Sparse Activations for Interpretable Disease Grading

Anonymous Submission[1] xx@xx
[1] ****

Editors: Under Review for MIDL 2023

Abstract

Deep learning models have shown great promise in classifying medical images and rely primarily on well-known saliency map techniques, such as GradCAM to interpret decisions. However, these techniques sometimes fail and provide coarse localization maps even for images with small regions of interest (ROIs) such as diabetic retinopathy (DR) lesions in fundus images or breast cancer tumors on mammography images. Here, we modify the bag-of-local-features models called BagNet into a fully convolutional network to classify medical images with small ROIs and then include a sparsification constraint based on the L1-norm to improve interpretability and produce an informative, sparse, fine-grained, and class evidence maps that detect and highlight accurately the discriminative regions of the input. We evaluated the classification performances of our models on the publicly available Kaggle retinal dataset, which features small DR lesions such as microaneurysms, and hemorrhages. The evaluation showed classification performance comparable to or better than the average state-of-the-art for DR grading. We also evaluated quantitatively the localization properties of the models on 15 annotated images sampled from the test set, which showed that the sparse model has better localization precision than the dense model. Our results put barriers on the high-performance black box models that use heavy post-attention techniques instead of design-interpretable models to understand and interpret the decisions on medical images.

Keywords: Diabetic retinopathy, sparse activations, interpretability

1. Introduction

While machine learning (ML) tools have been approaching expert-level performance in many medical imaging tasks (De Fauw et al., 2018; Shen et al., 2019; Mahoro and Akhloufi, 2022), their lack of interpretability is a key obstacle to widely applying them in clinical practice (Teng et al., 2022) and a potential ethical issue (Grote and Berens, 2020). The most common approach towards explaining deep ML models are saliency maps (Zeiler and Fergus, 2013; Simonyan et al., 2014; Zhou et al., 2015; Springenberg et al., 2015; Selvaraju et al., 2020), which post-hoc identify image regions that influence the classification result of a trained classifier. However, saliency maps have recently been shown to be poorly reproducible for medical images (Arun et al., 2021) and do not provide actionable insights, given that they do not accurately reflect the network's actual decision mechanisms. Instead, inherently interpretable models could provide a path forward for safety-critical tasks (Rudin, 2019), but few such solutions exist which at the same time provide sufficiently high accuracy.

In this paper, we develop an inherently interpretable deep learning model for clinically relevant tasks that rely on identifying and characterising small lesions or other anomalies in large search regions. Specifically, we focus on the detection and grading of Diabetic

© 2023 A. Submission.

Retinopathy (DR) on retinal fundus images, but other potential applications include screening for other retinal diseases and breast or lung cancer.

DR is a microvascular complication of diabetes characterized by the progressive presence of one or more small retinal lesions such as microaneurysms, hemorrhages, or hard and soft exudates (ICO, 2017). It is the leading cause of blindness in the working-age population and the third leading cause of visual impairment worldwide, and early diagnosis and treatment can slow its progression (ICO, 2017; Wong et al., 2018). It is therefore recommended that diabetes patients undergo regular monitoring, and automated models could facilitate mass screening and help clinicians use their time more efficiently (Ting et al., 2016).

Numerous black-box DR detection methods have been proposed (Rao et al., 2020; Alyoubi et al., 2020; Tavakoli and Kelley, 2021; Huang et al., 2022) with competitive and often superior performance than humans (Ruamviboonsuk et al., 2019). For such methods, the interpretation of their classification is mostly aided by saliency maps (Wang and Yang, 2019; Chetoui and Akhloufi, 2020) such as class activation mapping (CAM) (Zhou et al., 2015) and gradient-weighted CAM (Selvaraju et al., 2020). Moving towards a more interpretable model for detecting DR, Papadopoulos et al. (2021) used a multiple-instance learning approach with an attention mechanism to combine features extracted from different image patches. The attention mechanism also provides a heatmap to visualise the contribution of different image regions. Although all of these methods provide some visual evidence of suspicious image regions, they suffer from limitations: (1) none of them directly explains the decision mechanism, (2) the resulting heatmaps are coarse, and small DR lesions may require more fine-grained explanations to clearly indicate salient regions of the image, and (3) to serve as useful and actionable feedback to clinicians, sparse heatmaps may be more useful as they easier to digest and less distracting than typical dense and cluttered ones.

We overcome these key limitations and propose an inherently interpretable model for DR detection. Our approach is based on a bag-of-local-features model (BagNet) (Brendel and Bethge, 2019), which has already been shown to be effective in ophthalmology for gender classification from fundus images (Ilanchezian et al., 2021). BagNet relies solely on local evidence, which makes it relevant for classification or detection tasks where the regions of interest are very small and distributed over the image, as is the case in DR. Our main contributions can be summarized as follows:

1. We propose a number of simple but effective changes to the BagNet architecture and training regime (Sec. 2.2, illustrated in Fig. 1), leading to fine-grained, sparse, and easily interpretable heatmaps which directly and correctly reflect the neural network's decision mechanism. As far as we are aware, no previous studies have used sparsity (Sec. 2.3) to aid interpretability.

2. We evaluate our method on publicly available fundus images both for the binary task of referable DR detection (Sec. 3.2) and multi-class DR grading (Sec. 3.3), and experimentally show that it performs comparably to state-of-the-art models.

3. We propose a clinically relevant evaluation strategy to assess the usefulness of heatmaps (Sec. 2.4) and show that our method leads to high precision in identifying task-relevant image regions (Sec. 3.2.2).

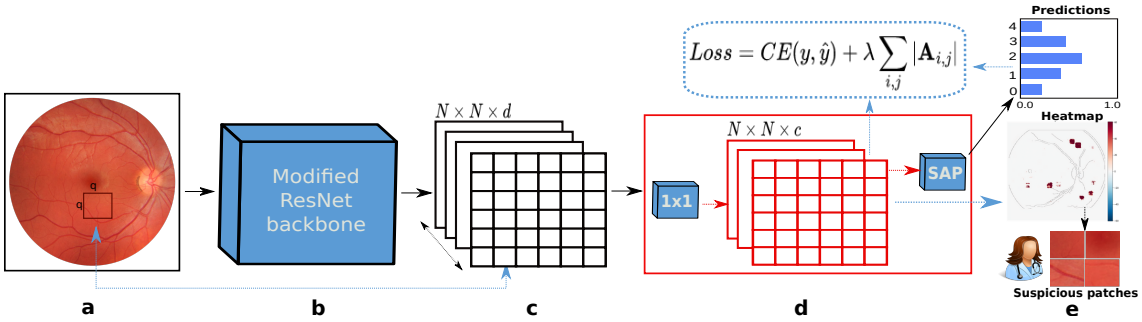

Figure 1: High-level overview of the proposed interpretable DR detection pipeline. **(a)** Example fundus image from the Kaggle dataset. The square inside is a small patch of size $q \times q$. **(b)** The modified ResNet-50 Block obtained by replacing most of the 3×3 filter with 1×1. **(c)** The feature maps of size $N \times N \times d$ of the last ResNet layer. **(d)** The final feature map obtained from the previous one with n kernels of size 1×1 where n is the number of classes. **(e)** From top to bottom, we denote respectively the class evidence score obtained by a simple global average pooling over the class dimension of the final activation; a class-specific selected from the n activation maps; some suspicious topK patches sampled from the input and based on the salient regions of the heatmap.

