compare the human and artificial attention maps. We used the gradient method to obtain for each image the map of pixels involved in the decision process of the ResNet34 deep neural network.

In Table 1, we show the proportion of shared variance between the attention maps of human novices, human experts, and the ResNet34 model extracted with the vanilla gradient algorithm. The proportion of shared variance corresponds to the average squared coefficient of correlation between pairs of attention maps. Consistently with Figures 4 and 5, attention maps are more similar when they contain a pathologic lesion and when comparing attention maps between human experts (35% of shared variance). The amount of shared variance between human and machine attention is very low for nonpathological images (1% and 2% for novices and experts, respectively). It is higher for pathological images, and machine attention is two times more similar to experts (13%) than to novices (7%). This is consistent with the classification performance of the deep learning model, which is closer to that of the expert group than to that of the nonexpert group (accuracy = 84.1%, see in Table A1 in Appendix A). We also ran the same analysis, but instead of computing pairwise attention map comparisons, we computed for each expert (resp. novice) the correlation between their attention map and the average attention map of all other experts (resp. novices). This led to similar results. On nonpathological images, the shared variance between experts was 26%, same as between novices. On pathological images, the shared variance between experts was 39%, while between novices, it was 29%.

To assess the statistical significance of these results, we computed the same linear mixed model as in Dispersion and distance to center and Comparison of the attention maps between humans (fixed effect: label, expertise, and their interaction; random effect: images and observers), this time for the NSS score between the machine and human attention maps. This model is applied to 55,000 observations (250 images × 22 observers × 10 cross-validations). As shown in Figure 5, the same effects as in the human attention comparison are visible. We found that the label has a

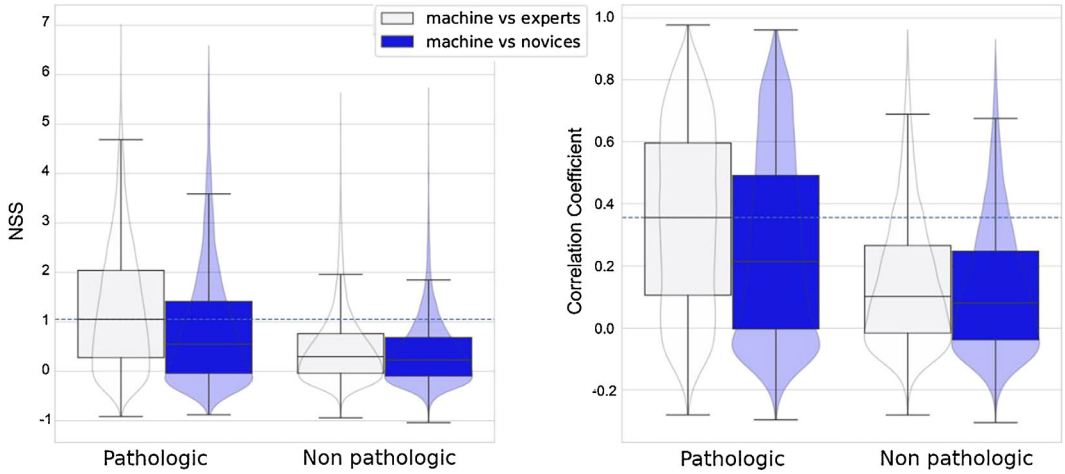


Figure 5. Comparison of the similarity between the spatial distribution of human and machine attention (higher NSS or correlation coefficient [CC] means more similar), for pathological and nonpathological images. The label "machine vs. nonexpert" corresponds to the comparison between the nonexpert group and the artificial attention extracted by the gradient method. The label "machine vs. experts" corresponds to the same comparison with the group of experts. These results were obtained for the ResNet34 network and the vanilla gradient attention extraction method.

significant effect on the NSS score, $t(55{,}000) = 10.508$, $p =$ 1e-26, Cohen's $d = 0.66$, 95% CI [0.65, 0.68], with more similar attention maps for pathological images than for nonpathological images. We did not find an effect of human expertise on the NSS score, $t(55{,}000) = -1.00$, $p = 0.31$, Cohen's $d = 0.22$, 95% CI [0.20, 0.24]. As expected, both artificial and human attention distributions are quite dependent on the image content.

To better understand the interaction between label and expertise on the comparison between artificial and human attention, we computed two independent linear mixed models with expertise as a fixed effect and random intercepts for participants and images. The first model uses only pathological images and the second only nonpathological images. For both models, we find a significant effect of expertise with $t(22{,}440) = -2.7379$, $p = 0.006$, Cohen's $d = 0.34$, 95% CI [0.31, 0.36] for pathological images and $t(22{,}440) = -2.0972$, $p = 0.04$, Cohen's $d = 0.12$, 95% CI [0.09, 0.14] for nonpathological images.

We verified that these results hold when controlling for interobserver consistency. Indeed, a strong heteroscedasticity between experts' and novices' attention maps could bias the results. We computed the same linear mixed models as above, adding the dispersion displayed in Figure 3 as a fixed effect. As in the previous paragraph, we find a significant effect of expertise with $t(22{,}440) = -3.05$, $p = 0.002$, Cohen's $d = 0.34$, 95% CI [0.31, 0.36] for pathological images and $t(22{,}440) = -2.25$, $p = 0.02$, Cohen's $d = 0.12$, 95% CI [0.09, 0.14] for nonpathological images. Another way to look at these results is to notice that on pathological images, the machine attention accounts for 0.13/0.35 = 37% of the between-expert explainable variance, while only accounting for 27% of the between-novice explainable variance. On nonpathological images, the machine attention accounts for 0.02/0.18 = 11% of the between-expert explainable variance, while only accounting for 5% of the between-novice explainable variance.

The results observed with the ResNet34 network are consistent with the results obtained with the VGG16 and VGG19 networks. A significant label effect is present on all data (VGG16: $t(55{,}000) = 12.687$, $p =$ 7e-37; VGG19: $t(55{,}000) = 14.338$, $p = 1e-46$), and an expertise effect is visible on pathological images (VGG16: $t(22{,}440) = -2.3538$, $p = 0.018$; VGG19: $t(22{,}440) = -2.3509$, $p = 0.018$).

With other machine attention extraction methods

Our results are stable across deep neural networks and artificial attention extraction methods. The different linear mixed models were calculated for the three deep neural networks and for the three other methods of post hoc attention extraction. The comparisons between human attention and machine attention extracted with the other methods are presented in Figure 6. The results with the NSS metric are shown in Figure C1 in Appendix C. For these methods of post hoc attention extraction, we observe a significant effect of the label on all the images and a significant effect of the expertise on the pathological images. However, it is only with the gradient method that the effect of expertise on nonpathological images is significant. The effect of expertise on pathological images is less strong for the other methods than for the gradient method. With the guided back-propagation method, we obtain a significant label effect on the NSS score with $t(55{,}000) = 12.015$, $p =$ 3e-33. When the test is performed only on pathological images, an effect of expertise is also significant with $t(22{,}440) = -2.2355$, $p = 0.025$. For the GradCAM method, the effect of the label is significant on all the images on the NSS metric with $t(55{,}000) = 9.1345$, $p =$ 6e-20, and an effect of the expertise visible on the pathological images $t(22{,}440) = -2.085$, $p = 0.0370$. For the guided GradCAM method, the effect of the label as for the other models is significant on the NSS metric, $t(55{,}000) = 9.1059$, $p =$ 8e-20, and an effect of the expertise is significant on the pathological images, $t(22{,}440) = -2.085$, $p = 0.037$.

The results of the different methods on VGG16 and VGG19 with the CC and NSS metrics are shown in Figure C2 and C3 in Appendix C.

Evolution of the similarity between human and machine attention across learning

In the previous section, we saw that the behavior of artificial attention was closer to the attentional behavior of experts on pathological images. As the performances of the three deep neural networks are closer to those of the experts than those of the novices, we will question here the influence of the network performances on this comparison. We saw in the section "Human comparison" that participants with higher image classification performances tend to look at the same areas. Thus, by comparing attentional maps extracted at different times during training and therefore on networks with different levels of performance, we seek to verify the hypothesis that attentional behaviors evolve with the level of expertise of the network and that the higher their performance, the closer its attentional behavior is to the behavior of an expert.

