Supplementary Materials

FGR-Net: Interpretable Fundus Image Gradeability Classification Based on Deep Reconstruction Learning

1 Appendix 1: Classification Performance

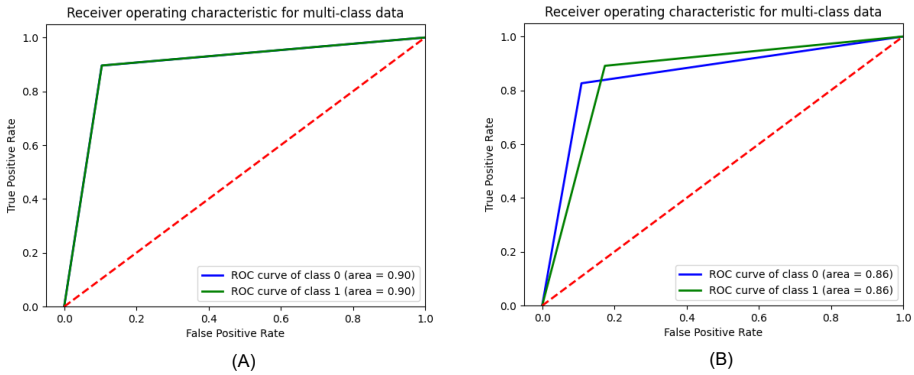

Fig. 1. Receiver operating characteristic (ROC) curves with the best model from two datasets. (A) EyePACS dataset with two classes. (B) Hospital dataset with two classes.

Figure 1 shows a receiver operator characteristic (ROC) curve [1] showing the diagnostic ability of FGR-Net to classify fundus images. Figure 1 shows ROC curves with the testing set of Eyepaccs and our private dataset. The EyePACS and private datasets demonstrated that the model achieved an Area Under Curve (AUC) of 0.9 with classes 0 and 1. In turn, our private dataset yielded an AUC value of 0.86 for the two classes. For these experiments, it is evident that combing an autoencoder network with a classifier network helps the model extract more patterns and features about the quality of fundus images.

Figure 2 also shows the ROC curve to show the relationship between sensitivity and specificity, resulting in the FGR-Net model of a three-class problem with the Eye-Quality (EyeQ) dataset. The model achieved an AUC of 0.94 with class 0, 0.88 with class 1, and 0.91 with class 2. The results confirm the results shown in the confusion matrix.

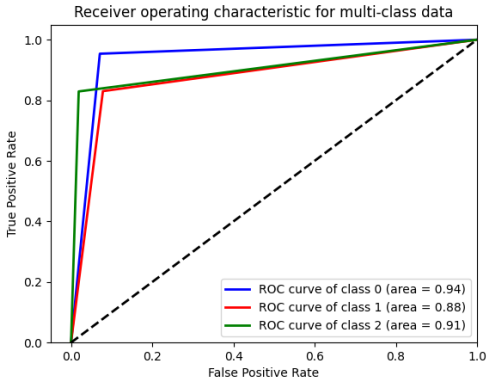

Fig. 2. Receiver operating characteristic (ROC) curve with the best model with the Eye-Quality (EyeQ) dataset of three classes.

2 Appendix 2: Explainability Experiments

2.1 Average Gradient and GradCAM saliency maps

In Figure 3 we show the average of 50 left-oriented fundus images for each class and model, as well as the corresponding average Gradient and GradCAM saliency maps. Given that fundus images can be left or right-oriented, and this presupposes a strong prior for the model, using a set of images with the same orientation allows understanding the average behaviour of the model. Since left and right fundus images are symmetric, and we use random horizontal flips as a data augmentation step, it is expected that the model learns this bi-modality and therefore the same results should be obtained with either orientation.

The average GradCAM saliency map shows that while the Autoencoder model focuses mostly on the optic disc,the Normal model pays a lot of attention on an arbitrary point on the upper right border of the fundus image that has no clinical significance. The Normal model does, however, focus slightly along the line of the main blood vessels that pass through the optic disc, but comparatively less than the Autoencoder model.

The average Gradient visualization,on the other hand, shows more sensitivity to small blood vessels in the whole image for the Normal model. Coincidentally with the GradCAM visualization, it confirms the undesired sensitivity focused on the pixels in upper-right border of the fundus disc. The Autoencoder model shows more focused areas of interest, specially in the top-right portion of the fundus, but also in the main blood vessel line that pass through the optic disk.

