# Visual interpretability analysis of Deep CNNs using an Adaptive Threshold method on Diabetic Retinopathy images

George Ioannou Tasos Papagiannis Thanos Tagaris Georgios Alexandridis
Andreas Stafylopatis
Artificial Intelligence & Learning Systems Laboratory
School of Electrical & Computer Engineering
National Technical University of Athens
15780, Zografou, Greece
geoioannou@islab.ntua.gr, tasos@islab.ntua.gr, thanos@islab.ntua.gr,
gealexandri@islab.ntua.gr, andreas@cs.ntua.gr

## Abstract

*Deep neural networks have been dominating the field of computer vision, achieving exceptional performance on object detection and pattern recognition. However, despite the highly accurate predictions of these models, the continuous increase in depth and complexity comes at the cost of interpretability, making the task of explaining the reasoning behind these predictions very challenging. In this paper, an analysis of state-of-the-art approaches towards the direction of interpreting the networks' representations, is carried out over two Diabetic Retinopathy image datasets, IDRiD and DDR. Furthermore, these techniques are compared in the task of image segmentation of the same datasets. This is to discover which method can produce the better attention maps that can solve the problem of segmentation without actually training the network for the specific task. To accomplish that we propose an adaptive threshold method that transforms the attention masks in a more suitable representation for segmentation. Experiments over multiple architectures were conducted to ensure the robustness of the results.*

## 1. Introduction

Diabetic Retinopathy (DR) is an eye condition that affects the retina, the light sensitive layer of the eye which converts the light into electric signals, causing vision issues such as blurring, eye pain and blindness. Depending on the grade of the damage in the blood vessels there are four stages of DR, namely the mild, moderate, severe and proliferative. As there are no apparent symptoms in the early stages of the disease, automated diagnosis techniques are of high importance in order to detect it and prevent its developing as soon as possible.

A wide variety of deep neural network architectures have been tested in diabetic retinopathy screening as well as in medical image analysis in general, most of which rely on Convolutional Neural Networks (CNNs) [8] with additional modifications. On top of that, several algorithms and different approaches have been proposed for visualization and interpretation of such classifiers. The aim of this work is to investigate the ability of state-of-the-art interpretability methods to focus on the parts that have a major impact on the model's final prediction over CNN-based network architectures in the task of classifying diabetic retinopathy images. Additionally, the attention maps produced by these methods are compared with segmentation masks denoting the regions of the image which are significantly affected by the disease. For this purpose an adaptive thresholding method is employed, attempting to make the maps more objectively comparable.

The remainder of this paper is structured as follows; Section 2 overviews related approaches on the DR image classification task as well as on models' interpretability. Section 3 describes the different datasets, architectures and interpretation techniques that were implemented in the experimental phase. Section 4 introduces the proposed thresholding method and presents the results. Finally, the paper concludes in Section 5 and discusses potential future work.

## 2. Related Work

Several attempts have been made in the direction of automated recognition for diabetic retinopathy over the last years, most of which are based on deep learning. In 2016, Gulshan et al. trained a deep neural network based on the Inception-v3 architecture for DR detection in retinal fundus images [4]. The model was evaluated on two DR im-

age datasets, namely the EyePACS [6] and Messidor [2]. In terms of performance, the specificity and sensitivity of the algorithm at the high-sensitivity operating point were $93.4\%$ and $97.5\%$ respectively in the EyePACS dataset [6] and $93.9\%$, $96.1\%$ in the Messidor [2].

Ensemble methods have also been proposed, combining deep learning as well as classical machine learning algorithms in order to improve the individual models' predictive ability. In [12], a set of CNN-based models including ResNet, DenseNet, Inception and Xception were trained on the Kaggle dataset, effectively classifying retina images among all different stages of DR. Another ensemble model relying on Logistic Regression, $k$-NN, Decision Trees and Random Forest variations has been developed in [13]. Both approaches proved the superiority of the ensemble models over the individual algorithms and achieved high performance.