To test this hypothesis, we recorded the weights of the networks at different times during their training on the CrohnIPI database. For each network, weights were recorded for the first 10 epochs, as well as for the initial state where the network is pretrained on ImageNet. Weights were then recorded every 5

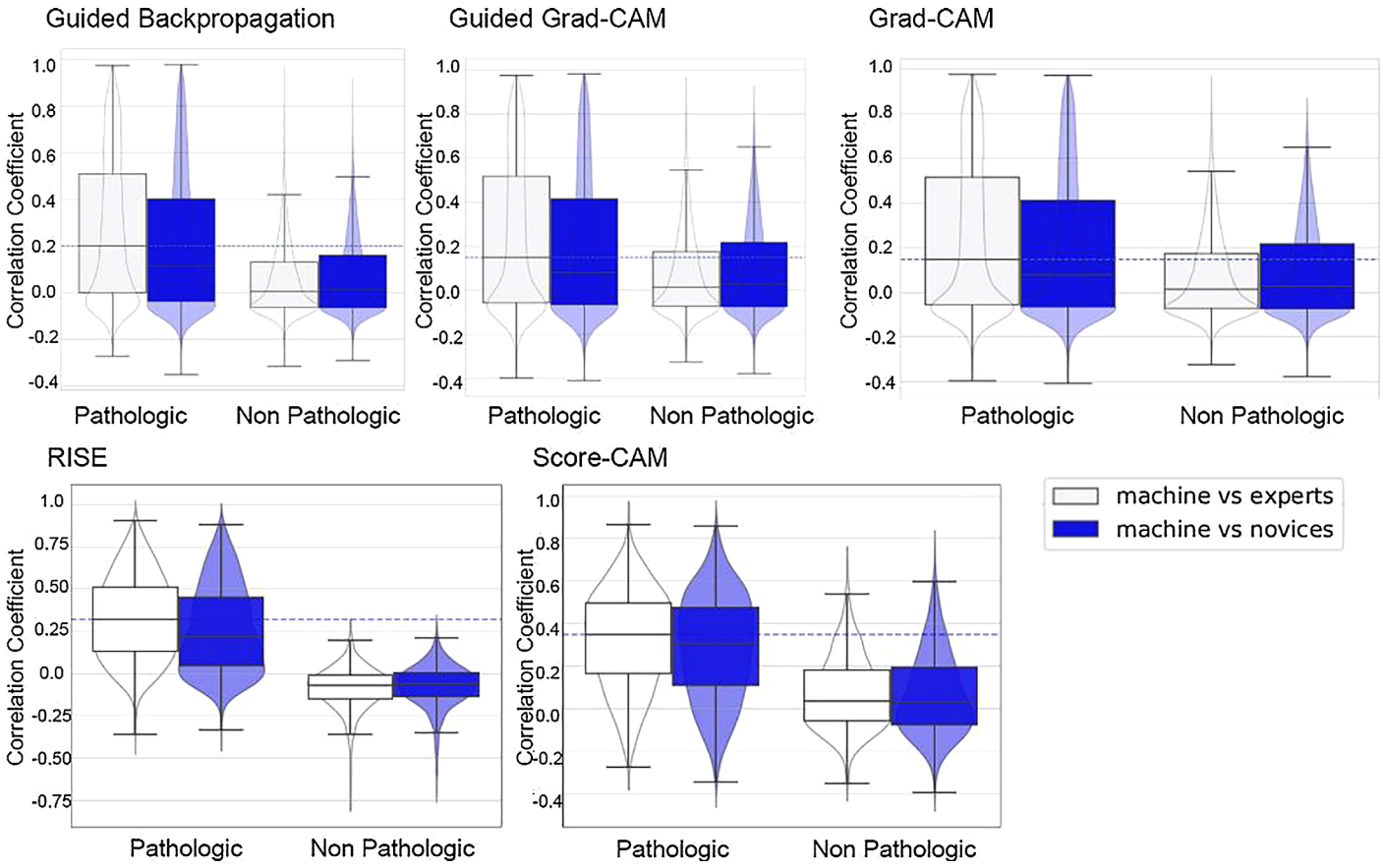


Figure 6. Comparison of the similarity between the spatial distribution of human and machine attention (higher correlation coefficient [CC] means more similar), for pathological and nonpathological images. The label "machine vs. nonexpert" corresponds to the comparison between the nonexpert group and the artificial attention extracted by the different methods. The label "machine vs. experts" corresponds to the same comparison with the group of experts. These results were obtained for the ResNet34 network.

epochs until 50 epochs after the network reached the minimum error on the validation set. For each of the 10 cross-validations performed for each of the three networks, the attentional maps obtained with each of the four methods were compared with the maps of all 22 participants in the eye-tracking experiment.

To further investigate the relationship between expertise and attentional behavior, we computed artificial attentional maps at different time points during network formation and performed the same comparison with human attention maps. Figure 7 shows that as the network becomes more accurate, its artificial attention maps (vanilla gradient) become closer to the attention maps of human experts. The similarity between artificial attention maps and human novices' attention maps also increases with network accuracy but at a slower rate. The same observation can be made when varying the sensitivity of the network and across the training process, but not its specificity due to a ceiling effect.

A scattering effect can be observed on the comparison across learning and be attributed to diverse factors. Foremost, the instability of the training process can play a role in the observed phenomenon, and this effect could have been mitigated by lowering the learning rate. The instability of artificial attention over training can further increase this phenomenon. Indeed, nonpathological images do not contain specific cues, challenging the network to keep a consistent attentional behavior. However, when we look at the evolution of comparison in the context of pathological images, the scattering is more moderate, with $SD = 0.00916$ on pathological images compared to $SD = 0.0212$ on all images (measurement produced on comparison with experts function of accuracy).

Similar results were obtained for the other artificial attention extraction methods and the other networks (see Figures E1 to E3 of Appendix E).

This validates our hypothesis: There is a relationship between the expertise of humans and machine and their attentional behavior. The higher the performance of the machine, the closer its attentional behavior is to human experts.

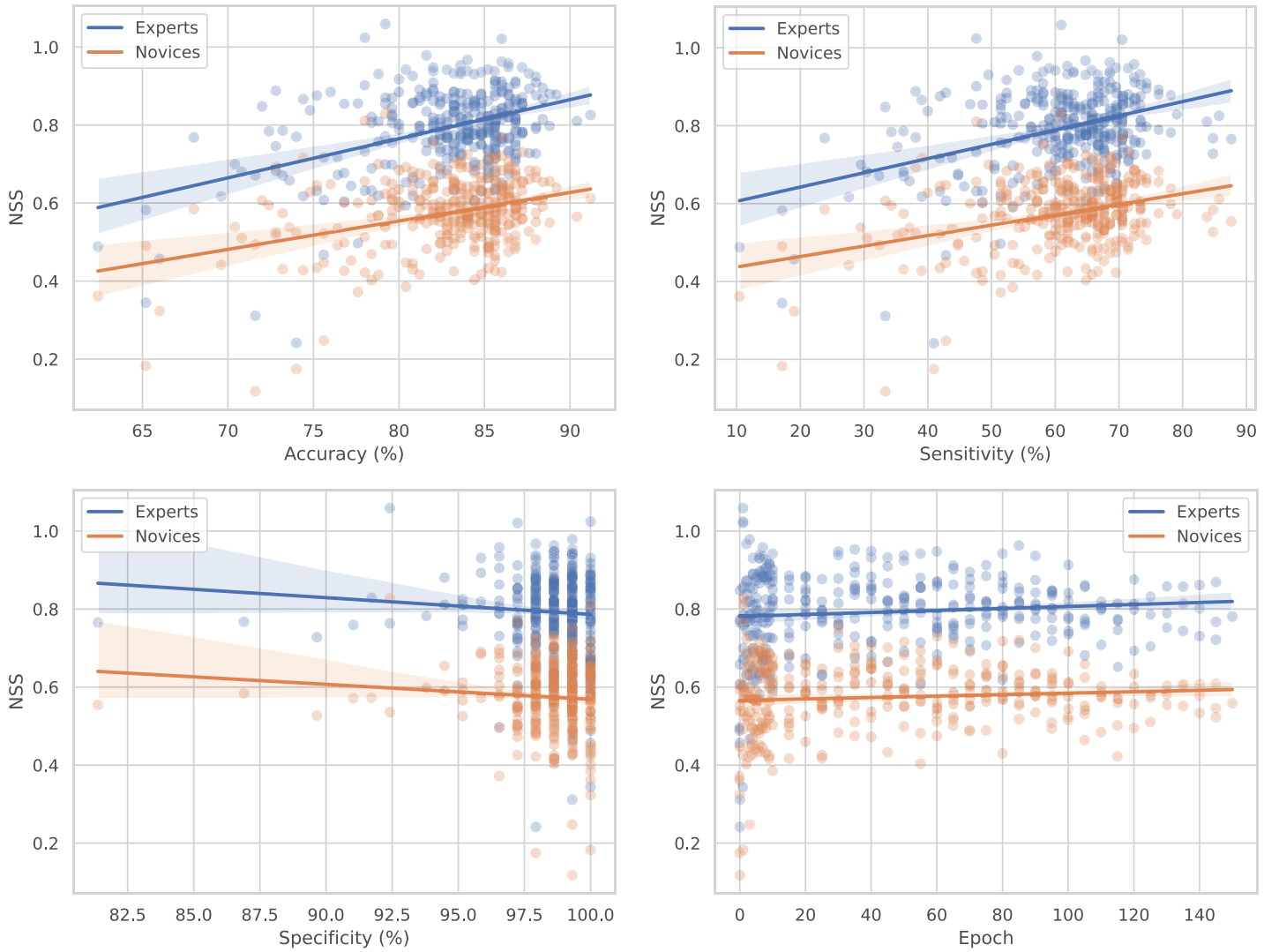


Figure 7. Association between the performance of the network and the similarity between its attention maps and the attention maps of human experts and novices. The higher the NSS, the more the attention maps of the network are similar to the humans'. Each point corresponds to the average NSS for all the images according to the performance reached during the training of the network. The curves correspond to linear regressions for novices and experts. The artificial attention is computed with the vanilla gradient method on the ResNet34 network. Similar curves for the other artificial attention extraction methods are available in Appendix D.

