Global explainability in spatially aligned image modalities

Justin Engelmann [1 2] **Amos Storkey** [3 *] **Miguel O. Bernabeu** [2 *]

Abstract

We introduce the concept of spatially aligned image modalities which is common in medical imaging and might be particularly useful for global model explanations. In a spatially aligned image modality each pixel position corresponds to a similar relative position on the imaged object across all images. We introduce the concept of a spatially aligned global explanation (SAGE) that explains a model globally rather than per image. Concretely, we propose spatial averaging of per-image explanations as a simple method of producing SAGEs as well as Progressive Erasing Plus Progressive Restoration (PEPPR) for quantitatively validating that the SAGEs faithfully reflect how the examined DL model works. We conduct experiments on a dataset for disease detection in retinal images with an ultrawide field-of-view. We find that the SAGEs generally match domain knowledge and give insight into what patterns of pathology the model focuses on. We further introduce an artificial "shortcut" signal into the data and find that this can be detected by examining the SAGEs. We hope that these methods may aid researchers to examine DL models for possible failure modes and to generate new insight about the underlying data.

1. Introduction

Deep learning (DL) models provide excellent performance on many computer vision problems. Accordingly, they have risen in popularity and are increasingly used in practice. This includes critical applications like healthcare where it is very important to understand and validate DL models. Unfortunately, DL models are black boxes that are inherently hard to explain, especially compared to traditional approaches such as using linear models or small decision trees with hand-crafted image features. DL models typically have millions of parameters and complex architectures with many non-linearities (Choo & Liu, 2018). Thus, there is a pressing need for explainability. A number of approaches for per-image explanations have emerged that can help practitioners understand why a model made a specific prediction for a given image. This allows to validate the model's prediction and to guide an expert's attention to regions of interest.

In applications where experts are present and we want to understand and validate a particular prediction, per-image explanations are of great value. They can, for instance, allow a clinician to understand whether a model has been misled by a spurious pattern or spotted real pathology the clinician had missed. However, in other settings it might be too labour-intensive (e.g. early disease screening) or slow (e.g. autonomous vehicles) to have each individual prediction validated by an expert. Instead of explaining the model per image, we would like to explain it globally so that we can validate that it behaves in a sensible way that matches our understanding of the domain and does not show signs of leveraging undesirable data artefacts (Roberts et al., 2021), also known as "shortcuts" (Geirhos et al., 2020; DeGrave et al., 2021). Such a global explanation could give us confidence that the model is generally working correctly and thus lessens the need for examining per-image explanations for each new prediction.

In this work, we focus on spatially aligned image modalities where each pixel position corresponds to a similar relative position on the imaged object across all images. For example, in a dataset of cropped and centred portraits, noses, eyes and ears will be at similar positions for all images. Such spatial alignment is uncommon for natural images, but common in medical imaging where it can be a natural by-product of the data acquisition process (e.g. retinal imaging or chest X-rays) or achieved through explicit registration (e.g. in brain magnetic resonance imaging). We propose to leverage this alignment to generate Spatially Aligned Global Explanations (SAGEs). In domains where we have good understanding of which regions of the images should be informative, we can examine the SAGEs to validate that the examined DL model works in a sensible fashion that

[*]Equal last authors [1]UKRI CDT Biomedical AI, School of Informatics, University of Edinburgh, United Kingdom [2]Centre for Medical Informatics, University of Edinburgh, United Kingdom [3]School of Informatics, University of Edinburgh, United Kingdom. Correspondence to: Justin Engelmann <justin.engelmann@ed.ac.uk>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2022 by the author(s).

matches our domain knowledge. Additionally, the SAGEs might also be used for knowledge discovery. Finally, we propose Progressive Erasing Plus Progressive Restoration (PEPPR) as a way to quantitatively verify that these global explanations faithfully reflect how the model makes its predictions and to investigate which image regions contain information about the target variables.

1.1. Explainability - What, why, and a brief taxonomy

While the concept of an explanation itself is philosophically complex (Bromberger, 1992; Thagard, 1978), explainable AI is concerned with explaining models to increase our understanding of how they work. Depending on who is addressed, different explanations will be suitable. In addition to increasing someone's understanding of a model objectively, another goal of explainable AI is often to also subjectively to increase their trust and acceptance of the model. This is a legitimate aim so long as we generate well-founded, appropriate trust. Explanations can be misleading and generate unfounded trust (Lakkaraju & Bastani, 2020), and this must be avoided.

In the aim of explaining "how a model works", many subtly different questions are tangled up. Each relates to a different sense of explainability and calls for a different method. For the present work, we primarily distinguish two types of explainability: local and global. Local explainability is concerned with explaining the model's prediction for a specific input, i.e. the question of *"What in this particular image does the model consider evidence for its prediction?"* As a brief note on terminology: we consider "local" to be the natural counterpart to "global" but it could also be understood to refer to a specific region of the data manifold. Thus, we instead use the term "per-image" to avoid this potential confusion. In computer vision, per-image explainability is commonly accomplished through generating so-called saliency heatmaps that highlight image regions that were key for the model's prediction. Such per-image explanations are particularly useful if critical decisions are to be based on the model's prediction. For example, if the model is used to assist a clinician in assessing medical images and the model predicts the presence of disease in a scan. Here, a per-image explanation allows the clinician to try to comprehend the model's prediction and might draw their attention to regions of possible pathology.