More recently, an architecture combining ResNet and Random Forest has been introduced for the task of DR degree classification [18]. In this method, the ResNet's averaged pooling layer has been used for the extraction of high level features from diabetic retinopathy images. Eventually, these feature maps were passed to a Random Forest classifier for the final class prediction. This approach outperformed state-of-the-art algorithms, achieving an accuracy of $96.0\%$ and $75.09\%$ on the Messidor[2] and EyePACS datasets [6] respectively.

Apart from the disease grade classification task, a wide variety of techniques have been developed under the scope of visualization and interpretation of the models' results. In [14], Gradient-weighted Class Activation Mapping (Grad-CAM) was introduced, an approach relying on the gradients of the last convolutional layer of a CNN-based model for producing an interpretable, localization map. This technique, which is compatible with any network architecture, makes use of the last layer's activation maps' gradients in order to create a heatmap, mapping each neuron to its corresponding importance weight for each class prediction. Subsequently, the map is upscaled to the dimensions of the original image, underlining the features that have the greatest impact on the classification process.

An enhanced version of the aforementioned method, Grad-CAM++, was presented in [1]. The authors used a weighted combination of the gradients to tackle the issue of small areas and multiple occurrences of a specific class fading away in the final saliency map. The results indicated that Grad-CAM++ outperformed the initial formulation of the algorithm in terms of object recognition explainability metrics as well as in human evaluation tests, and is considered to be among the state-of-the-art techniques concerning visual interpretability.

Sundararajan et al. proposed another gradient-based method to correlate the model's prediction with its inputs, called Integrated Gradients (IG) [15]. In this approach, a baseline is used in order to desaturate a well trained network and overcome the noisy-like gradients issue. This is achieved by scaling down the image's brightness and calculating the gradients on the new interpolated images. Finally, importance scores are attributed to the input features depending on their impact to the prediction, by averaging over these gradients.

In [10], Lundberg and Lee presented the concept of using Shapley values from cooperative game theory in order to measure the impact of each feature to a model's prediction. Shapley Additive Explanation values (SHAP) describe the features' average marginal contribution over all possible combinations and can therefore be considered as features' importance. Several approaches for estimating the Shapley values were also proposed, in order to overcome the computational expense due to the high complexity and make it possible to apply this technique in practice. In this paper, the above four methods are analyzed, however, there are many techniques that try to interpret the predictions of a model [19].

## 3. Experimental Framework

In this section, the experimental framework of this paper is presented. We use two DR datasets for training the neural networks and, then, we compare the interpretability methods in different setups. The next subsections present the whole process in more detail.

| Grades of DR | IDRiD | | DDR | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| No DR | 134 | 34 | 3133 | 1880 |
| Mild | 20 | 5 | 315 | 189 |
| Moderate | 136 | 32 | 2238 | 1344 |
| Severe | 74 | 19 | 118 | 71 |
| Proliferative (PDR) | 49 | 13 | 456 | 275 |
| Ungradable | - | - | 575 | 346 |
| Total | 413 | 103 | 6835 | 4105 |

Table 1. Distribution of the DR grading labels of the images for IDRiD and DDR

| Types of masks | IDRiD | DDR |
|---|---|---|
| MA | 81 | 570 |
| HE | 80 | 601 |
| EX | 81 | 486 |
| SE | 40 | 239 |
| OD | 81 | - |
| Total | 81 | 757 |