Discussion

We built a carefully labeled image dataset for a specific medical imaging classification task. We recorded and compared the visual attention of medical experts, novices, and deep neural networks while performing this classification. Several biases have been identified in the literature when comparing human and machine visual perception (Lai et al., 2020). To avoid them, we tried to design the human and machine classification tasks as closely as possible. First, since the number of parameters was identical for each image, we were able to fix the number of computations performed by the network. Assuming that the number of computations performed almost simultaneously by the machine corresponds to the image viewing time for humans, we also set the same viewing time (2 seconds) for each image. Then, each image was visualized by the network independently of the previous and following images, and without any information about the patient. Similarly, we presented the images to humans independently of their clinical context and in a different random order for each participant. Third, the parameter values were fixed and calculated from previous training on our dataset. Similarly, no feedback was given to participants, preventing them from learning during the experiment.

We found that visual attention strategies were more similar between experts than between novices. This indicates that high performance in our medical diagnostic task correlates with a specific attentional behavior. We compared expert attention with artificial attention, which was extracted from state-of-the-art convolutional neural networks with six different post hoc attention extraction methods. We showed that as the network's performance gets closer to the performance of the experts, its attention maps also get more similar to the attention maps of experts, mainly on pathological images. This correlation was strengthened when the network improved in the classification task.

We were surprised by the performance of some novices well above chance, even without any feedback from the experimenters. It might be interesting to evaluate the evolution of their performance throughout the task with an unsupervised deep learning algorithm. Novices, when evaluating endoscopic images, initially learn to distinguish normal from abnormal images and then group abnormal images into subcategories.

The comparison of the different methods of post hoc attention extraction allows us to evaluate the one most likely to help medical experts in their diagnosis. Faced with a ground truth composed of oculometric data from medical professionals, we observe that the results obtained by the different methods are not equal. It is possible to classify them from the most to the least similar to human attention given the different experiments performed: first the gradient method, then the guided back-propagation method, followed by guided GradCAM, and finally GradCAM.

In the article by Adebayo et al. (2018), a comparison of post hoc attention methods was performed through two tests. The first test consisted of comparing the maps obtained for a randomly initialized network and for a network trained on a specific task. If the method was effective, the obtained salience results should have been different, showing that the method was dependent on the model's parameters. The second test consisted of training two similar models on a dataset containing the same images but whose labels have been randomly swapped for the second. Through this experiment, the authors sought to identify whether the attention extraction method was indeed sensitive to the relationship between image and label. They concluded that, among the different methods of artificial attention extraction, only the gradient method and the GradCAM method passed both tests.

Here, we show that the artificial attention maps obtained with the gradient method are closer to the attention maps of experts than to the attention maps of nonexperts, on both pathological and nonpathological images. The presence of an effect of expertise on nonpathological images can be explained by the fact that, unlike other methods that only account for the parts of the image that contributed positively to the prediction, the gradient method also accounts for the elements that decreased the final prediction score. Although the algorithm answered "not pathological," the method shows areas that made it doubtful and thus allows us to account for recognition errors. For the other methods, all the errors are assimilated to detection errors, since the parts "observed" by the network but that did not lead to the conclusion of a presence of pathology are erased by applying a ReLU on the gradients. It can be observed in Figure G1 in Appendix G that for the gradient method and unlike the other methods, the difference between artificial and human attentional behaviors is small (i.e., the NSS is high) on the false negatives. Although the image has been misclassified by the network, its attentional behavior is as close to experts' as if the image was pathological and correctly classified.

We also performed a stability test across each method by observing if different trainings lead to different artificial attention maps. To do so, the attention maps of the networks were computed for each of the six methods, for each of the three networks, and for five distributions of the training data. Once these maps were computed, we used the Pearson correlation coefficient (CC) to evaluate for a given network, for a given method, and for each image if the networks pay attention to the same areas according to the network initialization. Thus, for each of the maps obtained, for each of the five distributions, the CC was calculated with the four other maps obtained with the other distributions. The more stable the attention maps are from one initialization to another, the closer the correlation score (CC) is to 1. The results of this experiment show that the gradient method is more stable on pathological images than the three other methods. All the scores of this experiment are presented in Appendix F. This stability across the training set is important as it guarantees that its artificial attention maps truly are indicative of the algorithm's decision process.

A critical difference between our experimental design and the reality of the clinical practice of gastroenterologists is the way VCE images are presented. We sequentially presented 250 independent images, 2 seconds per image, while gastroenterologists usually look at a sequence of consecutive frames that they can pause, rewind, or accelerate. Hence, gastroenterologists have access to the context of each image (what is just before and just after), which is likely to modify the way they look at it compared to an isolated image. This difference between our experiment and doctors' real-life practice might attenuate the effect of their expertise. Future studies could focus on how this contextual effect impacts attentional deployment, both in humans and in machines.

Keywords: attention, eye tracking, deep learning, medical imaging

Acknowledgments

Supported in part by an unrestricted grant from the IBD patients' association François Aupetit and the interdisciplinary project CrohnIPI of Nantes University (https://www.afa.asso.fr/).

Data availability: Instructions to download the eye-tracking dataset are available at https://crohnipi.ls2n.fr/en/crohn-ipi-project/.

Commercial relationships: none.
Corresponding author: Rémi Vallée.
Email: remi.vallee88@gmail.com.
Address: Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France.

*AB, NN, HM, and AC are co-senior authors of this article.

Footnote

[1] https://theeyetribe.com/theeyetribe.com/about/index.html.

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine, 1*(1), 39.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2013). Machine bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., . . . Blundell, C. (2020). Agent57: Outperforming the atari human benchmark. In *International conference on machine learning, Vienna* (pp. 507–517). PMLR.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA, USA: Conference Track Proceedings.

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology, 55*, 55–64.

Borji, A. (2021). Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(2), 679–700.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., . . . Schrüfer, P. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer, 113*, 47–54.

Brown, J. M., Campbell, J. P., Beers, A., Chang, K., Donohue, K., Ostmo, S., . . . Kalpathy-Cramer, J. (2018). Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications* (Vol. 10579, pp. 149–155). Houston: SPIE.

Buetti-Dinh, A., Galli, V., Bellenberg, S., Ilie, O., Herold, M., Christel, S., . . . Dopson, M. (2019). Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnology Reports, 22*, e00321.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(3), 740–757.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Journal of the American Medical Association, 318*(6), 5120177–5120518.

Chen, Y.-H., Zhou, D., & Weltman, D. I. (2018). Crohn disease imaging. *Medscape*, https://emedicine.medscape.com/article/367666.

Codevilla, F., Santana, E., Lopez, A. M., & Gaidon, A. (2019). Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9329–9338). Seoul.

Das, A., Agrawal, H., Zitnick, C. L., Parikh, D., & Batra, D. (2017). Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding, 163*, 90–100.

Dave, D., Naik, H., Singhal, S., & Patel, P. (2020). Explainable ai meets healthcare: A study on heart disease dataset. *arXiv preprint* arXiv:2011.03195.

de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences

in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences, 116*(24), 11687–11692.

de Maissin, A., Vallée, R., Flamant, M., Fondain-Bossiere, M., Berre, C. L., Coutrot, A., … Bourreille, A. (2021). Multi-expert annotation of Crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network. *Endoscopy International Open, 9*(7), E1136–E1144.

Eliakim, R. (2017). The impact of panenteric capsule endoscopy on the management of Crohn's disease. *Therapeutic Advances in Gastroenterology, 10*(9), 737–744.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing Higher-Layer Features of a Deep Network* (University of Montreal Technical Report 1341). Montreal: University of Montreal.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., … Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

Gal, E., Geller, A., Fraser, G., Levi, Z., & Niv, Y. (2008). Assessment and validation of the new Capsule Endoscopy Crohn's Disease Activity Index (CECDAI). *Digestive Diseases and Sciences, 53*(7), 1933–1937.

Gatoula, P., Dimas, G., Iakovidis, D. K., & Koulaouzidis, A. (2021). Enhanced CNN-based gaze estimation on wireless capsule endoscopy images. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems, Aveiro (CBMS)* (pp. 189–195). IEEE.

Gerhard, H. E., Wichmann, F. A., & Bethge, M. (2013). How sensitive is the human visual system to the local statistics of natural images? *PLoS Computational Biology, 9*(1), e1002873.

