Dr. Pradeep, 2026, 14:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Evaluating Localization and Decision Faithfulness Of Grad-Cam Explanations In Diabetic Retinopathy Grading

Dr. Pradeep[1], Abhendra Pratap Singh[2], Shree Bhagwan[3], Shivam Tiwari[4], Sushant Kumar[5]

[2]department Of Mechanical Engineering, Hmritm, Hamidpur, New Delhi 110036, India

[1,3,4,5] department Of Computer Science And Engineering, Hmritm, Hamidpur, New Delhi 110036, India

Abstract: Gradient weighted Class Activation Mapping (Grad-CAM) is widely used to show which image regions influence a deep learning model in diabetic retinopathy (DR) grading. A heatmap that overlaps annotated lesions does not prove that the model used those regions for its decision. This study measures two properties of Grad-CAM explanations separately. These are lesion localization and decision faithfulness. Two ResNet-50 models were built. Model A was trained on APTOS 2019 and Model B was obtained by training Model A further on IDRiD grading images. The explanations were evaluated on 81 IDRiD images that carry expert lesion masks. Localization was measured with Intersection over Union (IoU) the Dice coefficient and the pointing game. Faithfulness was measured with a deletion test that compares Grad-CAM guided removal with random removal of the same area. On the 103 image IDRiD test set accuracy rose from 34.95% for Model A to 61.17% for Model B and quadratic weighted kappa rose from 0.563 to 0.746. At a threshold of 0.5 the mean IoU was only 0.043 for Model A and 0.045 for Model B. Even so Grad-CAM guided deletion reduced prediction confidence much faster than random deletion for both models. The two models did not differ significantly in any explanation metric and the correlation between localization and faithfulness was weak. These results show that lesion overlap alone is an incomplete measure of explanation quality in DR grading.

Keywords: Diabetic retinopathy, explainable artificial intelligence, Grad-CAM, lesion localization, decision faithfulness, ResNet-50.

I. INTRODUCTION

Diabetic retinopathy (DR) is a retinal complication of diabetes. Automated analysis of colour fundus photographs has become an important research area because it can support screening and disease grading. Work in this area depends on datasets that offer reliable disease labels and where possible detailed lesion annotations. The Indian Diabetic Retinopathy Image Dataset (IDRiD) provides fundus images with image level DR and diabetic macular edema grades along with pixel level annotations of retinal lesions [1]. The IDRiD segmentation and grading challenge created a common setting for algorithms that address these related tasks [2]. The APTOS 2019 dataset contains fundus photographs graded by clinicians on five DR severity levels and includes variation in image acquisition conditions [3].

Convolutional neural networks are widely studied for image classification. Large image collections such as ImageNet have supported the pretraining of these models [4]. Residual learning made it practical to train much deeper networks and forms the basis of the ResNet architecture [5]. This study uses a ResNet-50 backbone with ImageNet pretrained weights for five class DR grading.

Surveys describe a wide range of machine learning and deep learning approaches for automated DR detection and classification [6]. A predicted grade however does not show which retinal regions led to it. For medical images a reader needs to see the evidence behind a prediction. Classification performance and the quality of the visual explanation therefore have to be judged as two separate things [7].

Several attribution methods have been proposed to make deep networks easier to interpret. Gradient weighted Class Activation Mapping (Grad-CAM) uses the gradients that flow into a convolutional layer to

© 2026 Dr. Pradeep, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

produce a coarse map of the regions that influence a predicted class [8]. Integrated Gradients is another attribution method and is built on the axioms of sensitivity and implementation invariance [9]. A heatmap can look convincing without being a true account of what the model used. Samek et al. proposed region perturbation as an objective way to evaluate heatmaps instead of relying on visual inspection alone [10]. The deletion test used in this study follows the same idea.

This study asks whether a Grad-CAM map that overlaps expert lesion annotations also marks the regions that the model depends on. Two ResNet-50 models are compared. One is trained on APTOS 2019 and the other is adapted to IDRiD. Localization is measured against expert annotations. Faithfulness is measured with deletion experiments and random controls. The relation between the two is then tested directly. The contributions of this work are as follows.

- A paired evaluation of localization and faithfulness on the same Grad-CAM maps for 81 IDRiD images that carry expert lesion masks.
- A comparison of a model trained on APTOS only with the same model adapted to IDRiD to see whether higher accuracy gives better explanations.
- Robustness checks that use random controls and blur replacement in the deletion test and removal of the retinal edge band.
- An exploratory analysis by DR grade and a comparison of Grad-CAM with Grad-CAM++ and LayerCAM.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes the datasets and methods. Section IV presents the results. Section V discusses the findings and the limitations. Section VI concludes the paper and Section VII outlines future work.

II. RELATED WORK

1. Evaluation Of Visual Explanations

Petsiuk et al. introduced RISE which estimates the importance of image regions by observing how randomly masked inputs change a model prediction [11]. Adebayo et al. ran a set of sanity checks to test whether saliency maps respond to the learned parameters and the training data of a model [12]. Their work showed that a map can look plausible and still say little about what the model has learned. Yeh et al. defined measures of infidelity and sensitivity so that an explanation can be judged by how it relates to model behaviour under input perturbation [13].