Global explainability, on the other hand, is concerned with explaining how the model generally works, i.e. the question of *"What does the model tend to focus on when making a particular kind of prediction?"* Global explanations can help us understand the model generally which allows to validate that it works in a desirable fashion and is thus suitable for being applied in practice. For instance, if the model generally focuses on image features that contain information about the target variable, this would suggest that it works as desired; whereas if it focuses on features that should not be informative about the target variable, this would indicate that the model might be relying on undesirable shortcut artefacts. These artefacts are informative in the training data but might be uninformative or simply not present during inference, leading to severely degraded model performance (Geirhos et al., 2020; DeGrave et al., 2021). Manually assessing many per-image explanations can build towards global explanations. However, this approach is labour intensive and examining all images is not feasible for modern image datasets that can contain anywhere between tens of thousands and hundreds of millions of images (Sun et al., 2017). Conclusions that are drawn from examining a small number of examples, however, are susceptible to biases such as confirmation bias. Real world datasets and per-image explanations can both be noisy in their own right. This could then lead to an actual yet unexpected pattern being dismissed as being spurious, whereas a spurious yet expected pattern is accepted.

The distinction between per-image (or local) and global explainability is the most relevant to the present work. Additionally, the explainability methods we consider here are all post-hoc rather than ante-hoc in the sense that they explain a model after it has been trained. However, global explainability as we are considering it here can be used to explain the model generally and thus increase our confidence that it will work correctly before new instances are observed during inference. While per-image explanations are also very useful in practice to validate a particular model decision, not all use-cases of DL models might have an expert present at inference time. Another distinction is between model-centric and data-centric approaches. We focus on validating a particular DL model in this work and thus on model-centric explanations. However, data-centric explainability is also useful as it can help validate a dataset and allow for knowledge discovery. In the next section, we briefly explain how our methods could be easily modified to be data-centric instead.

1.2. Spatially aligned image modalities & spatial averaging of per-image explanations

An image modality is aligned if each pixel/voxel corresponds to the same or similar relative position of the imaged object and thus samples implicitly share a common coordinate system. Natural image datasets are rarely aligned, but fortunately such alignment is common in medical imaging where the need for model validation and potential for knowledge discovery is particularly large. Figure 1 illustrates this for ultra-widefield retina images. After flipping all right eyes horizontally, the images are aligned such that relevant regions of the retina appear share coordinates across images, including but not limited to the visually apparent

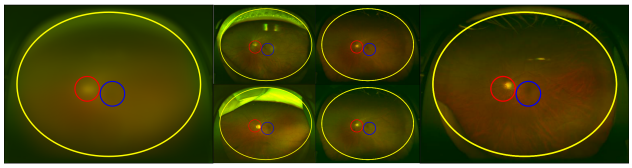

Figure 1. **Example of an aligned image modality.** In these ultra-widefield retina images, each pixel corresponds to a similar position of a human retinas. The coloured ellipses at identical coordinates in each image indicate the approximate positions of: The retina itself, outside of this region no relevant information should be found; the optic disc, where blood vessels go through the retina, visible as a bright spot; the fovea, a small pit responsible for the sharpest vision, visible as a dark spot. **Left:** The pixel-wise average of 1,664 validation set images. We can make out the optic disc clearly and the fovea faintly (when zoomed in), indicating that these images are well-aligned. **Middle and Right:** Example validation images.

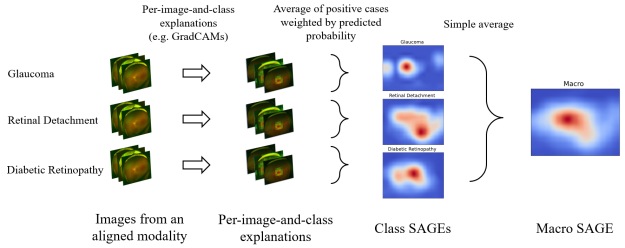

Figure 2. **From per-image to global explainability through aggregation.** In an aligned image modality, the pixel-wise aggregates of per-image explanations can be used as global explanations.

landmarks indicated in Figure 1. No further preprocessing or registration is done to sptially align the images, apart from simply flipping right eyes horizontally. Image modalities with fixed reference frames like camera data from a self-driving car might also exhibit alignment. Even modalities that are not intrinsically aligned could be aligned through post-processing. For instance, we could first use a model to detect objects, crop them, and then input these objects into an object identification model, which might then deal with aligned inputs.