Guo, S. S., Zhang, R., Liu, B., Zhu, Y., Ballard, D., Hayhoe, M., … Stone, P. (2021). Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems, 34*, 25370–25385.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas (pp. 770–778).

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). Natural image statistics: A probabilistic approach to early computational vision. *Computational Imaging and Vision* (Vol. *39*). Springer Science & Business Media.

Iddan, G., Meron, G., Glukhovsky, A., & Swain, P. (2000). Wireless capsule endoscopy. *Nature, 405*(6785), 417.

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications, 12*(1), 1872, doi:10.1038/s41467-021-22078-3.

Kümmerer, M., & Bethge, M. (2021). State-of-the-art in human scanpath prediction. *arXiv preprint* arXiv:2102.12239.

Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J., & Shao, L. (2020). Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia, 23*, 2086–2099.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron, 93*(2), 451–463.

Li, Y., Liu, M., & Rehg, J. M. (2020). In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 6731–6747.

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision, 120*, 233–255.

McAlindon, M. E., Ching, H.-L., Yung, D., Sidhu, R., & Koulaouzidis, A. (2016). Capsule endoscopy of the small bowel. *Annals of Translational Medicine, 4*(19), 369, https://doi.org/10.21037/atm.2016.09.18.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems,* Montréal, 27.

Muddamsetty, S. M., Jahromi, M. N. S., & Moeslund, T. B. (2021). Expert level evaluations for explainable AI (XAI) methods in the medical domain. In *International Conference on Pattern Recognition,* Taiwan Cham: Springer International Publishing (pp. 35–46).

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*, 2397–2416.

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise Randomized input sampling for explanation of black-box models. *arXiv preprint* arXiv:1806.07421.

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2023). Explanation strategies for image classification in humans vs. current explainable AI. *arXiv preprint* arXiv:2304.04448.

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., . . . Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Medicine, 15*(11), 1–17.

Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021). Human attention in fine-grained classification. *British Machine Vision Conference,* Aberdeen.

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision, 19*(3), 1, https://doi.org/10.1167/19.3.1.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). Italy: Venice, doi:10.1109/ICCV.2017.74.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Sharma, S., Kiros, R., & Salakhutdinov, R. (2015). *Action recognition using visual attention*. CoRR, abs/1511.04119.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., . . . Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science, 362*(6419), 1140–1144.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. CoRR, abs/1312.6034.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR May 7–9, 2015, Conference Track Proceedings*. San Diego, CA, USA.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron, 103*(6), 967–979.

Sood, E., Kögel, F., Strohm, F., Dhar, P., & Bulling, A. (2021). Vqa-mhug: A gaze dataset to study multimodal neural attention in visual question answering. *arXiv preprint* arXiv:2109.13116.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. A. (2015). Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR, May 7–9, 2015, Workshop Track Proceedings*. San Diego, CA, USA.

Tanner, J., & Itti, L. (2019). A top-down saliency model with goal relevance. *Journal of Vision, 19*(1), 1–16.

Tavakoli, H. R., Ahmed, F., Borji, A., & Laaksonen, J. (2017). Saliency revisited: Analysis of mouse movements versus fixations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6354–6362). Honolulu, HI, USA, doi:10.1109/CVPR.2017.673.

Thakoor, K., Koorathota, S., Hood, D., & Sajda, P. (2020). Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images. IEEE Dataport, https://dx.doi.org/10.21227/qg30-1p45.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., . . . Hu, X. (2020). Score-cam score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 24–25). Seattle, WA, USA.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19). Munich Germany.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., . . . Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the 32nd International Conference on Machine Learning, in Proceedings of Machine Learning Research, 37*, 2048–2057.

Yang, D. H., Keum, B., & Jeen, Y. T. (2016). Capsule endoscopy for crohn's disease: Current status of diagnosis and management. *Gastroenterology Research and Practice,* 2016.

Appendix A: Performances of each participant on the classification task

Participant ID	Accuracy (%)	Specificity (%)	Sensitivity (%)
Novices			
Random people 7	44.3	24.4	73.1
Random people 4	55.1	66.9	38.2
Random people 3	58.6	47.9	74.0
Random people 5	61.9	39.3	94.1
Junior doctor 4	64.9	53.2	81.0
Senior doctor 4	71.7	49.6	100
Random people 1	69.3	52.4	93.9
Junior doctor 1	69.7	49.7	98.0
Random people 2	69.8	77.1	59.4
Random people 6	73.2	78.6	65.3
Junior doctor 2	76.1	73.2	80.2
Experts			
Senior doctor 2	80.8	69.3	97.0
VGG16	**81.4**	**98.6**	**59.4**
VGG19	**82.2**	**97.5**	**60.6**
Senior doctor 8	93.7	96.8	89.7
ResNet34	**84.1**	**98.1**	**66.6**
Senior doctor 5	89.9	87.9	92.8
Senior doctor 9	92.1	91.4	93.0
Junior doctor 5	90.2	94.4	84.2
Senior doctor 6	93.4	89.4	99.0
Senior doctor 7	93	95.9	88.8
Senior doctor 10	91.5	93.1	89.2
Senior doctor 3	93.9	97.9	88.1
Senior doctor 1	94.7	93.1	97.1

Table A1. Participants and networks image classification performance (pathologic or nonpathologic). Participants (and networks) were assigned to the expert group if they had at least 80% accuracy. Otherwise, they were assigned to the novice group. Values in bold correspond to the performance of deep neural networks.

Appendix B: Attention between humans with CC

We ran the same analysis with correlation coefficient (CC) as for NSS and found similar results. We found a significant effect of image label, $t(10{,}776) = 12.298$, $p < 0.0001$, and of its interaction with the expertise level of observers, $t(10{,}776) = -11.813$, $p < 0.0001$. The effect of expertise was not significant, $t(10{,}776) = 0.86$, $p = 0.38$. When we computed the same linear mixed model on only pathological images, we found a significant effect of the expertise: $t(4{,}442) = 12.298$, $p = 0.05$.

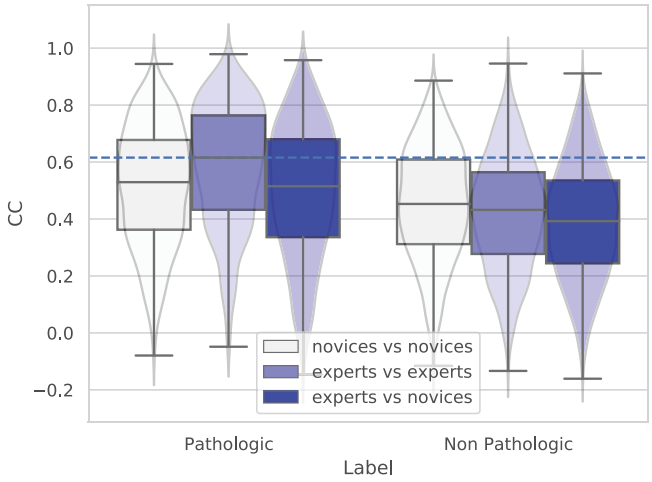


Figure B1. Illustration of the distribution of CC over human attention comparison, grouped by labels. The "novices vs. novices" label corresponds to the intranovice comparison, scoring at what point novices are looking at the same image areas. "Experts vs. experts" is the same as the "novices vs. novices" one inside the expert group. The "experts vs. novices" label corresponds to the intergroup comparison, scoring at one point experts and novices glimpse the same areas.

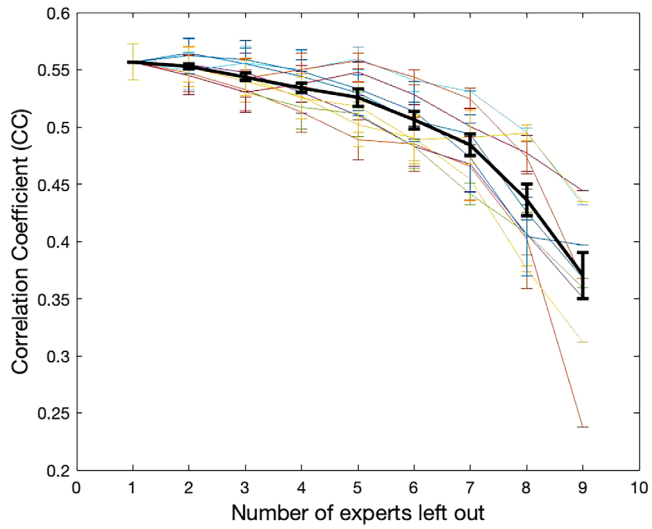
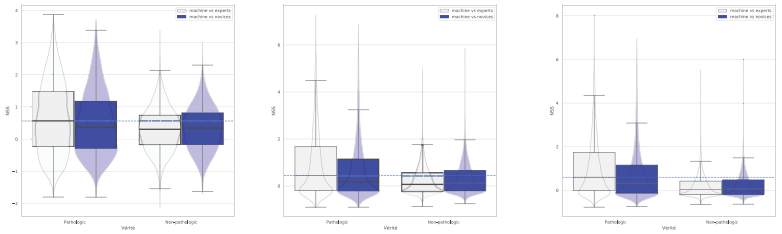