Table 2. Distribution of different types of segmentation masks for IDRiD and DDR

### 3.1. Diabetic Retinopathy Datasets

The first dataset that this study will use is the Indian Diabetic Retinopathy Image Dataset (IDRiD) [11]. IDRiD is a fundus image dataset of the Indian population and it contains three tasks: DR grading, segmentation and localization. In this paper, we are going to focus on the DR grading and segmentation tasks. The DR grading set consists of 516 images stored in JPEG format. The images have a resolution of $4288 \times 2848$ pixels which is considered to be of very high quality. The purpose of the task is to predict the grade of severity of the retinopathy dividing the dataset into 5 classes. The severity ranges from grade 0 (which indicates no apparent DR) to grade 1 (mild DR), grade 2 (moderate DR), grade 3 (severe DR) and, lastly, grade 4 (proliferative DR). The dataset is split to a training set, consisting of 413 images, and a test set of 103. Table 1 shows the distribution of classes in both sets. The segmentation task consists of 81 images, divided in 54 for training and 27 for testing. There are 5 different kind of segmentation masks: Microaneurysms (MA), Haemorrhages (HE), Hard Exudates (EX), Soft Exudates (SE) and Optic Disk (OD). The distribution of the above can be found on Table 2.

The second dataset to be considered is DDR [9]. Similar to IDRiD, there exist the same 3 tasks for this dataset, as well. The DR grading task consists of 1.3673 color fundus images taken from 147 hospitals. Images are categorized to 6 classes, with the first 5 being the same as in IDRiD and the $6^{th}$ class containing images with poor quality that cannot be categorized anywhere else. This class is called Ungradable. The distribution of the classes in the training and the test sets can be seen in Table 1. Unlike IDRiD, image resolutions vary. For example, there exists images of $1.137 \times 1.470$ pixels along with images of $3456 \times 5184$ pixels. On the other hand, the segmentation task consists of 757 images, provided with the same kind of segmentation masks, like IDRiD, except for the Optic Disc. The distribution of the masks is shown on Table 2.

### 3.2. Network architectures and Interpretability

In the current work, three different CNN-based network architectures were examined in order to verify the experiments' objectivity. More specifically the DenseNet [5], InceptionV3 [16] and EfficientNetB0 [17] models were trained for DR grading classification on IDRiD and DDR datasets separately, resulting in 6 models. All models were initialized with pretrained weights from ImageNet [3]. Image augmentation was performed using random zooms, rotations and flips. All images, from both datasets, were resized in $300 \times 300$ pixels, in order to make it easier to train the networks. The networks were trained for 20 epochs using the Adam optimizer [7]. The 6 networks achieved the test accuracies shown in Tables 3 and 4 for both datasets, respectively. It is highlighted that EfficientNetB0 is the most suitable architecture for this task with an overall accuracy of $60.19\%$ on IDRiD and $73.56\%$ on DDR. A common shortcoming of all three networks seems to be their difficulty in detecting the first stage (mild) of the disease, like most state-of-the-art approaches, while they perform quite satisfactorily in the moderate and no-DR classes.

Concerning interpretability, we focus on four state-of-the-art techniques to describe the functionality of the models and explain the logic behind the images' classification. More specifically, four gradient-based algorithms are examined in this study, GradCAM, GradCAM++, Integrated Gradients and SHAP with the expected gradients' approximation method. Each method is applied on every network with respect to the original image in order to produce an attention map, denoting the areas of the image that contributed the most to the network's prediction. Layers of different depth were used for the maps' construction, depending on the networks' architecture. Particularly, GradCAM and GradCAM++ were applied on the $79^{th}$, $15^{th}$ and $15^{th}$ layer of DenseNet, Inception and EfficientNet respectively. The Integrated Gradients method is independent of selecting a specific layer as the gradients are calculated with respect to the interpolated images in contrast to GradCAM and GradCAM++ which make use of the selected layers' outputs. Finally, the SHAP algorithm has been applied to network input.

Figure 1 shows the attribution masks produced by the aforesaid techniques after being applied on the Efficient-

| Architecture | No DR | Mild | Moderate | Severe | PDR | Total Accuracy |
|---|---|---|---|---|---|---|
| DenseNet | 0.7647 | 0.0 | 0.7812 | 0.1578 | 0.1538 | 0.5436 |
| InceptionV3 | 0.8235 | 0.2 | 0.6250 | 0.5789 | 0.0 | 0.5825 |
| EfficientNetB0 | 1.0 | 0.0 | 0.5937 | 0.3684 | 0.1538 | 0.6019 |