We propose to leverage such alignment by pixel-wise spatial averaging of per-image explanations into SAGEs. As the same pixel position corresponds to the same relative position of the imaged object, these averaged per-image explanations should then highlight which regions of the input images are generally used by the model to make its predictions. This is then a measure of where in the input images information about the target variables is found that is also used by the examined model. In principle, we could also extend this approach by then averaging SAGEs across different models to move from model-centric to data-centric global explanations. This would ask the question "What regions of the images contain relevant information about the target variable?" In the present work, we focus on examining a particular fitted model and thus model-centric explanations, asking the question "What regions of the images does the model extract relevant information about the target variable from?"

2. Methods

2.1. Per-image explanations

For a given fitted model f with model parameters θ, an per-image-and-class explanation method takes an input image X_i belonging to sample i and a target class c as inputs and yields an explanation E_i^c for the model's predictions of the target class for this image $\hat{p}_i(c) = (f_\theta(X_i))_l$. The explanation E_i^c is a matrix of pixel-wise importance scores with the same dimensions as the input image where higher values indicate that an input pixel was more important to the model's prediction of the target label. We use Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al., 2017) to generate image- and label-wise explanations because it is a well-established method (e.g. (Matsuba et al., 2019; Nagasato et al., 2019; 2018)) and because it is generally faithful to how the explained model works. Other methods have been shown to be akin to simple edge detectors and to yield very similar heatmaps even when replacing the trained model weights with random weights (Adebayo et al., 2018a;b). Human vision is generally biased towards edges and thus merely highlighting edges on an image might appear to be a reasonable explanation of the model's prediction, even if it is not actually faithful to the workings of the model. This also motivates the need for validating the obtained heatmaps beyond manually assessing a few examples.

2.2. Spatially aligned global explanations (SAGEs)

We aggregate per-image-and-class explanations into a per-class SAGE. Specifically, we take the average of all true positive instances from a validation or test set weighted by the model's predicted probability of the target label $\hat{p}_i(l)$. We denote the number of positives instances of the target label as N_c.

$$SAGE(c) = \frac{1}{N_c} \sum_{i}^{N_c} \hat{p}_i(c)\, E_i^c$$

We weigh by the predicted probability $\hat{p}_i(c)$ as per-image explanations for examples where the model does not predict the target label with high confidence are very noisy. This is expected as explanations for these predictions are also conceptually unsound. If the model thinks that the class in

question is not present in the image, then few things if any are evidence of this class being present. Thus, we assign less weight to them. We further only include positive instances of the target label, as otherwise the global explanations for rare labels could be dominated by noisy per-image explanations where the model does not predict the target label with high confidence. Predicted probability weighting does mitigate this but not fully, especially if the DL model is not well-calibrated or if a class is particularly rare. Using a validation or test set instead of training examples avoids aggregating spurious patterns of a model that has started to overfit. Generating global explanations using the training set and comparing them to those obtained on the validation or test set could help diagnose what a model tends to overfit on but we will leave this for future work.

These label-wise global explanations $SAGE(c)$ can then be further aggregated into a macro SAGE $SAGE^{macro}$ which shows the image regions that are most important to the model generally. We take the simple, unweighted average of label-wise global explanations to preserve information about all labels even if the data is imbalanced. We denote the number of classes as $|c|$.

$$SAGE^{macro} = \frac{1}{|c|} \sum_{l}^{|c|} SAGE(c)$$

2.3. Validating SAGEs through Progressive Erasing Plus Progressive Restoration (PEPPR)

The global explanations should not only appear consistent with domain knowledge, but also reflect faithfully how our model makes its predictions. Qualitative evaluation alone can be subject to confirmation biases where explanations are accepted that do not reflect the model's actual workings (Adebayo et al., 2018a). To validate the global explanations quantitatively, we propose Progressive Erasing Plus Progressive Restoration (PEPPR). First, we threshold the overall global explanation at different quantiles to obtain a series of binary masks. We then use these masks to progressively erase the least important image regions globally until we are left with a blank image, and evaluate model performance at every step - without retraining as our goal is to explain and validate the fitted model f_θ. Then we take the inverse of these masks, starting with a blank image, and progressively restore the least important regions. This yields two curves of threshold quantile versus performance that allow us to validate whether the global explanation is faithful to the model's workings and to better understand which image regions are informative. A detailed example is presented in Section 3.4. We suggest erasing by either replacing the removed pixels by their average across the training set, or by random noise if the model was trained with RandomErasing (Zhong et al., 2020) as augmentation. Erasure-based tests have been used to validate per-image explanations (Samek et al., 2016), but to our knowledge only starting with the most important regions and moving towards less important directions. We propose doing it in both directions so that PEPPR is sensitive to duplicated information. For instance, one area of an image might be the most important in the sense that it contains the most information about the target variable, however, some of this information might be duplicated in other image regions.