2. Developing an interpretable-by-design disease classification network

Below we describe the proposed model and pipeline (Fig. 1) introducing the modifications we made to the BagNet model, the sparsity constraints we introduced, and how the heatmap and candidate patches were obtained.

2.1. Backbone architecture and baseline model

We used a BagNet (Brendel and Bethge, 2019) as a baseline interpretable-by-design classification model. The BagNet is a variant of ResNet-50 (He et al., 2016) and is obtained by replacing many of the 3×3 convolutional kernels of a ResNet-50 with 1×1 kernels and reducing the strides. This leads to a final feature map F of size $N \times N \times d$ where d is the number of feature channels (Fig. 1c, typically $d = 2048$). A spatial average pooling layer reduces these activations to the $1 \times d$ (see Appendix Fig. 5d), and a linear layer subsequently provides the logits of the final predictions $l = \mathbb{R}^{1 \times c}$ where c is the number of classes (see Appendix Fig. 5e). The final prediction can be viewed as the average of all the local evidence.

These architecture modifications in the BagNet have two effects: First, each pixel in the final activation map has a reduced effective receptive field of size $q \times q$. Therefore, the model makes its predictions based on small local image patches of size $q \times q$, implicitly. Secondly, reducing the stride in the convolutional layers prevents downsampling effects and results in relatively high-resolution (i.e. fine-grained) activation maps compared to the original ResNet.

2.2. Enhancing the architecture with an interpretable decision-making mechanism

To interpret predictions in the original BagNet model, one cannot directly examine the final feature maps (Fig. 1c), as these represent high-dimensional features rather than class evidence at each location. Rather, it is necessary to construct activation maps with multiple forward passes for individual image patches of size $q \times q$.

Here, we introduce explicit class evidence layers to obtain actual class evidence maps with a single forward pass and make the local information representation explicit. To this end, we reorganised the two final network operations. As both the spatial average pooling and the final fully connected layer are summing, swapping their order can be done without affecting the resulting logits:

$$l_c = \sum_{i=1}^{d} w_{i,c} \left(\sum_{x,y} \frac{1}{N^2} F_i^{x,y} \right) = \sum_{x,y} \frac{1}{N^2} \left(\sum_{i=1}^{d} w_{i,c} F_i^{x,y} \right) = \sum_{x,y} \frac{1}{N^2} A_c^{x,y}. \tag{1}$$

Importantly, Eq. 1 can be implemented by replacing the (swapped) FC layer by a 1×1 convolution with c output channels. The final class-wise evidence maps A_c directly represent the contribution of individual input patches to the final prediction. The final class score is then obtained by simple spatial averaging (Fig. 1d), resulting in a c-dimensional vector, which is then transformed by the softmax function into class probabilities (Fig. 1e).

This modification of the architecture makes the model more interpretable by design and offers some advantages: (1) it intuitively improves the transparency of the model by providing activation maps (without further post-processing) showing the contribution of each patch to the final prediction; (2) it provides for each class a different class evidence map showing the contribution of each patch to the classification of the input into that class (which is more useful for multi-class tasks); and (3) it allows for the application of sparsity constraints based on the activation maps of each class (see below).

2.3. Sparsity constraints on class evidence layer further improve interpretablity

We found that the original BagNet produces very dense heatmaps, with many positive and negative activations, indicating that clinically irrelevant input patches contribute to the prediction. This behavior makes it difficult for a human to discern how the prediction was formed and to efficiently verify its correctness.

By introducing explicit class evidence layers (Sec. 2.2) we can directly place constraints on the class evidence map (Fig. 1d) containing the per-patch scores to induce spatial sparsity. To achieve that, we propose to place an L1 regularisation constraint on the class evidence maps A, leading to the following loss function:

$$loss(\mathbf{y}, \hat{\mathbf{y}}) = \mathrm{CE}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \sum_{i,j} |\mathbf{A_{ij}}|. \tag{2}$$

The sparsity of the activation maps depends on the hyperparameter λ. Enforcing sparsity in class evidence in this way is not a post-hoc measure, but rather forces the classification model to focus on the most relevant regions.

	Acc.	**AUC**	**Spec.**	**Sens.**
ResNet-50	0.942	0.960	0.993	0.810
Dense BagNet	0.936	0.957	0.991	0.779
Sparse BagNet	0.928	0.937	1.0	0.750

Table 1: Classification performance on referable DR detection task between the proposed approaches and the baseline ResNet-50 model on the test set.

2.4. Potential use in a clinical workflow

Compared to the initial BagNet, the modified BagNet makes it easier to derive class evidence maps, while being less computationally and memory intensive. The resulting low-resolution activation maps (Fig. 1d) can be up-sampled to the input size and overlaid on the input (Fig. 1e) for easy visualisation and interpretation by clinicians. The class evidence map is then used to extract suspicious patches (Fig. 1d) based on activation map scores and presented to the clinician for further investigations. A clinician can use the global prediction, the class evidence maps, and suspicious patches to either strengthen their trust in the model or reject a decision.

3. Experiments and Results

Below we describe the experiments including the dataset and preprocessing, then we describe the metrics used to evaluate both the classification model and the heatmap. We finally present and discuss the main results.

3.1. Data

We used retinal fundus images from the Kaggle Diabetic Retinopathy challenge (Kaggle, 2015) with reference DR grades ranging from 0 (no DR) to 4 (proliferative DR). We removed poor-quality images from the dataset using an ensemble of EffcientNets models (Tan and Le, 2020) trained on the ISBI2020 [1] challenge dataset. The resulting dataset after the quality filtering contained $45,923$ images, which we split into training (75%), validation (10%), and test sets (15%), making sure images from the same patients were allocated to the same set. We are particularly interested in detecting referable DR (grade 2 or higher) as this is where machine learning algorithms can significantly help to prevent vision loss in early screening settings. Therefore, we also group the DR grades into $\{0, 1\}$ vs $\{2, 3, 4\}$ binary labels.