Table 3. Test Accuracies of each class and Total Accuracy among the 3 architectures for the IDRiD dataset

| Architecture | No DR | Mild | Moderate | Severe | Proliferative | Ungradable | Total Accuracy |
|---|---|---|---|---|---|---|---|
| DenseNet | 0.8202 | 0.0 | 0.6279 | 0.0 | 0.2145 | 0.8063 | 0.6635 |
| InceptionV3 | 0.9904 | 0.0 | 0.4211 | 0.1267 | 0.5636 | 0.9682 | 0.7130 |
| EfficientNetB0 | 0.9574 | 0.0 | 0.5610 | 0.014 | 0.4727 | 0.9682 | 0.7356 |

Table 4. Test Accuracies of each class and Total Accuracy among the 3 architectures for the DDR dataset

Net model (trained on IDRiD), for a sample DR image of the IDRiD dataset, as well as their overlays on the original image. Concerning the GradCAM and GradCAM++ algorithms, the latter highlights a larger region of the image, as expected, making it difficult to extract useful information. Regarding the Integrated Gradients, it tends to produce smaller areas of attention, although the main highlighted regions are common with the ones indicated by the above-stated methods in most cases. SHAP mask values are of low order of magnitude and thus, are not easily comparable with the rest of the approaches. Also, from Figure 1, we can see that there are barely any highlighted areas of attention in the SHAP masks.

All four techniques have their merits, however, from a medical perspective they fail to capture the important areas of the eye that can be beneficial in specific disease diagnosis. Due to this drawback, this study will focus on transforming the attribution masks of each method, in order to achieve a better interpretation performance regarding a lesion segmentation task.

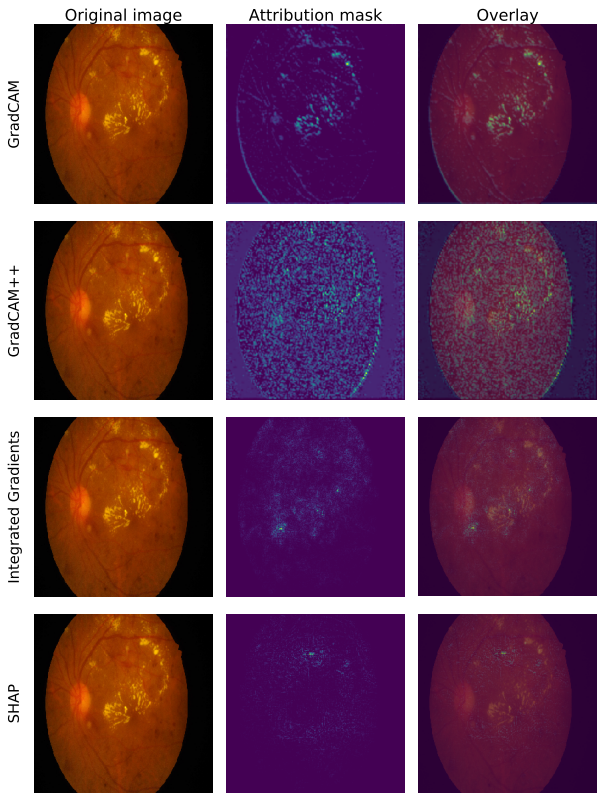


Figure 1. Attribution masks on DR image (EfficientNet model)

## 4. Adaptive Threshold and Segmentation

As demonstrated above, it is difficult to evaluate the attention masks in their current form since they are not comparable with each other. This is mainly due to the significantly different sizes of the areas being indicated as crucial for the classification by each method. On top of that, same values on different attribution masks might indicate different order of importance. For example, a high value in a SHAP mask can be considered low in comparison with the values of a GradCAM mask. Subsequently, these algorithms are compared under the scope of segmentation with respect to the different types of observed lesions.