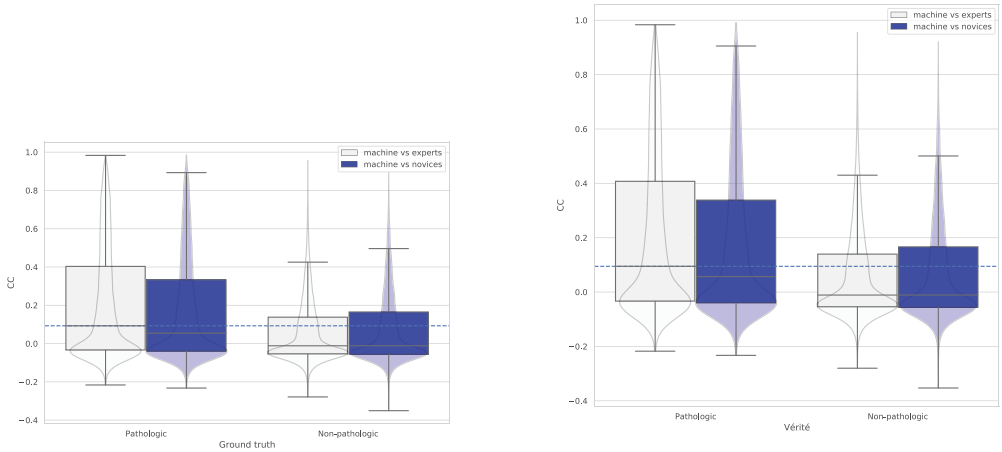
Figure B2. Effect of leaving M human experts out on interobserver consistency. The 10 colored lines correspond to different random permutations of the left-out experts. The black line corresponds to the average over the 10 colored lines. Error bars correspond to the standard errors.

Appendix C: Artificial vs. human attention for different methods for different networks with NSS and CC

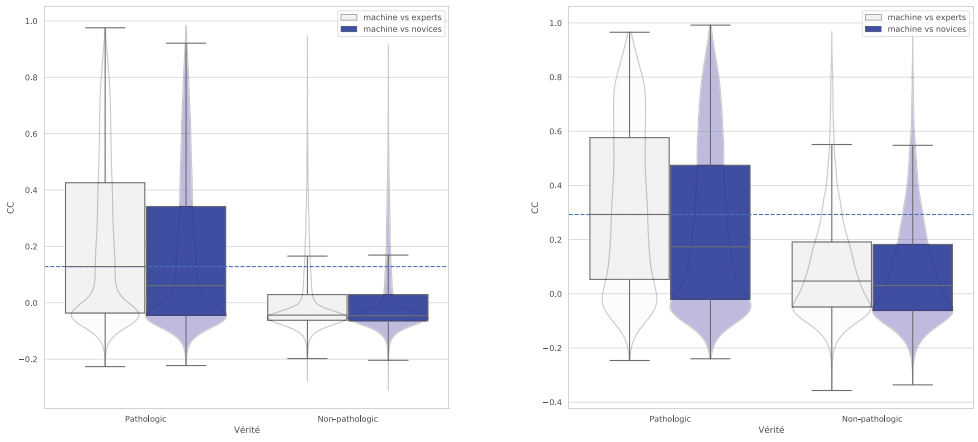


(a) GradCAM (b) Guided GradCAM (c) Guided backpropagation

Figure C1. Comparison of artificial attention extracted with different methods with human attention as a function of expertise level and image label. These results were obtained for the ResNet34 network and the comparison was performed using the NSS metric. The higher the NSS score, the closer the areas of attention between machine and humans are.

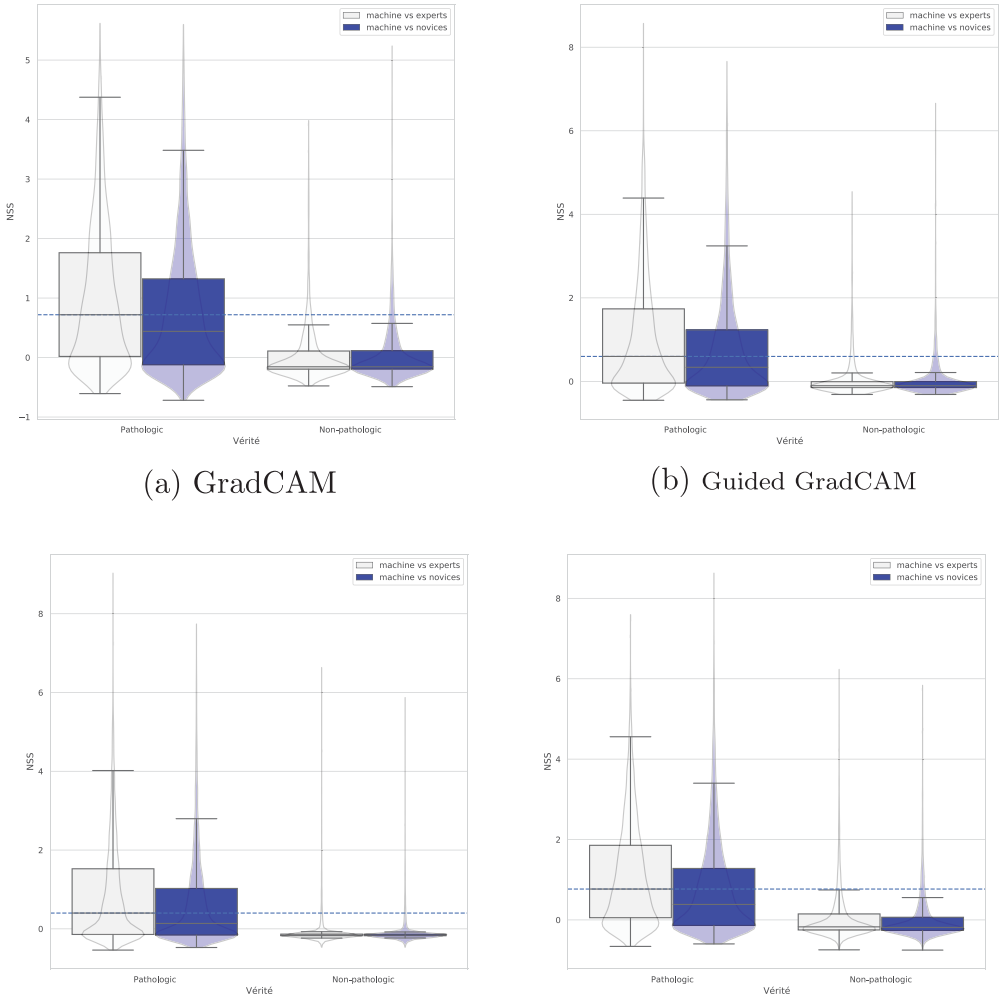


(a) GradCAM (b) Guided GradCAM



(c) Guided backpropagation (d) Vanilla gradients

Figure C2. Comparison of artificial attention extracted with different methods with human attention as a function of expertise level and image label. These results were obtained for the VGG16 network, and the comparison was performed using the CC metric. The larger the CC score, the closer the areas of attention between machine and human are.



(a) GradCAM (b) Guided GradCAM

(c) Guided backpropagation (d) Vanilla gradients

Figure C3. Comparison of artificial attention extracted with different methods with human attention as a function of expertise level and image label. These results were obtained for the VGG19 network, and the comparison was performed using the NSS metric. The larger the NSS score, the closer the areas of attention between machine and human are.

Appendix D: Post hoc attention methods description

The goal of post hoc attention extraction algorithms is to allow visualization of the parts of the input that lead to the final decision, focusing on the information contained in the model. This type of attention is extracted post hoc, the supervised learning algorithm having already been trained on a dataset beforehand. These methods do not require retraining and additional optimization steps. Three types of methods can be found: perturbation methods, back-propagation methods, and methods based on the activation of feature maps.

GradCAM

GradCAM is an algorithm based on the work of Erhan et al. (2009) and Mahendran and Vedaldi (2016), which shows that the deeper the convolutional layer (thus close to the output layer; i.e., the softmax layer for a classification task), the more high-level visual constructs it captures. Thus, by focusing on the information captured by the last convolutional layers, it extracts semantic information specific to each of the classes and, by the spatial nature of the convolution operation, determines the attention areas of a deep neural network. To do this, it uses the information that flows in the last convolution layer: the activation maps obtained during propagation (see Figure D2) and the gradients obtained during back-propagation.