We preprocessed the images by fitting a circular mask to the field of view and cropping its bounding box. All images were resized to 512×512 and the image intensities were normalised by the mean and standard deviation of the training set. For complementary analyses, an in-house experienced ophthalmologist provided detailed lesion annotations on a selection of 15 test images.

3.2. Application to referable DR detection

We first evaluated our method for the clinically more relevant case of (binary) referable DR detection. We configured the backbone architecture (Sec. 2.1) such that the receptive field size was $q = 33$ as in Ilanchezian et al. (2021). The regularisation coefficient (Eq. 2) was set to $\lambda = 5 * 10^{-5}$ based on the tradeoff between classification performance on the validation set and sparsity (see Appendix Fig. 6a).

We compared our sparse BagNet against its dense baseline ($\lambda = 0$) and a ResNet-50 as a black-box state-of-the-art reference. The training procedures for all models were adopted from Huang et al. (2022), who comprehensively evaluated ideal hyperparameter choices for the specific application of DR grading using ResNet-50 (see App. A.1).

We found that the dense model achieved good accuracy and AUC, but comparatively, low sensitivity compared to the baseline model (Tab. 1). In contrast, the sparse model performed less well. Due to the variety of datasets, the splitting strategies, and the method used to filter out the images with low quality, there are no reference baseline classification metrics for DR grading. However, based on the state-of-the-art metrics reported by Papadopoulos et al. (2021), the average AUC of the models that use the Kaggle dataset to train and evaluate on referable DR task is approximately 0.942, which is lower than the AUC of the dense model. This result shows that the dense BagNet is a potentially good candidate model for DR detection and that a good hyper-parameter search could even improve the classification performances on DR detection.

3.2.1. Sparsity constraints declutter class evidence maps

Next, we examined the local evidence for classifier decisions provided by the models. For the BagNet variants, the class evidence maps could be used as heatmaps directly as described in Sec. 2.2. We also visualised the local contributions for ResNet-50 using the post-hoc technique GradCAM (Selvaraju et al., 2020). While some highlighted regions were similar to the BagNet versions (Fig. 2), the ResNet heatmaps were coarse due to the model's large receptive field. As the ResNet's heatmaps do not represent the specific local contribution to the model's decision-making process, we focused on the inherently interpretable BagNet versions for further analysis.

The constraints we imposed on our model led to much sparser heatmaps compared to the original BagNet (see right columns in Fig. 2), showing that the model decision was formed from few small regions of the retinal fundus. These regions seemed to be mostly a subset of the salient regions used by the dense model. On healthy images, the sparse model led to an almost complete absence of positive activations, which contrasts the mix of positive and negative evidence suggested by the dense model (see bottom row in Fig. 2). To assess this quantitatively, we measured how consistently the local class evidence supported to the final prediction. We calculated the proportion of pixels with negative scores among all non-zero scores on healthy test images, which confirmed the qualitative assessment (second column in Tab. 2). The same analysis for diseased eyes also showed an increased heatmap consistency compared to the dense model (third column in Tab. 2). However, on diseased eyes, there remained a large proportion of evidence for healthy tissue.

1. https://isbi.deepdr.org/challenge2.html

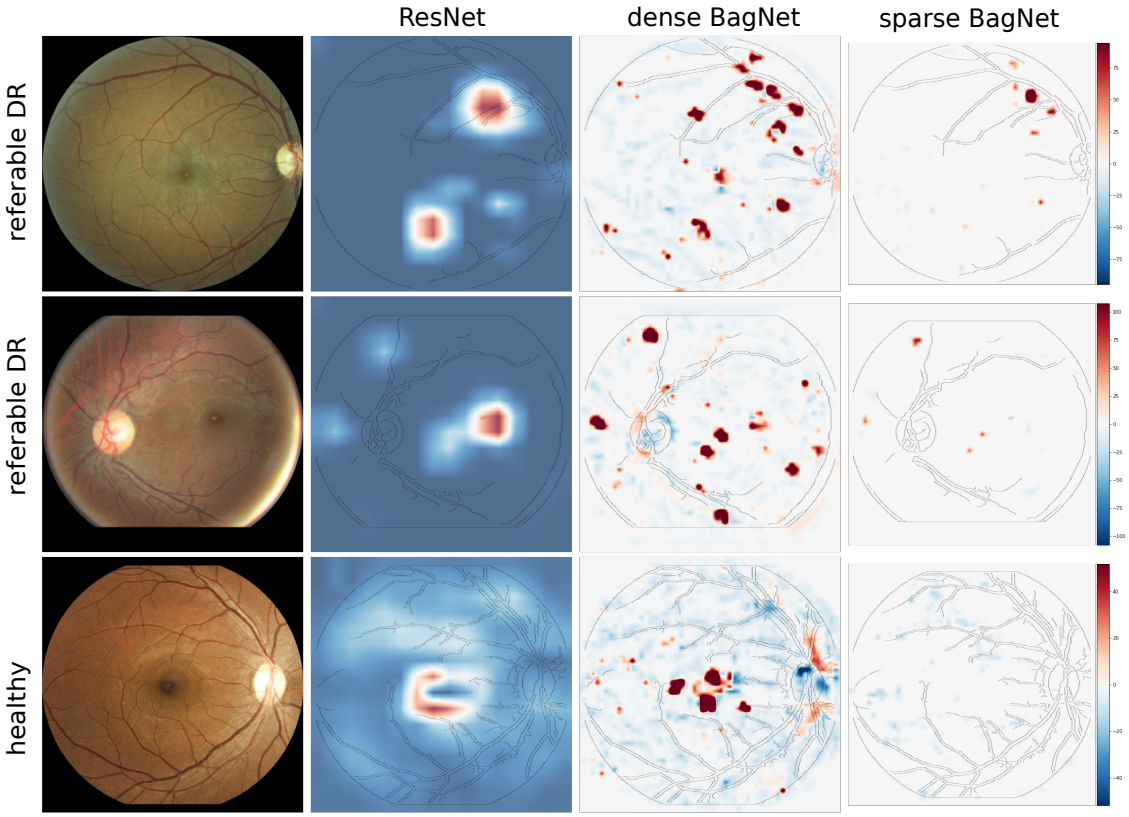

Figure 2: Examples of qualitative evaluation between the visualization maps produced by the GradCAM on the baseline ResNet-50 and the heatmaps provided by the dense and sparse BagNet on referable DR grading.

	Precision ↑	**T. fneg ↑**	**T. fpos ↑**
Dense BagNet	0.227	0.935	0.118
Sparse BagNet	0.767	0.993	0.448

Table 2: Quantitative heatmap evaluation on the test set. The first column refers to the localisation precision. The second column refers to the total fraction of negative activation on non-referable fundus images. The third column refers to the total of positive activation on referable fundus images.