2.4. Related work

A large literature on explainable AI has emerged (Gilpin et al., 2018), with many methods focusing on post-hoc explanations of black box models (Guidotti et al., 2018). Global explanations have been considered before in the context of tree-based algorithms applied to tabular data (Lundberg et al., 2019). For DL models in computer vision, a number of methods for per-image explanations have emerged (e.g. (Shrikumar et al., 2017; Selvaraju et al., 2017; Lundberg & Lee, 2017)). Some data-centric methods go in the direction of global explanations, such as learning a generative model or a cycle-consistent image-to-image translator to identify key distinguishing features of different datasets (e.g. (DeGrave et al., 2021)). In terms of explaining a fitted DL model generally, feature map visualisations (Olah et al., 2017) and image generation from classifiers (Pal et al., 2021) are two kinds of approaches towards global explainability. However, applying them to validate a model has some challenges as there are many feature maps that can be visualised and a whole space of class conditional samples that can be generated. The methods we present here are limited to spatially aligned image modality but at the same time conceptually simple and easy to implement. Our approach generates one global explanation per class, which can be further summarises into a single macro explanation.

Spatially averaging per-image explanations has been considered in other contexts. (Lapuschkin et al., 2019) present an average of per-image explanations for non-aligned natural images in passing (Fig. 1 b)). However, this serves to make the point that it necessary to "focus the explanation on the predictions of individual examples" (contrasting with traditional feature selection approaches) as these aggregates are not very helpful for images that are not aligned. (Petsiuk et al., 2021) propose a method for per-image explanations of object detection models and present some class averages of these explanations. This requires detecting the object, cropping the bounding box, and aligning and resizing all boxes to a common size. This is necessary and sensible in the context of non-aligned images and object detection. However, in an aligned image modality, we are precisely interested in where in the images it is that the patterns used by classification models occur.

Erasure to test explanations was initially proposed starting with the most important regions (equivalent to progressive restoration) (Samek et al., 2016), and then later combined with starting with the least important regions as well (Petsiuk et al., 2018). PEPPR differs from these approaches only in two minor ways: First, we erase on a global level using a single global explanation rather than erasing at the image-level using an individual explanation for each image. Second, we propose to consider class-wise metrics rather than a single aggregated measure of performance to investigate whether different classes have characteristically sensitivities to PEPPR.

3. Experiments

3.1. Data: ultra-widefield retinal images

Ultra-widefield retinal images are a particularly interesting application for explainability methods as it is a relatively new modality that has been studied far less than traditional colour fundus photography and thus holds more potential for knowledge discovery. The much larger field of view (200 degrees, compared to 30-60 degrees regular retina images) raises the question which regions of the retina are useful particularly of those that are not imaged by colour fundus photography. The scale also means that signs of pathology could be missed relatively easily by practitioners. Thus, DL and especially model explanations could add significant practical value here.

We use the Tsukazaki Optos Public (TOP) dataset (Ohsugi et al., 2017; Masumoto et al., 2018b;a; 2019), a dataset of 13,047 ultra-widefield retinal images.[1] The data was collected at Tsukazaki hospital in Himeji, Japan, between October 11, 2011 and September 6, 2018. The study was approved by the Ethics Committee of Tsukazaki Hospital (No. 191014) and the dataset is released for research use only, with commercial use being explicitly prohibited.

The dataset has labels for eight retinal diseases of which we select the three most common ones: Diabetic Retinopathy, Glaucoma, and Retinal Detachment. Using fewer diseases simplifies the discussion considerably but is not a requirement for applying our method and we find characteristically similar results when considering all eight disease. The three selected diseases have characteristically different presentation. Diabetic Retinopathy manifests itself in numerous ways, primarily microaneurysms and hemorrhages which can occur across the retina, and neovascularization which occurs primarily around the optic disc (Victor, 2019). Thus, pathology related to Diabetic Retinopathy occurs around the optic disc but is not confined to this area. Glaucoma, on the other hand, is a condition where the optic nerve is damaged and should thus be tightly localised around the optic disc where the optic nerve is situated. Finally, Retinal Detachment can occur anywhere across the retina and unlike the other two diseases should not occur preferentially around the optic disc. However, Retinal Detachment does tend to occur more often on the temporal side (on right as the images are shown here) (Shunmugam et al., 2014). Thus, we have one tightly localised disease with no signs of pathology expected elsewhere (Glaucoma), one disease that is localised with signs of pathology occurring across the retina (Diabetic Retinopathy), and one disease that is localised in a different place from the other two diseases, also with signs of pathology occurring across the retina (Retinal Detachment).

After excluding the other disease, we are left with 4,894 healthy images, 3,261 images with Diabetic Retinopathy, 2,440 with Glaucoma, and 933 with Retinal Detachment. We then randomly split the data into train, validation and test sets containing 70, 15 and 15 % of the data, respectively. We split the data on a patient- rather than image-level such that each patient occurs in exactly one of the three sets to avoid data leakage across sets. We frame the problem as multi-label classification as diseases can co-occur, using a binary target label per disease.

These experiments aim to test our proposed methodology. Please note that while we chose this modality because we are familiar with the domain of retinal imaging, the model we train is not intended for clinical application as it is presented here, nor do we intend to present concrete biomedical findings at present.