2.2 Unsupervised learning on Saliency maps

While the main experiment section analyzes the saliency maps of typical examples via the Gradient and GradCAM methods, it would be more useful to analyse

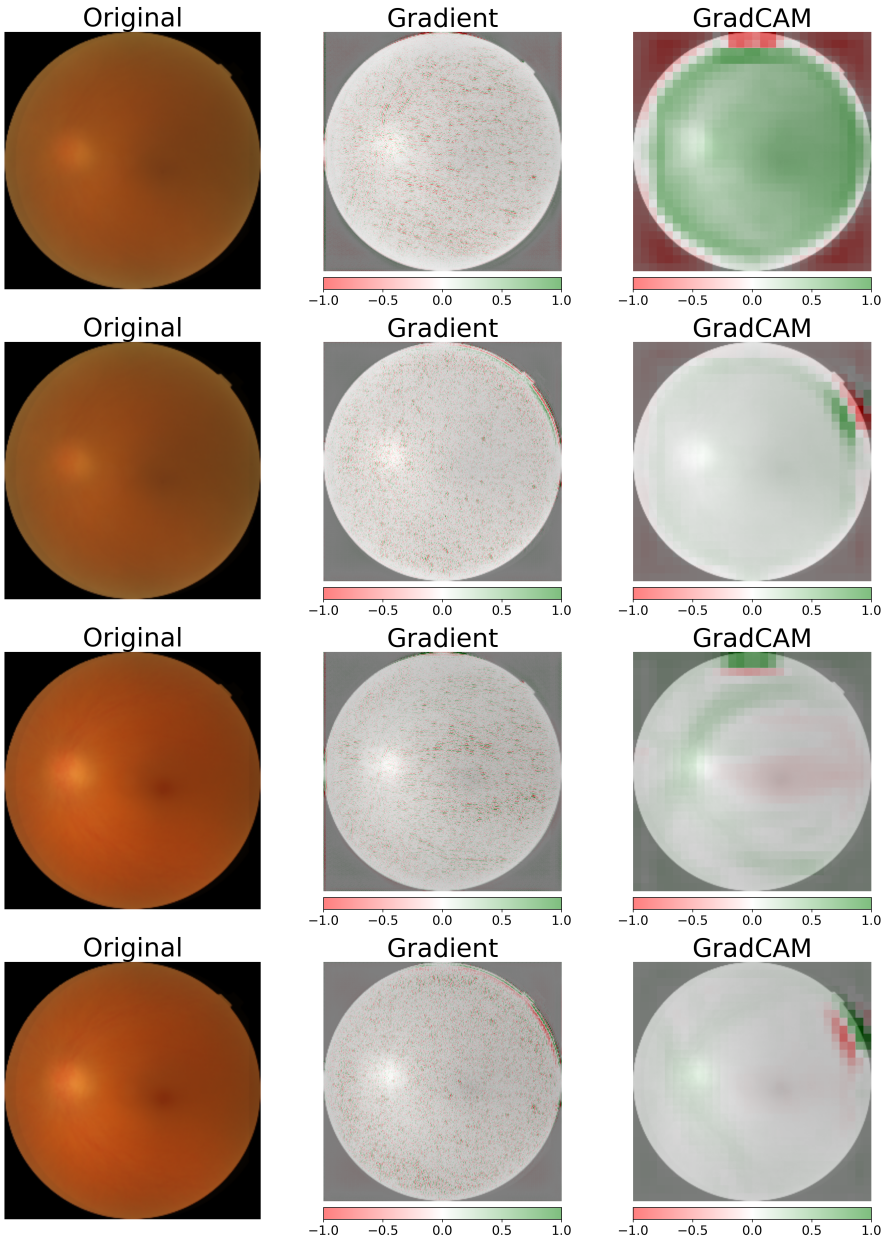

Fig. 3. Mean image (left), Gradient (middle) and GradCAM (right) of 50 left-oriented fundus images. Rows 1 and 2 correspond to saliency maps of class *Gradable* for the Normal and Autoencoder model, respectively. Rows 3 and 4 correspond to class *Ungradable* for the same models.