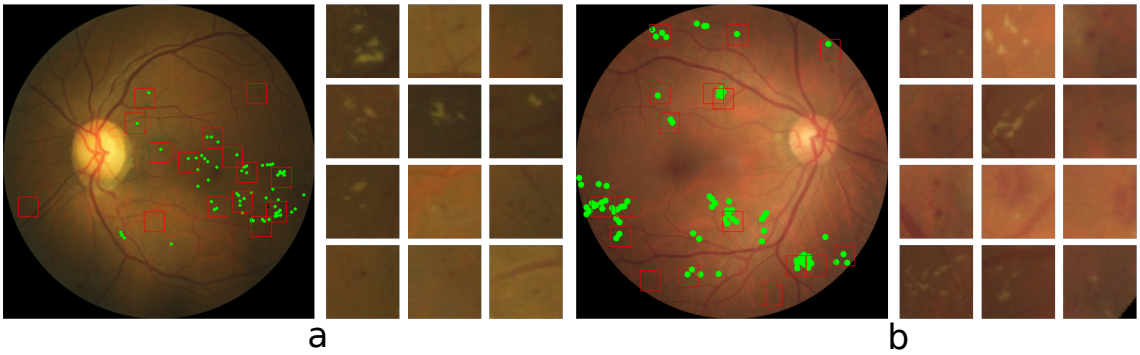

Figure 3: Annotations of two referrable DR samples with the top 12 suspicious patches. The green markers indicate the annotation of lesions by an experienced ophthalmologist, whereas the red box (enlarge on the right and sorted according to their decreasing classification scores) denotes suspicious patches provided by the sparse model based on the activation map scores.

3.2.2. Lesion detection precision of the sparse bagnet

To quantitatively assess whether the highlighted regions were clinically relevant for diagnosing DR, we calculated the heatmaps' precision at localising DR lesions on the subset of fully annotated test images. On these images, we extracted all non-overlapping input patches with positive scores and calculated precision as the proportion of patches that contained an annotated DR lesion.

The dense BagNet model contained many positive activations in the healthy area, resulting in low lesion localisation precision. In contrast, the sparsifying the model increased precision considerably (first column in Tab. 2).

We inspected the relevant patches identified by the sparse model on example annotated images (Fig. 3) and found high overlap between patches with positive scores (red boxes, magnified on the right) and actual lesions (green markers). Some false positive patches could be due to the lesion-like appearance of image artifacts or healthy variations in pigmentation.

3.3. Application to multi-class DR grading

Finally, we applied our method to the multi-class setting of DR grading, where individual severity grades were predicted. We used the same hyperparameters as for the binary task and set the number of output classes to 5. We set the regularisation coefficient (Eq. 2) of the sparse model to $\lambda = 6 * 10^{-6}$ based on the tradeoff between classification performance on the validation set and sparsity (see App. Fig. 8). We found that the dense and sparse models achieve comparable or slightly better accuracy (resp. 0.864, and 0.850) than the baseline ResNet-50 model (0.862).

Next, we qualitatively evaluate the class evidence maps provided by the dense and sparse model. In contrast to the GradCAM (see Appendix Fig. 8), the BagNet provides class-specific activation maps (Fig. 4) which highlights the salient regions for a given class (for example the regions containing hemorrhage for grade 2).

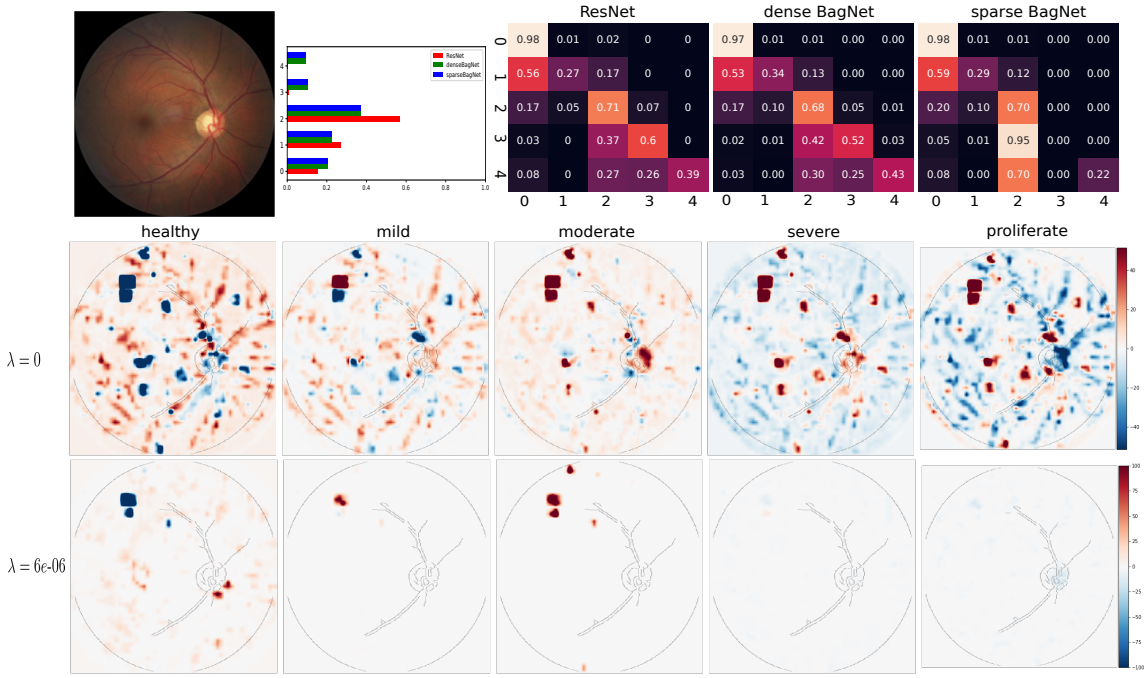

Figure 4: Overview of the multi-class evaluation. The first row shows an example of fundus image (grade 2) from the test set, followed by the classification probabilities obtained by each model (ResNet-50, dense and sparse BagNet), and the confusion matrix on the test set (normalized across rows). The second and third rows respectively show the class-specific activation maps of the dense and sparse model from healthy (grade 0) to proliferate (grade 4).

Although the sparse BagNet provided interpretable class evidence maps and achieved good accuracy, we noticed (see confusion matrix in Fig. 4) that it always failed to detect grade 3 and most often grade 4. Instead, it tended to classify these cases as moderate (grade 2). This result correlates with the significantly low kappa of the sparse model (0.780) compared to that of the ResNet-50 (0.826) and dense model (0.830).