3.2. Model training and implementation details

For our experiments, we fine-tune a simple ResNet18 (He et al., 2016) using pre-trained weights from ImageNet (Deng et al., 2009) with AdamW (Loshchilov & Hutter, 2017) with a learning rate $\eta = 5 \times 10^{-5}$, exponential decay rates $\beta_1 = 0.9, \beta_2 = 0.999$ and weight decay $\lambda = 5 \times 10^{-5}$ applied to the weights of the convolutional and fully-connected layers, but not to biases or batchnorm parameters. We train to minimise the label-wise binary crossentropy loss for 5 epochs, which is sufficient to observe convergence. Training took about 7 minutes per run with a batchsize of 64 on a single NVIDIA RTX 2060 6GB.

We use RandomErasing (Zhong et al., 2020) as our only data augmentation, which with probability $p = 0.33$ randomly replaces between 5 and 40% of the image with per-pixel random noise with an aspect ratio between $\frac{1}{3}$ and 3. The total erased area is divided between up to five different areas. Besides being a generally useful augmentation, training a model with RandomErasing induces robustness to the

[1] We would like to thank Hiroki Masumoto and all his colleagues at Tsukazaki Hospital for releasing this dataset for research use. This is a great contribution to artificial intelligence research in ophthalmology.

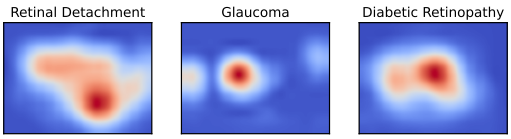

Figure 3. **Per-class SAGEs.** These explanations generally match domain knowledge. Retinal Detachment can occur across the entire retina. Glaucoma primarily affects the optic disc. Diabetic Retinopathy can present pathology across the retina but primarily affects the macula. However, for Glaucoma there is some importance at the left and right edges of the images, which is unexpected.

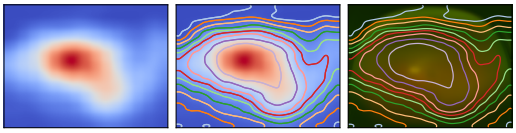

Figure 4. **Overall global explanation.** **Left:** The overall global explanation. **Middle:** The overall global explanation with contour lines indicating the most important regions in quantile steps of 10%. **Right:** The same contour lines overlaid on the average validation set image.

erasure of image regions in general. Thus, when conducting PEPPR the mere presence of such erased regions should not confuse our model and it should be able to perform well as long as informative regions are still present.

We use limited augmentations to avoid injecting spatial biases into our model. For instance, random cropping or rotations remove the edges of images and thus might encourage the model to focus more on the centre. As more and well-chosen augmentations are typically very beneficial for small datasets, using only RandomErasing also creates more challenging conditions for our methods. We chose ResNet as our DL architecture because it is well-established, efficient and performs well. We briefly experimented with larger ResNet variants but found no substantial performance benefit. We implemented our training process using the PyTorch (Paszke et al., 2019) and timm (Wightman, 2019) libraries. We use the pytorch-gradcam package for the implementation of per-image explanations (Gildenblat & contributors, 2021). We scale the SAGE for each class to the interval $[0, 1]$ but forgo scaling the individual GradCAMs so their relative magnitudes are preserved. As is common in the literature, we use bicubic interpolation when resizing the explanations form the dimensions of the final featuremap to the input image dimension.

We obtain the following label-wise Areas Under the Receiver Operating Characteristic Curve (AUCs) on the held-out test set: 0.9805 for Retinal Detachment; 0.9462 for Glaucoma; and 0.9099 for Diabetic Retinopathy. The performance of the DL model is not the focus of this work but these values represent very good model performance and thus indicate that our model training strategy was effective. As the model fits the data well, the obtained explanations should reflect the relationship between data and target labels well.

3.3. SAGE

Figure 3 shows the SAGEs we generated through probability-weighted averaging of GradCAMs for positive samples of each label as outlined in Section 2.2. We find that these generally match the domain knowledge we outlined in Section 3.1. Glaucoma is concentrated around the optic disc and has little importance allocated to other regions. Diabetic Retinopathy, too, is concentrated around the optic disc but with more importance elsewhere on the retina. Finally, the explanation for Retinal Detachment is focused on the temporal side (as it is shown here: on right) with some importance spread across the entire retina. However, these explanations also show patterns that do not match clinical evidence and thus might be signs of noise or artefacts: The explanation for Glaucoma has importance allocated to the edges of the retina, particularly to the left and right; and the explanation for Retinal Detachment has importance allocated to the bottom right corner of the images, which is an area that does not show the retina at all and thus should be uninformative. In the present work, we will use PEPPR to investigate whether the model relies on these regions. But if we wanted to apply this model in practice, then these unexpected patterns should also be investigated in more detail, for example by selecting examples where the per-image explanations have the most importance allocated to this region. This could reveal data issues, or potentially yield new domain insights.[2]

We also aggregate the class SAGEs into a macro SAGE, as shown in Figure 4. Despite some unexpected patterns for the class SAGEs, the macro SAGE also matches our domain knowledge. It correctly ranks the regions of the image that actually show the retina as generally more important than the non-retina regions. It identifies the area around the optic disc as the most important region, with a part of the temporal side also ranking highly. This matches the included retinal diseases.