The gradients are used to compute the importance of each of the activation maps A_k by computing the scalars α_k^c. This coefficient is specific to a class c and an activation map k. Thus, we compute the gradient of the output y^c before the softmax layer with respect to the activation map A^k. The obtained gradients are then globally averaged (*global average pooling* in English) as described in the Figure D1 and in the Equation 1:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{Global average}} \overbrace{\frac{\partial y^c}{\partial A_{ij}^k}}^{\text{Gradient}}. \quad (1)$$

Once the importance coefficients α_k^c of the activation maps A^k are computed, the map $\mathcal{L}_{GradCAM}^c$ for a given class c can be obtained by performing a weighted combination between these activation maps and their importance coefficient. In order to obtain only those maps that contribute positively to the target class c, a ReLU is applied to this weighted combination. Thus, through the use of a ReLU, only those features that contributed to increase the y^c output are captured

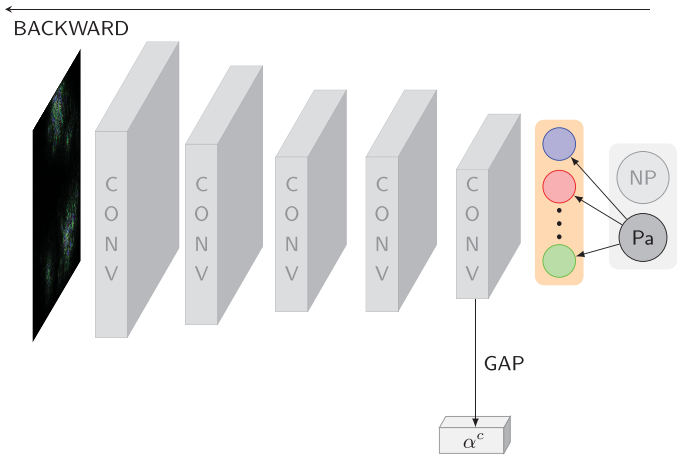


Figure D1. Diagram of the gradient maps obtained during the back-propagation, allowing the calculation of the coefficients of importance α_k^c.

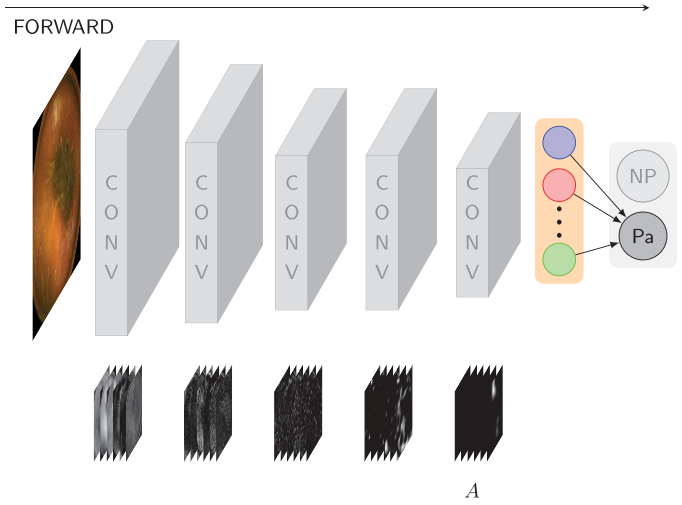


Figure D2. Diagram for obtaining feature maps during propagation.

while ignoring those that decreased it. The features that contribute negatively to a class are too likely to belong to another class. This results in the following Equation 2:

$$\mathcal{L}_{GradCAM}^c = \overbrace{ReLu}^{\text{Only positive influence}} \left(\sum_k \overbrace{\alpha_k^c}^{\text{Importance coefficient}} \overbrace{A_k}^{\text{Features maps}} \right). \quad (2)$$

Thus, each map will represent the weighted combination of activations that contributed positively to the final decision. The coefficients α_k^c represent a partial linearization of the network downstream of A

and capture the importance of a feature map k for a given target class c.

The size of the resulting maps thus depends on the input size and the depth of the target feature volume. Indeed, these maps have the same resolution as the activation maps that allowed their computation and are therefore smaller when the target volume is deep due to the use of max-pooling in deep convolution neural networks. In order to visualize the attention areas on the original image, they are often rescaled by bilinear interpolation, which induces that their attention areas are coarser.

Back-propagation algorithms

In the back-propagation algorithms, we can mention the two most famous ones, close to each other in the idea, but which have some differences: the so-called gradients algorithm (also called simple back-propagation) and the guided back-propagation algorithm (Springenberg et al., 2015). These methods apply like the other post hoc attention extraction techniques to already trained networks and do not change their weights. They are based on the same idea: to compute the gradient of the network's prediction with respect to the input by keeping the weights fixed. This allows one to determine which input elements (e.g., which pixels in the case of an input image) should be changed the least to affect the prediction the most. The difference between these three methods lies in how the gradients are calculated and how the back-propagation through the network takes place.

The idea of these methods is to find the weight matrix of the network w_c associating to each pixel of an input I a score function $S_c(I)$. Thus, we can define with b_c the biases of the network:

$$S_c(I) = w_c^T I + b_c.$$

The problem with deep neural networks is that the score function equation $S_c(I)$ is a highly nonlinear function of I. However, using a first-order Taylor expansion, in the neighborhood of a given image I_0, we can approximate the previous equation as follows:

$$S_c(I) \approx w_c = \frac{\partial I}{\partial S_c}\bigg|_{I_0}^T I + b,$$

with w the drift of S_c with respect to I at the operating point I_0.

According to Simonyan et al., 2013: "Another interpretation of calculating image-specific class salience using the derivative of the class score is that the magnitude of the derivative indicates which pixels need to be modified the least to affect the class score the most. These pixels can be expected to correspond to the location of the object in the image." This can also be seen as the pixels with the largest magnitude of the derivative of the score function with respect to a given image I_0 are the ones with the most influence on the final decision and thus the ones that allow the network to make its decision.

Gradients

In the gradient method, no changes are applied to the network, and the gradients are back-propagated to the input layer. Once the gradient map is obtained, we calculate the absolute value element by element of this map of the same dimension as the input image. If the image has three channels (RGB), we take the maximum of each channel.

Guided back-propagation

In the guided back-propagation method, as in GradCAM, we wish to observe only the positive influence of the weights on a class to avoid the gradients of another class interfering with the target class. For this purpose, the architecture of the trained convolution network is slightly modified. Only the positive influence of the input pixels on the target class c is kept. The gradients of the output with respect to the input are computed by ignoring all negative gradients. This translates into the implementation of the method by applying a ReLU on the gradients during back-propagation. The ReLUs of the model are exchanged with guided back-propagation ReLUs. These ReLUs work in a very similar way to the normal ReLUs during propagation but have the particularity of also applying to the gradient during back-propagation. Thus we obtain a map of the pixels contributing positively to a target class. We recall here that although the architecture of the network is modified by replacing the ReLus with guided back-propagation ReLus, the weights and biases remain unchanged.

Thanks to this algorithm, we obtain a saliency map, specific to the target class c with the same dimensions, and the same number of channels as the input image. If the input image has three channels (RGB), we take the maximum value of the gradient map on the three channels.

Guided GradCAM

The guided GradCAM algorithm is, as the name suggests, a combination of the guided back-propagation algorithm and GradCAM. Thus, to compute the importance coefficients α_k^c in the target layer k for a given class c, only positive gradients are used in the back-propagation by applying a ReLU to the error signal as in the guided back-propagation algorithm.

The map computed using the target layer activation maps weighted by the α_k^c importance coefficients will be used as a mask for the map obtained by guided back-propagation. This technique has the advantage of taking into account the activation maps computed during the propagation while keeping a good resolution since the final map is of the same size as the input map as in the guided back-propagation technique.

Score-CAM

Wang et al. (2020) addressed two issues related to the gradient-based methods GradCAM when explaining deep image classification networks. The first problem is the presence of noise in saliency maps caused by gradient saturation/evanescence. The second problem is the potential overemphasis on specific feature maps. To overcome these challenges, the authors introduced Score-CAM as an alternative approach (Wang et al., 2020). Score-CAM assigns weights to individual feature maps based on the class score observed when masking the areas activating the maps. As with GradCAM, the saliency map is computed by a pondered mean of the feature maps:

$$\text{Score-CAM}(x) = \sum_k w_{kc}^s \times f^k(x), \quad (3)$$

where x, w_{kc}^s, and $f^k(x)$ are respectively the input image, the weight of feature map k for class c, and the feature map k. The weights are computed as follows:

$$w_{kc}^s = y_c(x \cdot Norm(Upsample(f^k))) - y_c(x_b), \quad (4)$$

where $y_c(x)$ is the score of class c with input x, and x_b is a baseline input (e.g., a uniform black image). The function *Upsample* resizes the feature map f^k to match the resolution of the input image x, and the *Norm* operator is a min–max normalization that sets the minimum and maximum values of the feature map respectively to 0 and 1.