Even though the tasks of DR grading and lesion segmentation are different, there is an overlap in the grade of the retinopathy with the lesions found in areas of the eye. Any overlapping between these areas of the lesions and the attention masks can accurately show which interpretability technique (and which model in consequence) can spot the most significant parts of the eye that can classify the grade of the retinopathy. In order to evaluate the attribution masks, we combined the segmentation masks of all five lesions to a final unified mask for each image in the dataset. This can be seen in Figure 2, where an original image is shown along with the total segmentation mask (including all types: MA, HE, EX, SE and OD) and the overlay of the two.

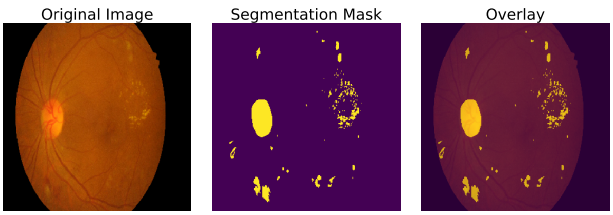


Figure 2. Unified segmentation mask applied on an original image

In order to overcome the above stated peculiarities of the different masks we performed a thresholding scheme on the attention masks to round the values of each pixel to $\{0, 1\}$. This is also required for the comparison with the segmentation groundtruths since the values of the attention masks are continuous in the range $[0, 1]$ while the segmentation masks consist of discrete binary values, as can be seen in Figures 1 and 2. For this task, we propose a method based on statisti-

cal properties of the images in order to produce an adaptive threshold, individually for each mask. More specifically, the threshold value is defined according to Equation 1:

$$threshold = median + k * std \quad (1)$$

where the median and standard deviation are calculated over the pixels of the mask. Parameter $k$ adjusts the threshold depending on the median and the percentage of pixels originally highlighted as significant for the model's prediction (i.e. non-zero pixels) according to the following formula:

$$k = [a + \ln(1 + pct)] * (1 - median) \quad (2)$$

In Equation 2, $pct$ stands for the percentage of non-zero values over the total size of the image while $a$ is a constant. The intuition behind $pct$ is that the threshold value should be larger as the percentage of important pixels in the mask increases in order to balance the amount of selected pixels and make it comparable among the different masks. The purpose of the logarithm function is to prevent the threshold from being too high for large values of percentage as this would lead to non informative masks, consisting of very few descriptive points. Concerning the contribution of the median, high values indicate left skewed distributions, meaning a smaller increase is needed in order to avoid cutting off a very large percentage of points. On the contrary, low median values demand a higher increase so as to not produce an excessive mask. Finally, the constant defining the lower bound of the parameter was set to $1.2$ after experimentation.

Figure 3 shows the segmentation masks produced by the interpretability methods on the trained models after applying the adaptive threshold. The original image and its segmentation mask are the ones depicted in Figure 2. It is clear that the new masks are of the same order of magnitude and can be easily compared, as opposed to the original masks, similar to the ones in Figure 1. Additionally, there seem to be many similarities among the masks produced by the same interpretability technique across the different architectures. Among the methods, GradCAM and GradCAM++ visibly underline similar patterns as the points with the major impact on the models' decision. On the other hand, the SHAP masks appear to be more scattered but seem quite similar to the IG masks in the majority of images.

The fact that all four techniques focus on quite similar parts of the images and share some regions of interest with the segmentation mask boosts the likelihood of detecting the most important areas. Thereafter, the Intersection over Union (IoU) score is calculated for the four interpretability methods separately, in order to evaluate the performance of each technique after the application of the proposed threshold. For this task, $81$ segmentation-labeled images of the IDRiD dataset were used. The segmentation masks of the DDR dataset's images consisted of very few and sparse points and thus they were not considered suitable for the purpose of this study.

| Interpretability Method | Training Dataset | | Overall |
|---|---|---|---|
| | IDRiD | DDR | |
| GradCAM | 0.0563 | 0.0634 | 0.060 |
| GradCAM++ | 0.0547 | 0.0560 | 0.055 |
| Integrated Gradients | 0.0544 | 0.0880 | 0.071 |
| SHAP | 0.0763 | 0.0985 | 0.087 |