[2]For example, in the case of Glaucoma, changes on the temporal (as shown here: right) side (Sihota et al., 2015) of the retina have been noted. We had been unaware of this previously before examining the label-wise global explanation generated with our methodology. However, we are unsure at present whether the label-wise for Glaucoma matches indeed matches this piece of clinical evidence, given that the Optical Coherence Tomography used (Sihota et al., 2015) might have a lower field of view than these ultra-widefield retinal images.

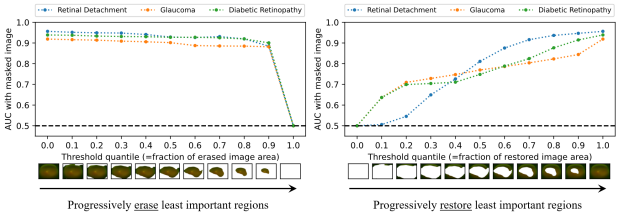

Figure 5. **Results of PEPPR. Both:** The images at the bottom show the mean validation image with the mask at the given quantile applied. The black dotted horizontal line indicates AUC=0.5 (equivalent to random guessing). **Left:** Progressive erasure of the least important regions. We observe that performance for all three labels barely dropped by the time half the image was erased. **Right:** Progressive restoration of the least important regions. We observe a monotonic increase in AUC. Note the particularly large increase in AUC for Glaucoma when restoring the most important 10% of the image containing the optic disc.

3.4. PEPPR

We conducted PEPPR using quantile steps of 10%. We replace erased pixels with random noise rather than the mean value across the training data as we trained our model with RandomErasing and thus it should be robust to encountering random noise. The results are shown in Figure 5. The progressive erasure (Figure 5 left) shows that the performance for each of the three classes drops as we erase more of the images, but not substantially so. A large drop in performance only occurs once the entire image is erased. This suggests, that the macro SAGE has indeed identified the image region that is most important for the model's performance. The progressive restoration (Figure 5 right) shows a near monotonic increase for all three classes as more regions are restored, which suggests that the macro SAGE provides a meaningful ranking of the importance of different image regions. While having the most important 10% of the images allows for performance close to the level of having full images, this level of performance is not matched even when the 50% of the images is available that the macro SAGE identifies as least important. The monotonic increase in performance for progressive restoration suggests that the periphery does contain meaningful information. However, this information does not allow for substantially better performance that just having the central region available. This suggests that the central region might contain all or most of the information in the periphery, but not vice versa.

We also note that with the least important 10% of the images, AUCs in excess of 0.6 can be achieved for Glaucoma and Diabetic Retinopathy. This is unexpected as these regions do not show the retina for many images and thus should be mostly uninformative. This could be a sign of a data artefact that should be investigated before a model were deployed in practice. However, the progressive erasure results suggest

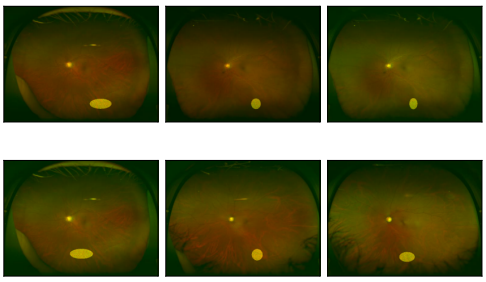

Figure 6. **Examples of the artificial short signal.**

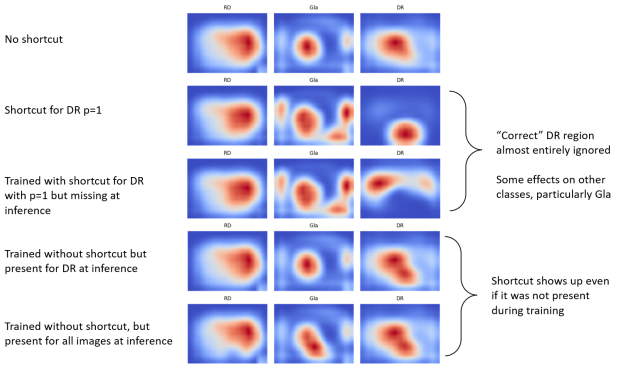

Figure 7. **SAGEs for the classes when varying presence of shortcut during training and inference.**

that even if there is a data artefact in those regions, our model works well if those regions are erased. Overall, the results of PEPPR suggest that the macro SAGE faithfully reflects how the model works and correctly ranks the regions by their relative importance.