4. Discussion and Conclusion

In this paper, we proposed an inherently interpretable classification model which provides sparse high-resolution class evidence maps. Enforcing sparse activations in the penultimate network layer directly caused fewer input regions to contribute to the classifier decision. The sparse model is therefore easier and potentially less time-consuming to inspect. We showed that the remaining relevant regions in the sparse model identified lesions with high precision, while healthy images yielded heatmaps with consistently negative evidence. On the other hand, diseased eyes contained both positive and negative evidence. This is expected, as the disease often manifests in small lesions, while many areas of the retina may appear healthy.

In our model, increased interpretability came at the price of a slight decrease in classification performance. This was expected and a direct consequence of the adding an additional constraint beyond cross-entropy to the objective function. However, appropriate tradeoffs and minimum performance will depend on requirements in real-world scenarios.

- mostly intended to work for small lesions. Use failures in severe grades as an example. On the other hand, for small lesions it is remarkable what subtle lesions the model can detect. This is useful in problem settings where local evidence for disease can be expected. We believe our approach lends itself to other clinical tasks that rely on small regions.

In summary, we believe our approach can be a useful tool to verify predictions and understand failure modes, and to facilitate clinicians' and patients' trust in ML systems.

although visualization-based interpretability methods provide a qualitative analysis that helps humans to understand the internal mechanism of DL models, we provided a quantitative evaluation of the proposed model

Acknowledgments

References

Wejdan L. Alyoubi, Wafaa M. Shalash, and Maysoon F. Abulkhair. Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, 20:100377, January 2020. ISSN 2352-9148. doi: 10.1016/j.imu.2020.100377. URL https://www.sciencedirect.com/science/article/pii/S2352914820302069.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*, 3(6), November 2021. doi: 10.1148/ryai.2021200267. URL https://pubs.rsna.org/doi/10.1148/ryai.2021200267.

Wieland Brendel and Matthias Bethge. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet, March 2019. URL http://arxiv.org/abs/1904.00760. arXiv:1904.00760 [cs, stat].

Mohamed Chetoui and Moulay A. Akhloufi. Explainable Diabetic Retinopathy using EfficientNET. pages 1966–1969. IEEE, July 2020. doi: 10.1109/EMBC44109.2020.9175664. ISSN: 2694-0604.

Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees,

Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, September 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0107-6. URL https://www.nature.com/articles/s41591-018-0107-6. Number: 9 Publisher: Nature Publishing Group.

Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90. ISSN: 1063-6919.

Yijin Huang, Li Lin, Pujin Cheng, Junyan Lyu, Roger Tam, and Xiaoying Tang. Identifying the key components in ResNet-50 for diabetic retinopathy grading from fundus images: a systematic investigation, October 2022. URL http://arxiv.org/abs/2110.14160. arXiv:2110.14160 [cs, eess].

ICO. International council of ophthalmology (ico) guidelines for diabetic eye care, 2017. URL https://icoph.org/eye-care-delivery/diabetic-eye-care/.

Indu Ilanchezian, Dmitry Kobak, Hanna Faber, Focke Ziemssen, Philipp Berens, and Murat Seçkin Ayhan. Interpretable gender classification from retinal fundus images using BagNets, June 2021. URL https://www.medrxiv.org/content/10.1101/2021.06.21.21259243v1.

Kaggle. Kaggle competition on diabetic retinopathy detection, 2015. URL https://www.kaggle.com/c/diabetic-retinopathy-detection/data. Accessed: 2022-11-30.

Ella Mahoro and Moulay A. Akhloufi. Applying Deep Learning for Breast Cancer Detection in Radiology. *Current Oncology (Toronto, Ont.)*, 29(11):8767–8793, November 2022. ISSN 1718-7729. doi: 10.3390/curroncol29110690.

Alexandros Papadopoulos, Fotis Topouzis, and Anastasios Delopoulos. An interpretable multiple-instance approach for the detection of referable diabetic retinopathy in fundus images. *Scientific Reports*, 11(1):14326, July 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-93632-8. URL https://www.nature.com/articles/s41598-021-93632-8. Number: 1 Publisher: Nature Publishing Group.

Mihir Rao, Michelle Zhu, and Tianyang Wang. Conversion and Implementation of State-of-the-Art Deep Learning Algorithms for the Classification of Diabetic Retinopathy, October 2020. URL http://arxiv.org/abs/2010.11692.

Paisan Ruamviboonsuk, , et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine*, 2(1):1–9, April 2019. ISSN 2398-6352. doi: 10.1038/s41746-019-0099-8. URL https://www.nature.com/articles/s41746-019-0099-8. Number: 1Publisher: Nature Publishing Group.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://www.nature.com/articles/s42256-019-0048-x. Number: 5 Publisher: Nature Publishing Group.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://arxiv.org/abs/1610.02391. arXiv:1610.02391 [cs].

Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Systems with Applications*, 128:84–95, August 2019. ISSN 0957-4174. doi: 10.1016/j.eswa.2019.01.048. URL https://www.sciencedirect.com/science/article/pii/S0957417419300545.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. URL http://arxiv.org/abs/1312.6034. arXiv:1312.6034 [cs].

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, April 2015. URL http://arxiv.org/abs/1412.6806. arXiv:1412.6806 [cs].

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. URL http://arxiv.org/abs/1905.11946. arXiv:1905.11946 [cs, stat].

Meysam Tavakoli and Patrick Kelley. *A comprehensive survey on computer-aided diagnostic systems in diabetic retinopathy screening*. December 2021. doi: 10.1088/978-0-7503-2060-3ch12. URL http://arxiv.org/abs/2208.01810. arXiv:2208.01810 [physics].

Qiaoying Teng, Zhe Liu, Yuqing Song, Kai Han, and Yang Lu. A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, 28(6):2335–2355, December 2022. ISSN 1432-1882. doi: 10.1007/s00530-022-00960-4. URL https://doi.org/10.1007/s00530-022-00960-4.

Daniel Shu Wei Ting, Gemmy Chui Ming Cheung, and Tien Yin Wong. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & Experimental Ophthalmology*, 44(4), May 2016. ISSN 1442-9071. doi: 10.1111/ceo.12696.

Zhiguang Wang and Jianbo Yang. Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation. In *Workshops at the thirty-second AAAI conference on artificial intelligence*. arXiv, December 2019. doi: 10.48550/arXiv.1703.10757. URL http://arxiv.org/abs/1703.10757. arXiv:1703.10757 [cs].