RISE

The method called RISE was proposed by Petsiuk et al. (2018). This method first involves dividing the image into a rectangular grid $H' \times W'$ and calculating Q random binary masks $m^q \in \{0, 1\}^{H' \times W'}$, where $m_{i'j'}^q \sim$ Bernoulli(0.5) and $q \in [1, ..., Q]$. Then, Q inferences are calculated by applying a different mask to the image each time:

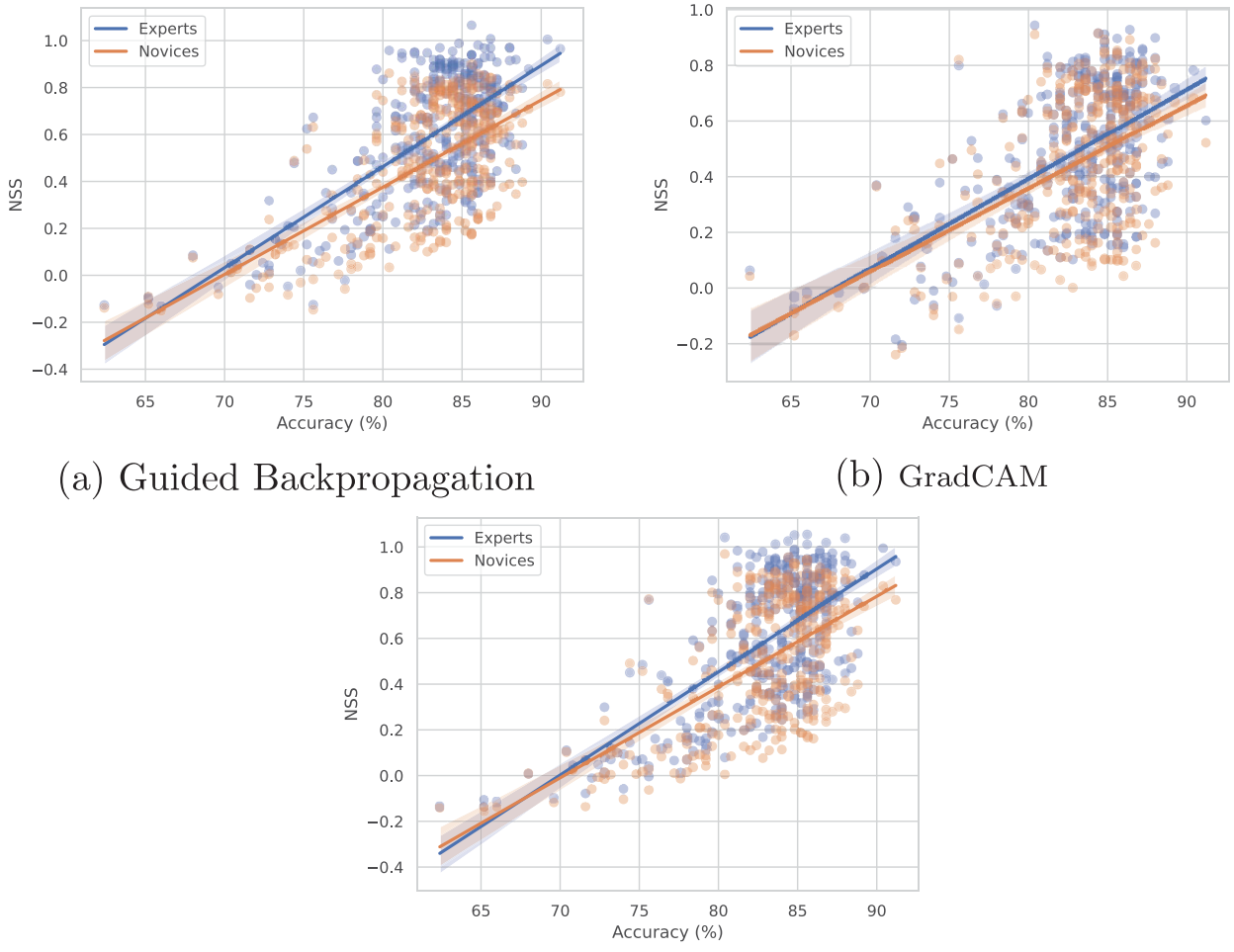
$$c_q = y_c(x \cdot Upsample(m^q)), \quad (5)$$

where the operator *Upsample* increases the resolution of the mask using nearest-neighbor interpolation to match the resolution of the input image and $y_c(x)$ is the score of class c with input x. This procedure of masking the image and measuring the score variation is repeated several thousand times in practice to estimate which areas of the image, when not masked, lead to the largest class score and thus are the most important for the decision. The final salience map is a weighted average of the masks, where the weights are the corresponding class scores:

$$S = \frac{1}{Q} \sum_q c_q \times m_q. \quad (6)$$

In this work, we apply RISE on a ResNet architecture and choose the same mask number used by the authors of the original RISE paper, which is $Q = 8{,}000$. We tried two mask resolutions: $H' \times W' = 7 \times 7$ and 14×14. We observed that the granularity of the 7×7 resolution is insufficient to highlight small objects like Crohn's lesions and therefore show only results with the 14×14 resolution.

Appendix E: Artificial attention and human attention through network training



(a) Guided Backpropagation

(b) GradCAM

(c) Guided GradCAM

Figure E1. Comparison between artificial attention and human attention as a function of the network accuracy. Artificial attention is obtained with three different post hoc methods: guided back-propagation, GradCAM, and guided GradCAM on three different networks on ResNet34. Results with VGG19 and VGG16 are similar.