3.5. Artificial shortcuts

To test whether our methods can detect if the model relies on "shortcuts", we conducted an experiment introducing an artificial shortcut that mimics reflection artefacts that occur in UWF images Figure 6. In our preliminary experiments, we introduce the artefact for Diabetic Retinopathy (DR) with p=1, and find that with this artefact present the SAGE for DR deviates from the pattern we would expect based on domain knowledge Figure 7. Thus, together with domain experts, a practitioner would notice this unexpected SAGE and further investigation would reveal this shortcut. However, we intend to do more comprehensive experiments to investigate what kinds of shortcuts can be detect with our method. An additional figure is reported in the appendix Figure 8.

4. Discussion & conclusion

We introduced the notion of an aligned image modality, and proposed aggregating per-image explanations into SAGEs. We further proposed PEPPR for quantitatively validating these explanations. We then applied these methods to ultra-widefield retina images, finding that the SAGEs are consistent with domain knowledge and that macro SAGE correctly ranked the image regions according to their importance for our model. The application of these methods is limited to image modalities that are aligned. This is rare for unprocessed natural images but common in domains like medical imaging or natural images that have been post-processed. Furthermore, our methods also assume that information about the target variables has a spatial dependence, e.g. that different labels tend to occur in different regions or that some regions should be entirely uninformative. However, in image modalities that are aligned, we would expect to find such characteristics.

While the results from our experiments are encouraging, there are many directions that future work could explore. First, we only used GradCAM to generate per-image-and-class explanations we used as input to our methods. Future work could investigate whether alternatives to GradCAM yield characteristically different SAGEs. Second, future work could generating SAGEs through PEPPR itself, similar to how AblationCAM (Ramaswamy et al., 2020) functions. This would take a data-centric perspective and frame global explainability in aligned modalities as a feature selection problem, where we want to select the most informative pixels or regions. Thus, this approach could leverage approaches from feature selection such as Relevance Determination and Max Information Gain. Third, future work could explore applying the methods to different datasets like chest x-rays, as well as three-dimensional data such as brain Magnetic Resonance Images. Finally, SAGEs might be used to contrast the behaviour of different models trained on the same data, e.g. with different augmentation strategies. In aligned modalities, we could also use learned attention (e.g. a ResNet with an attention pool) where we learn a query per disease and only use location embeddings but not tokens as the keys. This might allow to see which locations the model learns to expect possible pathology for specific diseases to occur.

In this work, we focused on presenting and testing the methodology we introduced. We hope that these methods will be a further tool in the toolbox for model explanation, and useful to applied work where it is used to validate DL models that are developed for critical applications and to discover new domain knowledge.

Data availability statement

The data is available from the authors of the Tsukazaki Optos Public Project upon reasonable request. Previously, it was publicly accessible via a project website where we obtained the copy used in this study. A subset containing images images of healthy eyes and eyes with RP used in a previous study (Masumoto et al., 2019) is publicly accessible directly online: `https://figshare.com/authors/Masahiro_Kameoka/6020591`.

Acknowledgements

We thank Antreas Antoniou and Elliot Crowley for their support, feedback and comments.

We thank Dr. Hiroki Masumoto, as well as Daisuke Nagasato, Shunsuke Nakakura, Masahiro Kameoka, Hitoshi Tabuchi, Ryota Aoki, Takahiro Sogawa, Shinji Matsuba, Hirotaka Tanabe, Toshihiko Nagasawa, Yuki Yoshizumi, Tomoaki Sonobe, Tomofusa Yamauchi and all their colleagues at Tsukazaki hospital for releasing the TOP dataset. This is a great contribution to AI research in ophthalmology for which we are most grateful.

This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

Adebayo, J., Gilmer, J., Goodfellow, I., and Kim, B. Local explanation methods for deep neural networks lack sensitivity to parameter values. arXiv preprint arXiv:1810.03307, 2018a.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity Checks for Saliency Maps. arXiv preprint arXiv:1810.03292, 2018b.

Bromberger, S. On what we know we don't know: Explanation, theory, linguistics, and how questions shape them. University of Chicago Press, 1992.

Choo, J. and Liu, S. Visual analytics for explainable deep learning. IEEE Computer Graphics and Applications, 38 (4):84–92, 2018.

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, pp. 1–10, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image

database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Ieee, 2009.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11):665–673, 2020.

Gildenblat, J. and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pp. 80–89. IEEE, 2018.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A Survey Of Methods For Explaining Black Box Models. ACM computing surveys (CSUR), 51(5):1–42, 2018.

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

Lakkaraju, H. and Bastani, O. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 79–85, 2020.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1):1–8, 2019.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610, 2019.

Masumoto, H., Tabuchi, H., Adachi, S., Nakakura, S., Ohsugi, H., and Nagasato, D. Retinal detachment screening with ensembles of neural network models. In Asian Conference on Computer Vision, pp. 251–260. Springer, 2018a.

Masumoto, H., Tabuchi, H., Nakakura, S., Ishitobi, N., Miki, M., and Enno, H. Deep-learning classifier with an ultrawide-field scanning laser ophthalmoscope detects glaucoma visual field severity. Journal of Glaucoma, 27 (7):647–652, 2018b.