Tien Y Wong, Jennifer Sun, Ryo Kawasaki, Paisan Ruamviboonsuk, Neeru Gupta, Van Charles Lansingh, Mauricio Maia, Wanjiku Mathenge, Sunil Moreker, Mahi MK Muqit, et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*, 125(10):1608–1622, 2018. URL https://icoph.org/eye-care-delivery/diabetic-eye-care/.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks, November 2013. URL http://arxiv.org/abs/1311.2901. arXiv:1311.2901 [cs].

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*, 126(10):1084–1102, October 2018. ISSN 1573-1405. doi: 10.1007/s11263-017-1059-x. URL https://doi.org/10.1007/s11263-017-1059-x.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization, December 2015. URL http://arxiv.org/abs/1512.04150. arXiv:1512.04150 [cs].

Appendix A. Baseline architecture

A.1. ResNet-50

A.2. BagNet-33

Baseline BagNet architecture

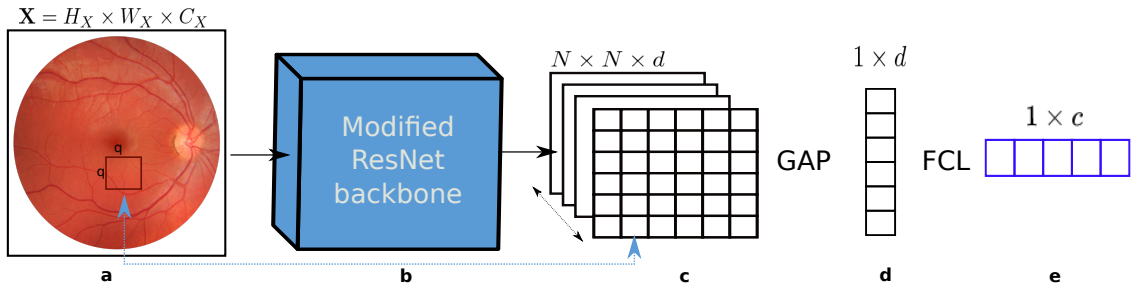

Figure 5: Baseline BagNet architecture

Appendix B. BagNet training details and performances

Moreover, common data augmentation techniques were used to counteract the small data size including the composition of rotation, flipping, cropping, and color jitter with a given probability.

B.1. Baseline implementations

B.2. Preprocessing and data augmentations

B.3. Hyperparameters

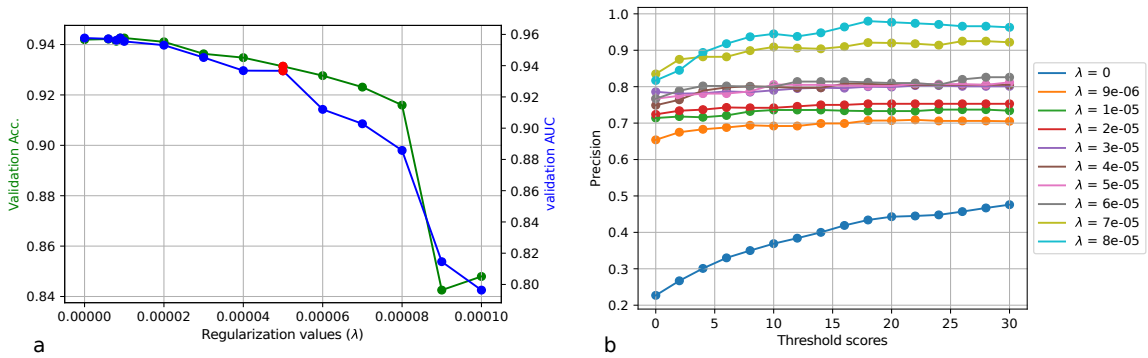

Figure 6: Comparison of the validation performances with different regularize values.

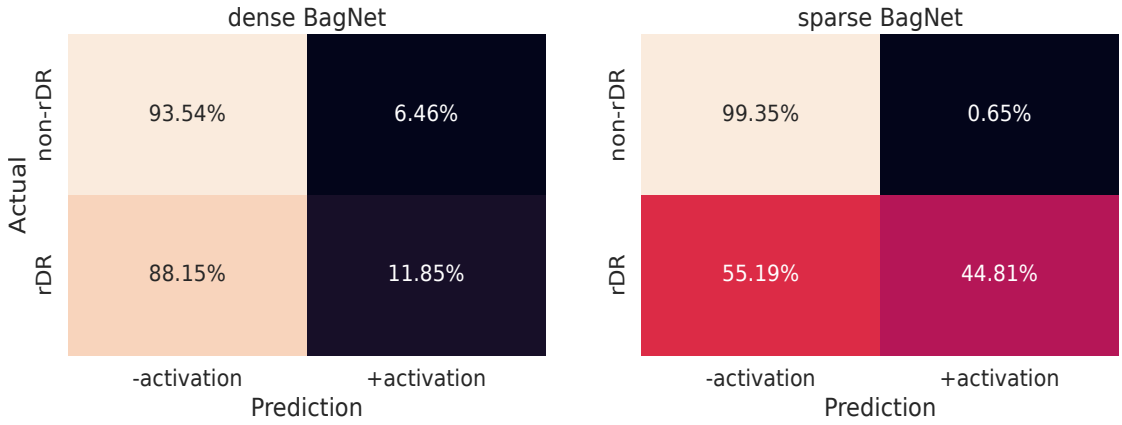

Figure 7: Heatmap evaluation.

B.4. Heatmap evaluation

Appendix C. Multi-class

	Acc.	**Kappa**
ResNet-50	0.862	0.826
Dense BagNet	0.864	0.830
Sparse BagNet	0.850	0.780

Table 3: Comparison of the classification performances on multiclass diabetic retinopathy detection task between the proposed approaches and the baseline ResNet-50 model on the test set.

Appendix D. Further visualizations

Appendix E. Figures/table

E.1. Classification performance

Table 1

- The results are reported on the test set.
- The models were saved based on the best performance of each individual metric on the validation dataset
- The overall classification performances of ResNet-50 and the dense BagNet are comparable except for the sensitivity which is clearly higher for ResNet-50 (Table 1).
- the regularizer strength ($\lambda = 5e$-5) for the sparse BagNet was chosen based on the tradeoff between the accuracy and the AUC on the validation set (Fig 6 an in Appendix B).