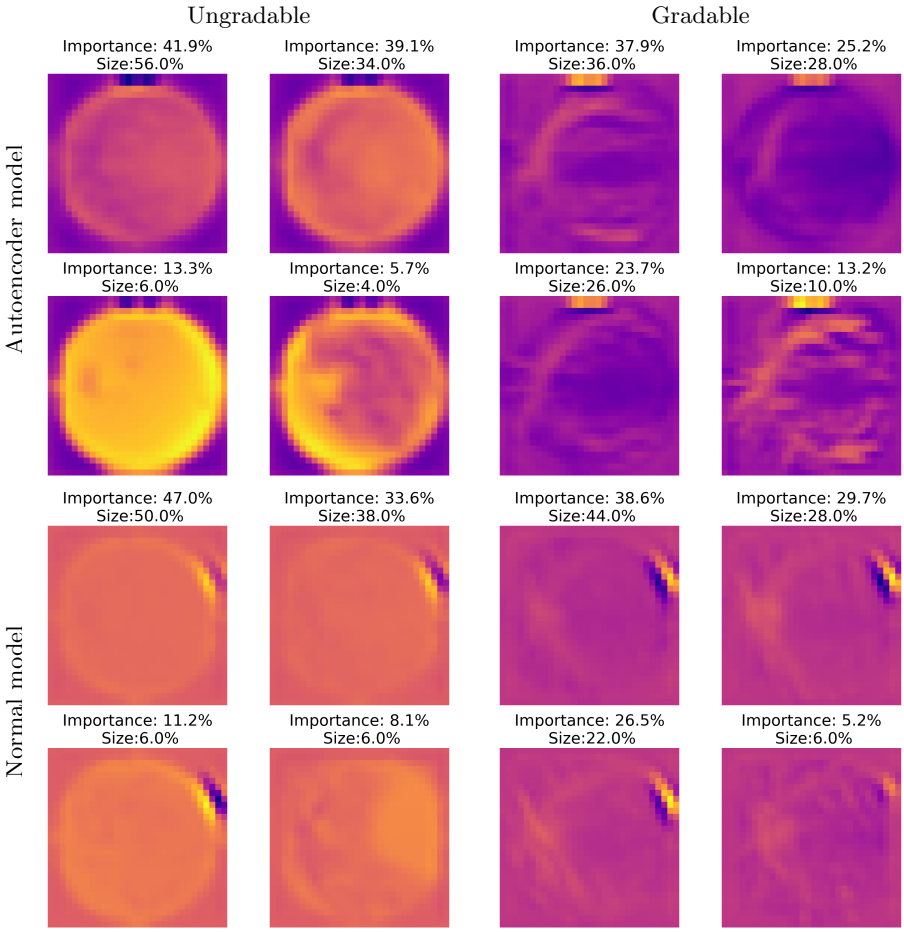

Fig. 4. K-Means clustering of a GradCAM saliency map, with $k = 9$, over a set of 50 left oriented examples for the Autoencoder model (top), Normal model (bottom), and classes *Gradable* (left) and *Ungradable* (right). Each image corresponds to a cluster center, and also shows the relative size (percentage of activations assigned to the cluster) and importance (percentage over the sum of absolute saliency values).

the set of features for a larger set of examples without recurring to a manual inspection of each. Therefore, we cluster the results of the GradCAM method for a set of 50 different, left oriented, fundus images. We repeat this process for each class, to obtain cluster centers with K-Means that represent prototypical feature activations, and for both models, as shown in Figure 4.

As we can see, the features for the Autoencoder model are much more relevant and well defined. The Normal model focuses mostly on the small peak on the top right of the fundus image, which is not relevant to determine the gradability of the image. On the other hand, the Autoencoder model does show the importance of the main blood vessels, specially for the *Gradable* class. It is also worth noting that the Autoencoder model also shows a small artifact as it focuses on a pattern on the top center of the images, but its importance is much smaller relative to the other areas of the image than in the case of the Normal model.

We also note that Non-Negative Matrix Factorization (NMF), another typical approach for aggregating activations [2], was also employed but gave inferior results in this case (Figure 4). Since the structure and alignment of the fundus images is very uniform, a centroid-based clustering such as K-Means focuses more on separating different types of activations based on their general structure (ie, blood vessels, textures), while NMF focuses more on separating the features based on the region where they are most activated (ie, top, bottom, border, etc).

References

1. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology **143**(1), 29–36 (1982)
2. Kolouri, S., Martin, C.E., Hoffmann, H.: Explaining distributed neural activations via unsupervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28 (2017)