Table 5. Intersection over Union score for interpretability masks

Table 5 presents the IoU scores for the masks produced by the models trained on the IDRiD and DDR datasets separately as well as the total IoU scores for each interpretability method over all different architectures. Among the examined approaches, SHAP clearly achieved higher IoU scores in both setups, around $0.087$. Concerning the other techniques, Integrated Gradients performance was close to SHAP with an overall IoU of $0.071$ while GradCAM and GradCAM++ reached $0.06$ and $0.055$ IoU scores, respectively. Naturally, the above performances are affected by the thresholds applied on the original masks. Different thresholding methods might produce modified attention maps, leading to different results. It is observed that, even though the evaluation set consists of images of the IDRiD dataset, models trained on DDR reach higher scores. This can be attributed to the fact that DDR is a bigger dataset, leading to more complete trained models. Because our models are trained on the classification task, the IoU scores of the interpretability methods are much lower than any state-of-the-art approach on segmentation. Thus, the previously mentioned scores are not comparable to models trained explicitly in the segmentation task.

## 5. Conclusion

The present work sets out to study and analyze the performance of state-of-the-art algorithms for visual interpretability of neural networks. Specifically, GradCAM, GradCAM++, Integrated Gradients and SHAP methods are examined on the task of Diabetic Retinopathy grading classification. Multiple architectures were trained along two datasets, IDRiD and DDR, to setup a complete framework for comparing the above methods. Additionally, a statistical based adaptive thresholding method has been proposed in order to transform the attention maps, aiming to make them comparable to the groundtruth segmentation masks. This can result in a more robust and objective way of evaluating attention masks and interpretability methods, in general. The interpretability masks produced after applying the thresholding technique are qualitatively comparable with each other as well as with the label segmentation mask. Obviously, in terms of quantitative metrics, such as IoU, they still lack the ability to perform in a competitive scenario

with other state-of-the-art segmentation techniques.

It would be quite an interesting path of work to examine if these interpretability methods can be modified in a way to compete in a segmentation task. Naturally, in order for this to be feasible, there is a need of datasets with aligned tasks of classification and segmentation. This way, a model trained for classification could, at the same time, be able to solve the segmentation problem, as well. A promising idea is to incorporate the segmentation masks in the classification task, leading to models that learn to focus on more significant areas of the image.

## References


[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2018. 2

[2] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein. Feedback on a publicly distributed database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, Aug. 2014. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 12 2016. 1