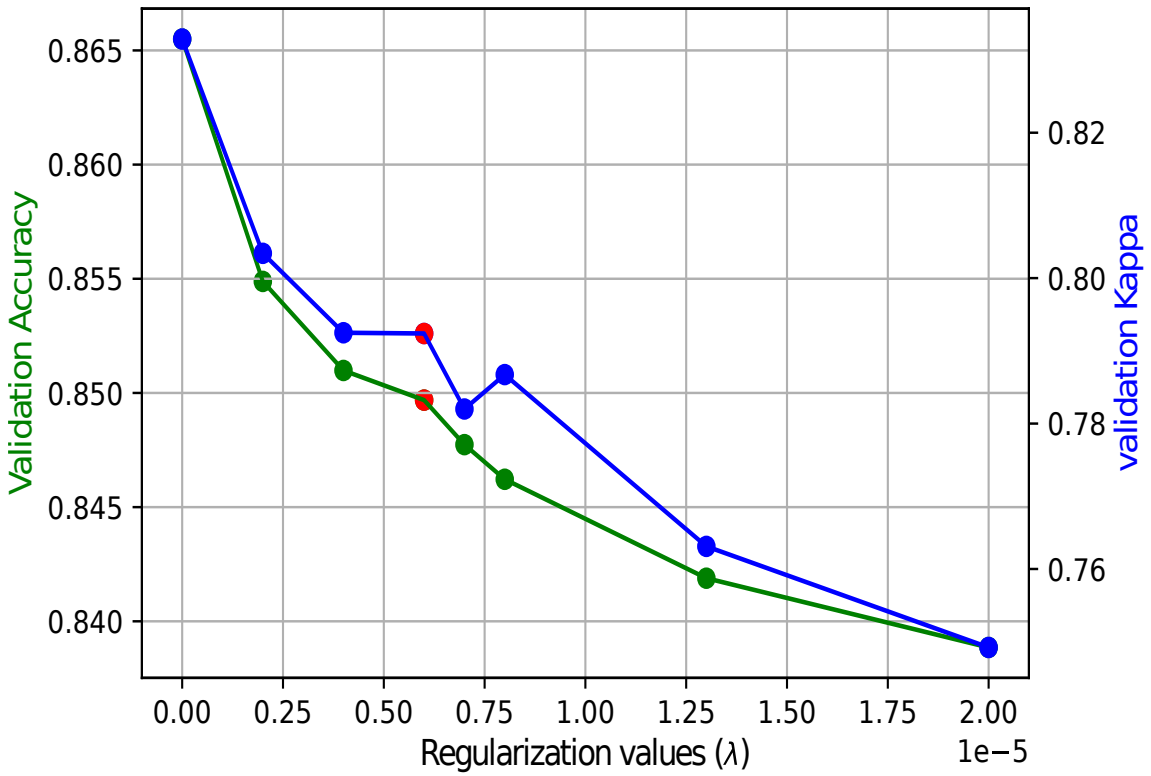

Figure 8: Comparison of the validation performance (Kappa vs accuracy) of the multiclass model on different regularize values.

- Due to the variety of datasets, the splitting strategies, and the method used to filter out the images with low quality, there are no reference baseline classification metrics for DR grading. However, from the state-of-the-art metrics reported by Papadopoulos et al. (2021), the average AUC of the models that use the Kaggle dataset to train and evaluate on rDR task is about 0.942.
- The results show that the dense BagNet's AUC is above the mean AUC of the reported SOTA, whereas the AUC of the sparse BagNet is still comparable to the mean SOTA's AUC.
- The classification results show that the BagNet is a good candidate model for rDR.
- A good hyper-parameter search could increase the classification performance of the dense and sparse model on rDR detection

E.2. Visualization maps

Figure 2

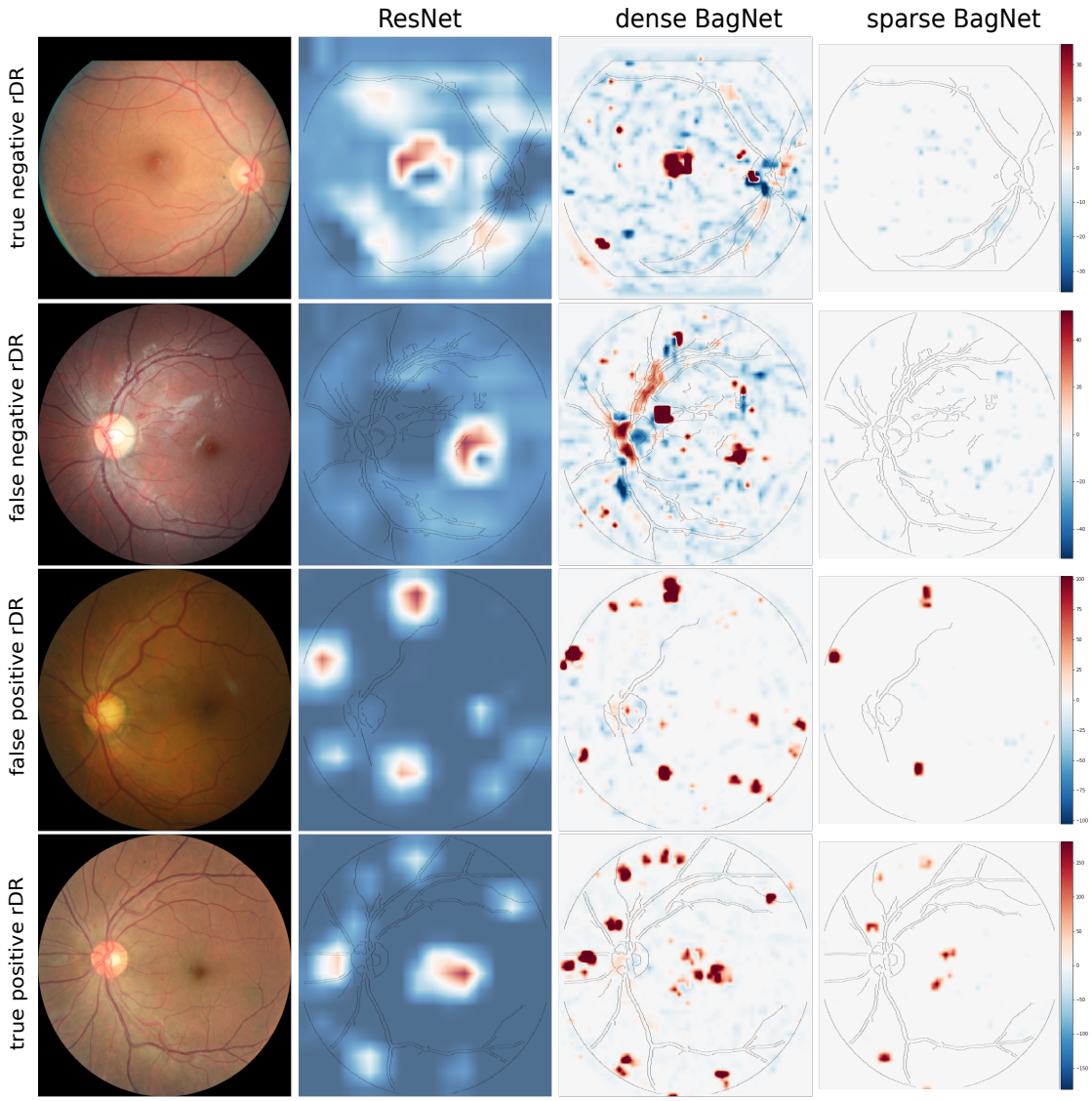

Figure 9: appendix

- The BagNet heatmap always points to the same location as the GradCAM with the difference of providing more fine-grained localization whereas the GradCAM provides coarse-grained localization.