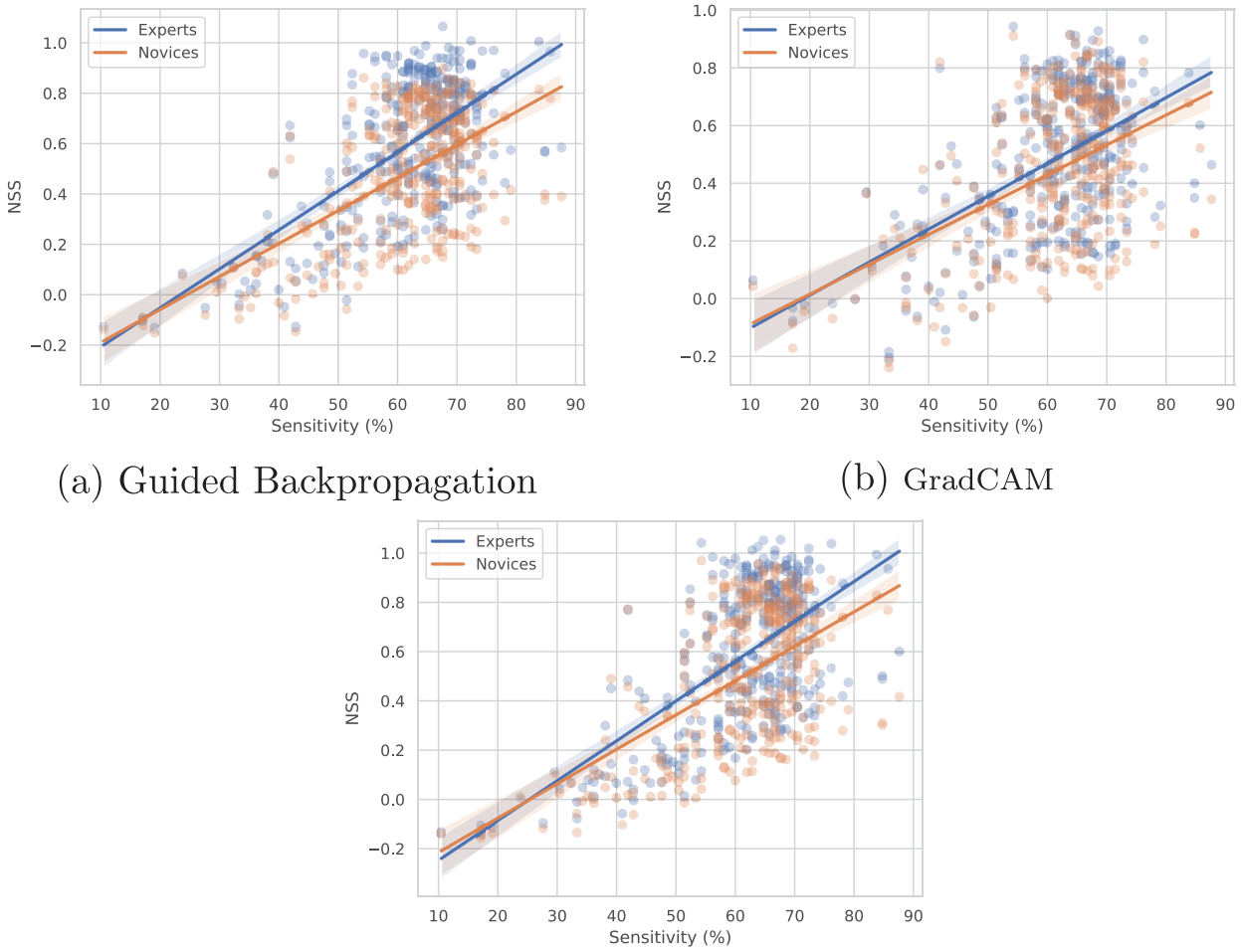
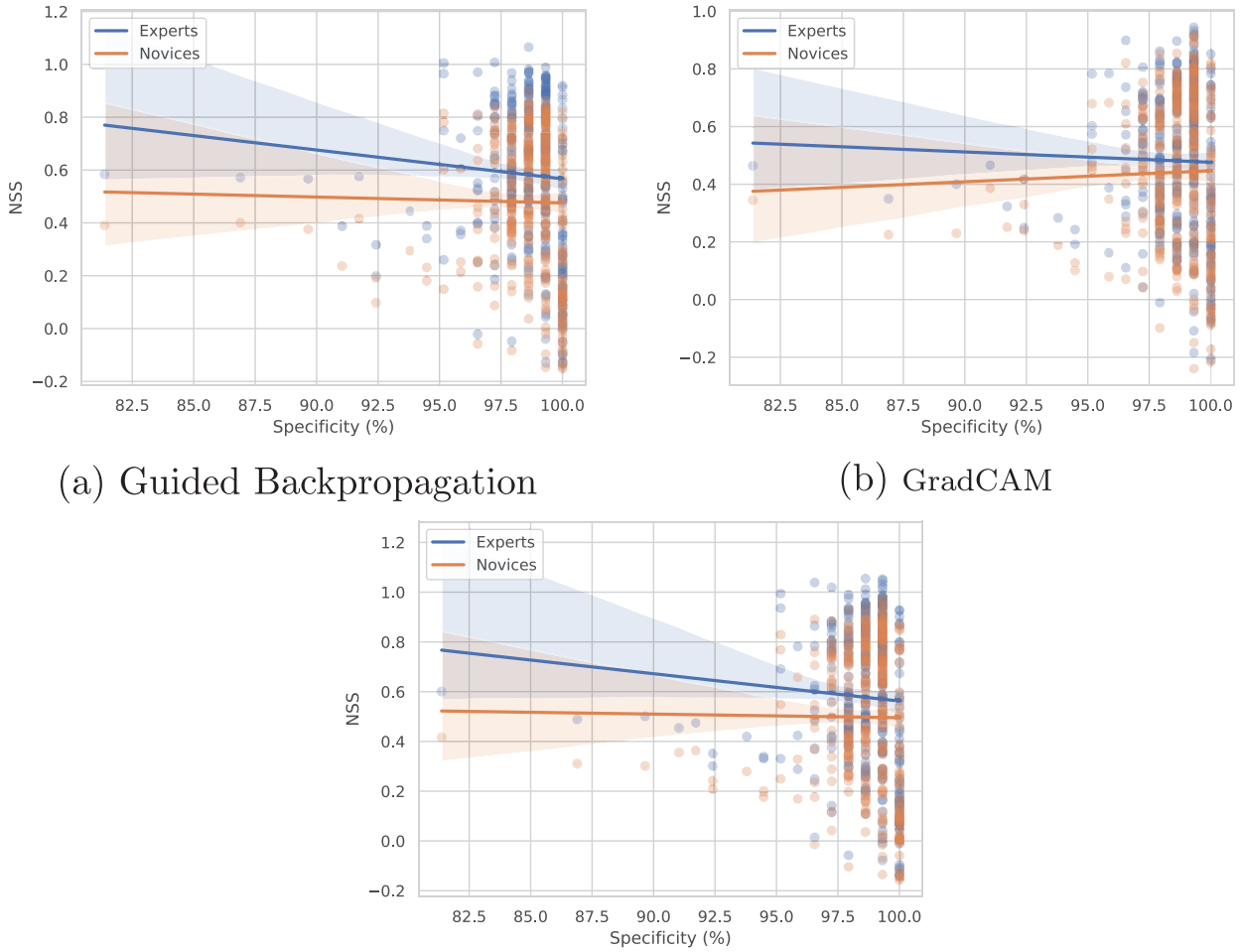


Figure E2. Comparison between artificial attention and human attention as a function of the network sensitivity. Artificial attention is obtained with three different post hoc methods: guided back-propagation, GradCAM, and guided GradCAM on three different networks on ResNet34. Results with VGG16 and VGG19 are similar.



(a) Guided Backpropagation

(b) GradCAM

(c) Guided GradCAM

Figure E3. Comparison between artificial attention and human attention as a function of the network specificity. Artificial attention is obtained with three different post hoc methods: guided back-propagation, GradCAM, and guided GradCAM on three different networks on ResNet34. Results with VGG16 and VGG19 are similar.

Appendix F: Stability experiment

Tables F1 and F2 summarize the stability results for the CC metric for nonpathological and pathological images, respectively. The results are consistent with the ones obtained with NSS. Although for equivalent networks, the stability results for the different methods are fluctuating, it is clear that the gradient method is more stable than the other methods on the pathological images. A notable difference in stability is also present between pathological and nonpathological images. Although trained differently, the networks tend on average to focus on closer areas when the image is pathological. This result seems logical because the pathological images contain localized features symptomatic of Crohn's disease.

Method	VGG16	VGG19	ResNet34
GradCAM	**0.32 (0.30)**	**0.44 (0.29)**	0.20 (0.36)
Guided GradCAM	0.18 (0.26)	0.26 (0.26)	0.25 (0.28)
Guided back-propagation	0.18 (0.330)	0.121 (0.238)	**0.44 (0.23)**
Gradients	0.16 (0.26)	0.15 (0.32)	0.37 (0.29)

Table F1. Summary table of the scores obtained with Pearson's correlation coefficient (CC) between the attention maps of the images **nonpathological** of the same trained network for different distributions of the training and validation set.

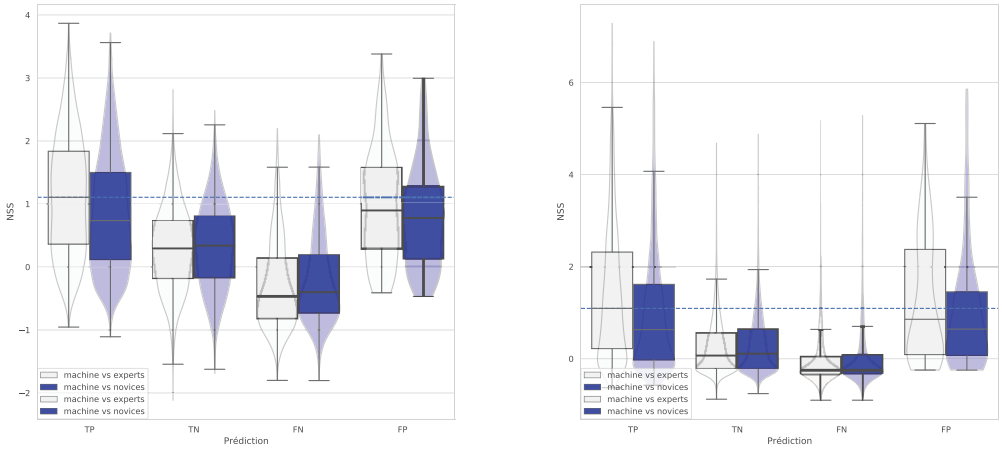
The results presented here seem to indicate a different attentional behavior of deep neural networks on pathological and nonpathological images. The lesions of Crohn's disease, whose identification is necessary for image classification, homogenize the attentional behaviors of the different networks.

Although the influence of the label is visible on the stability of the attentional behaviors of the deep neural networks, we observe that for the same network, the attentional behaviors rendered by the different methods are different. Although the weights of the network are perfectly similar, the different methods do not give us the same attentional areas for the same decision.

Method	VGG16	VGG19	ResNet34
GradCAM	0.25 (0.34)	0.35 (0.39)	0.42 (0.54)
Guided GradCAM	0.37 (0.39)	0.44 (0.37)	0.46 (0.42)
Guided back-propagation	0.42 (0.39)	0.50 (0.34)	0.56 (0.28)
Gradients	**0.60 (0.35)**	**0.59 (0.35)**	**0.69 (0.27)**

Table F2. Summary table of the scores obtained with Pearson's correlation coefficient (CC) between the attention maps of the images **pathological** of the same trained network for different distributions of the training and validation set.

Appendix G: Comparison between human and machine attention across the image classification confusion matrix



(a) GradCAM (b) Guided GradCAM

(c) Guided Backpropagation (d) Gradients

Figure G1. Comparison of different post hoc artificial attention extraction methods with human novice and expert attention as a function of image labels. The abbreviations TP, TN, FN, and FP respectively correspond to true positives, true negatives, false negatives, and false positives. These results show us that only the gradient method can account for the fact that when the algorithm makes a prediction error, its attentional behavior is close to that of humans who correctly classified the image. This indicates that the errors of the networks could mostly be recognition or diagnostic errors and not detection errors.