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017. 3

[6] Kaggle and EyePacs. Kaggle diabetic retinopathy detection, jul 2015. 2


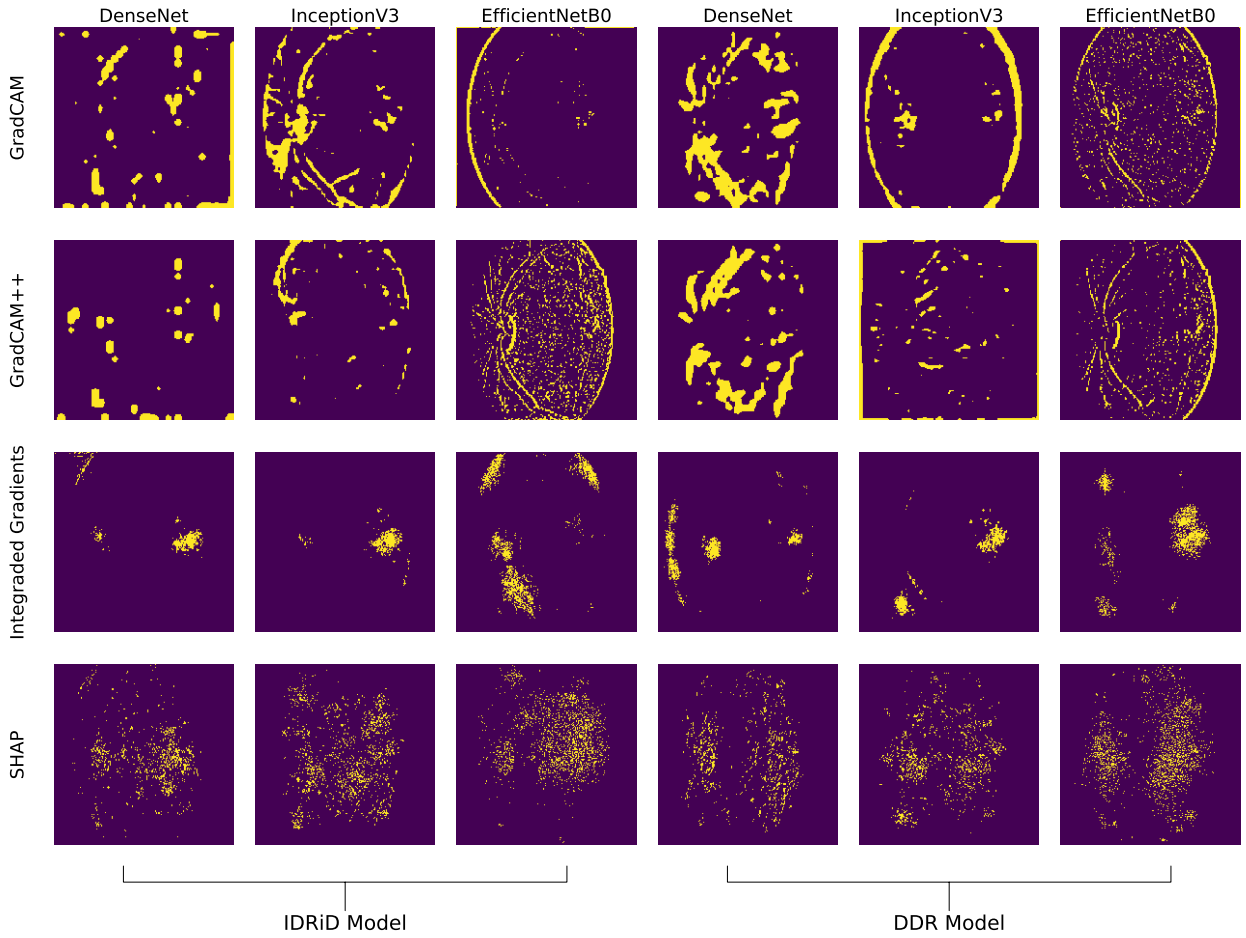


Figure 3. Segmentation masks with adaptive threshold for different interpretability techniques and network architectures

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[9] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.*, 501:511–522, 2019. 3

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017. 2

[11] Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, Jianzong Wang, Xinhui Liu, Liangxin Gao, Tianbo Wu, Jing Xiao, Fengyan Wang, Baocai Yin, Yunzhi Wang, Gopichandh Danala, Linsheng He, Yoon Ho Choi, and Fabrice Mériaudeau. Idrid: Diabetic retinopathy - segmentation and grading challenge. *Medical Image Anal.*, 59, 2020. 3

[12] Sehrish Qummar, Fiaz Gul Khan, Sajid Shah, Ahmad Khan, Shahaboddin Shamshirband, Zia Ur Rehman, Iftikhar Ahmed Khan, and Waqas Jadoon. A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, 7:150530–150539, 2019. 2

[13] G Thippa Reddy, Sweta Bhattacharya, S Siva Ramakrishnan, Chiranji Lal Chowdhary, Saqib Hakak, Rajesh Kaluri, and M Praveen Kumar Reddy. An ensemble based machine learning model for diabetic retinopathy classification. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–6, 2020. 2

[14] 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, Oct 2019. 2

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. 2

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016. 3

[17] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 3

[18] Muhammad Kashif Yaqoob, Syed Farooq Ali, Muhammad Bilal, Muhammad Shehzad Hanif, and Ubaid M Al-Saggaf. Resnet based deep features and random forest classifier for diabetic retinopathy detection. *Sensors*, 21(11):3883, 2021. 2

[19] Quanshi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers Inf. Technol. Electron. Eng.*, 19(1):27–39, 2018. 2