- In addition to the fine-grained localization, the BagNet most often points to more suspicious regions compared to the GradCAM.

- The regularized BagNet heatmap is qualitatively much more interpretable with fewer activation than the dense BagNet as it filters out the activation regions with low confidence.

- On the healthy image, the sparse BagNet mostly provides a heatmap with negative activation compared to the GradCAM and the dense BagNet which always fail by having positive activation
- Negatives activation indicates that the image is healthy
- the activation on the sparse heatmap can be seen as a subset of dense heatmap activation

E.3. Annotation and localization

Figure 3

- The sparse model picks most regions that contain lesions and achieve high precision
- Some candidate patches provided by the model do not contain annotated lesions and need more investigation.
- "False positive" predictions could be due to the lesion-like appearance of image artifacts or healthy variations in pigmentation.

E.4. Heatmap evaluations

Table 2

- Inspired by the pointing game metric (Zhang et al., 2018) which considers a single location from each image to compute the precision, we computed the precision for the top$K$ patches from each image where k is the number of patches with a positive score.
- The experiments showed that although the dense BagNet can generate an interpretable heatmap, it contains a lot of positive activations in healthy areas (low localization precision) for non-heathy images (referable), and very few positive activations (T. fpos) for non-heathy images.
- the sparse BagNet has a better localization precision which increases as the regularizer strength increase (Appendix B Fig. 6b).
- The sparse BagNet leads to very few positive activations region ($\approx 0.65\%$) on healthy images compared to the dense BagNet ($\approx 6.46\%$), and to higher positive action (T. fpos) on non-heathy image compared to the dense BagNet.
- The fpos with the sparse BagNet are limited because on a non-healthy image, only a few regions can be non-healthy (see Appendix B, Fig. 7)
- These results show quantitatively that the dense heatmap contains many suspicious regions with low confidence that need to be filtered out.

E.5. Multi-class heatmap

Figure 4

- The BagNet provides a class-wise heatmap and highlights the lesions for each class
- The sparse BagNet fails on detecting class three and predict severe cases as moderate. This behavior deserves more investigation
- most grade 4 are classified as grade 2 which leads to low kappa
- The BagNet allows the extraction of a class-specific heatmap that highlights the discriminating part that triggers the decision for each class.

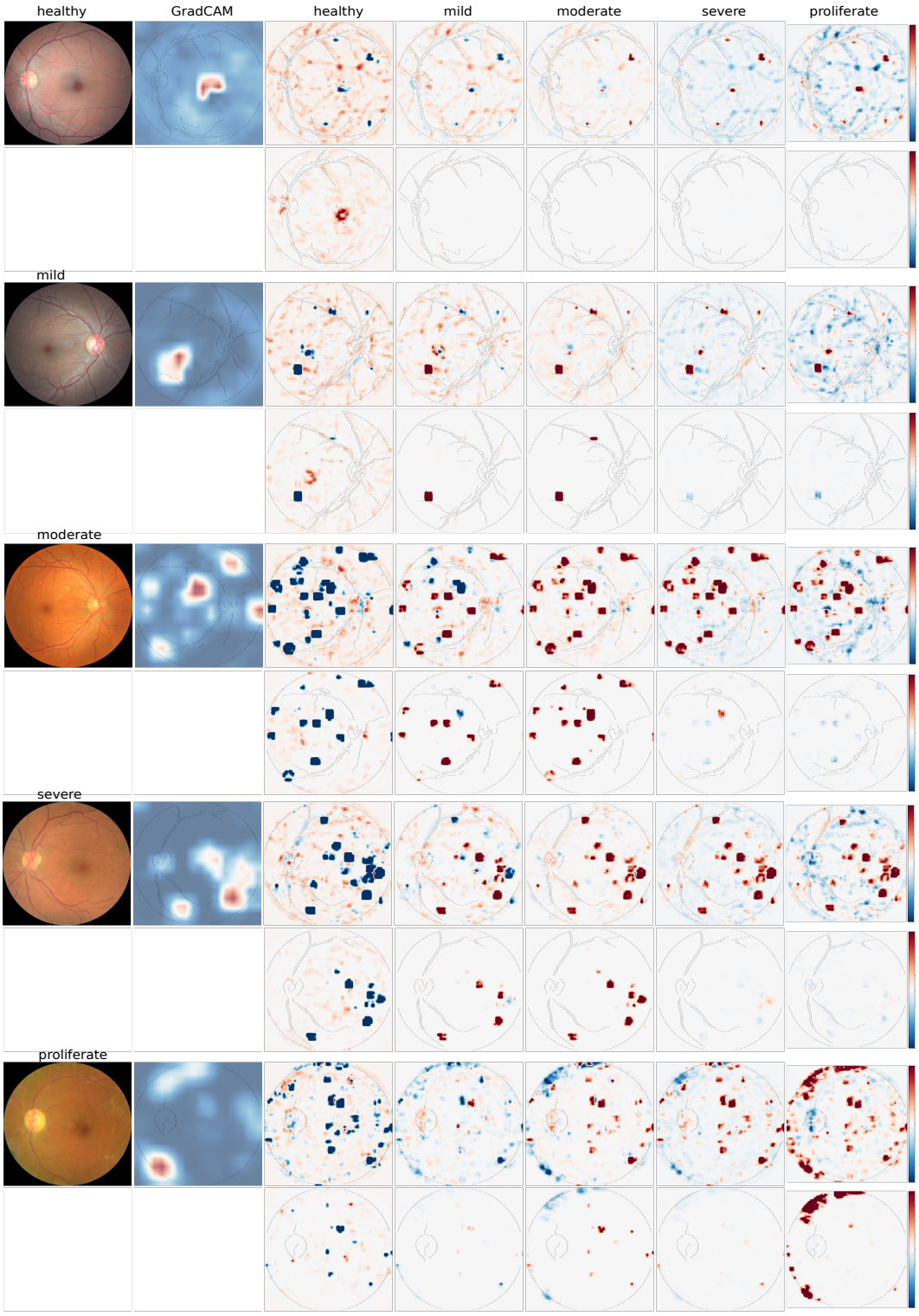

Figure 10: Multiclass heatmap