Masumoto, H., Tabuchi, H., Nakakura, S., Ohsugi, H., Enno, H., Ishitobi, N., Ohsugi, E., and Mitamura, Y. Accuracy of a deep convolutional neural network in detection of retinitis pigmentosa on ultrawide-field images. PeerJ, 7: e6900, 2019.

Matsuba, S., Tabuchi, H., Ohsugi, H., Enno, H., Ishitobi, N., Masumoto, H., and Kiuchi, Y. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. International Ophthalmology, 39 (6):1269–1275, 2019.

Nagasato, D., Tabuchi, H., Ohsugi, H., Masumoto, H., Enno, H., Ishitobi, N., Sonobe, T., Kameoka, M., Niki, M., Hayashi, K., et al. Deep neural network-based method for detecting central retinal vein occlusion using ultrawide-field fundus ophthalmoscopy. Journal of Ophthalmology, 2018, 2018.

Nagasato, D., Tabuchi, H., Ohsugi, H., Masumoto, H., Enno, H., Ishitobi, N., Sonobe, T., Kameoka, M., Niki, M., and Mitamura, Y. Deep-learning classifier with ultrawide-field fundus ophthalmoscopy for detecting branch retinal vein occlusion. International Journal of Ophthalmology, 12(1):94, 2019.

Ohsugi, H., Tabuchi, H., Enno, H., and Ishitobi, N. Accuracy of deep learning, a machine-learning technology, using ultra–wide-field fundus ophthalmoscopy for detecting rhegmatogenous retinal detachment. Scientific Reports, 7(1):1–4, 2017.

Olah, C., Mordvintsev, A., and Schubert, L. Feature visualization. Distill, 2(11):e7, 2017.

Pal, A., Phan, R. C.-W., and Wong, K. Synthesize-It-Classifier: Learning a Generative Classifier Through Recurrent Self-Analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5161–5170, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc., 2019.

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

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K. Black-box explanation of object detectors via saliency maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11443–11452, 2021.

Ramaswamy, H. G. et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 983–991, 2020.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3):199–217, 2021.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transactions on Neural Networks and Learning Systems, 28(11):2660–2673, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. In International Conference on Machine Learning, pp. 3145–3153. PMLR, 2017.

Shunmugam, M., Shah, A. N., Hysi, P. G., and Williamson, T. H. The pattern and distribution of retinal breaks in eyes with rhegmatogenous retinal detachment. American Journal of Ophthalmology, 157(1):221–226, 2014.

Sihota, R., Naithani, P., Sony, P., and Gupta, V. Temporal retinal thickness in eyes with glaucomatous visual field defects using optical coherence tomography. Journal of Glaucoma, 24(4):257–261, 2015.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, pp. 843–852, 2017.

Thagard, P. R. The best explanation: Criteria for theory choice. The Journal of Philosophy, 75(2):76–92, 1978.

Victor, A. A. Optic Nerve Changes in Diabetic Retinopathy. In Ferreri, F. M. (ed.), Optic Nerve, chapter 4. IntechOpen, Rijeka, 2019. doi: 10.5772/intechopen.81221. URL https://doi.org/10.5772/intechopen.81221.

Wightman, R. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models, 2019.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 13001–13008, 2020.

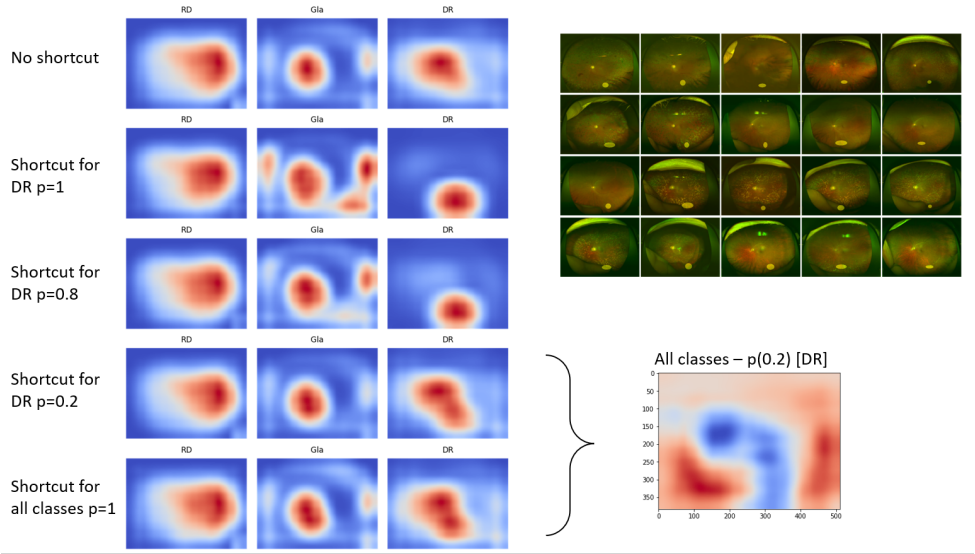

Figure 8. **SAGEs for the classes when varying the probability of the shortcut occurring.**