Other studies address explanation evaluation more broadly or for medical images. Gomez et al. examined metrics used to evaluate saliency maps [14]. Lamprou et al. evaluated deep learning interpretability methods for medical images under the scope of faithfulness [15]. Brima and Atemkeng combined visual explainability with statistical quantitative analysis for saliency driven medical imaging [16]. Lago et al. described a framework for the systematic assessment of explainability features in medical imaging [17]. Ozer et al. studied consistent explainable image quality assessment for medical imaging [18].

2. Grad-Cam In Diabetic Retinopathy

Jiang et al. proposed a multi label deep learning model that combines DR classification with Grad-CAM based localization of different lesion types [19]. Van Craenendonck et al. compared ten heatmapping techniques for DR lesion detection on IDRiD using expert pixel level lesion annotations. They found that the best combination of network and heatmapping method differed across lesion types and they reported considerable disagreement between the heatmaps and the expert annotations [20]. Meng et al. proposed a framework in which ophthalmologists take part in refining the localization of retinal disease for fundus image classification [21]. Shorfuzzaman et al. developed an explainable ensemble model for DR grading that uses Grad-CAM and SHAP explanations and was evaluated on APTOS Messidor and IDRiD data [22].

Ghadai et al. described a clinically interpretable pipeline for DR classification that uses EfficientNet with advanced data augmentation and Grad-CAM [23]. Persada et al. reviewed visual explanation methods based on class activation mapping for DR [24]. Cascone et al. benchmarked explainable AI methods for DR analysis based on Vision Transformers [25].

3. Clinical Interpretation And Reliability Of Heatmaps

Malerbi et al. compared Grad-CAM heatmaps from retinal images of the same eyes taken before and after pupil dilation. They reported that small differences in image quality were linked to changes in model scores and heatmap patterns [26]. Biswas et al. studied whether MobileNet based DR classifiers follow decision criteria similar to those of human retina specialists. They used Grad-CAM and Integrated Gradients together with lesion annotations and found that the classifiers did not consistently follow the International Clinical Diabetic Retinopathy grading criteria. Activations appeared in both lesion and non lesion regions [27]. Zhang et al. applied seven interpretability methods to four network architectures on a clinically annotated dataset of 2060 fundus images. They assessed the results with qualitative and quantitative measures that include an entropy based evaluation of saliency maps [28]. Civit-Masot et al. introduced an explainable ensemble for five class DR grading with a confidence quality factor and configurable heatmaps [29].

4. Research Gap

A heatmap can overlap with expert lesion annotations without the model prediction depending on that region. In the same way the removal of a heatmap region can change a prediction strongly even when the region overlaps the annotations poorly. Localization and faithfulness are therefore different properties and need separate measurement. This study measures both for the same Grad-CAM maps of two ResNet-50 DR grading models. Localization is measured against expert lesion masks and faithfulness through deletion experiments with random controls. The relation between the two is then tested directly.

III. MATERIALS AND METHODS

1. Study Design

This is a quantitative experimental study. Fig. 1 shows the evaluation pipeline. A fundus image is preprocessed and graded by a ResNet-50 model. Grad-CAM then produces a map for the predicted class. The map is evaluated for localization against expert lesion masks and for faithfulness through a deletion test. The two sets of scores are then compared by correlation analysis. The properties were analysed separately because overlap with an annotated lesion does not show that the region decides the prediction.

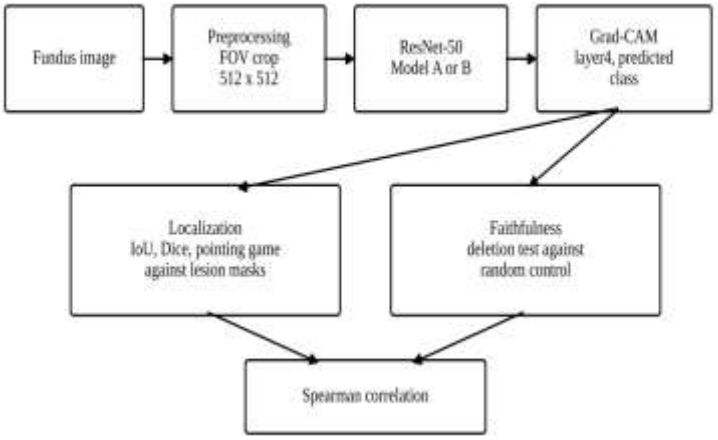

Fig. 1. Evaluation pipeline used in this study.

2. Datasets And Cohorts

Three public datasets were used. APTOS 2019 contains 3662 fundus images labelled with five DR severity grades from 0 to 4. A stratified split of 80 to 20 gave 2929 training images and 733 validation images. These images were used only for Model A.

The IDRiD grading set contains 516 images. Of these 413 form the training partition and 103 form the held out test partition. The test partition was never used for training or validation or checkpoint selection of either model. To keep the explanation cohort independent 57 training images that overlap with the segmentation cohort were removed from the pool used for Model B. The remaining 356 images were divided into 284 training images and 72 validation images with stratification and random seed 42.

The IDRiD segmentation set contains 81 images with expert pixel level annotations for microaneurysms (MA) haemorrhages (HE) hard exudates (EX) and soft exudates (SE). It also provides an annotation of the optic disc. Of these 81 images 57 overlap with the IDRiD training partition and three overlap with the test partition. The other 21 appear only in the segmentation set. Reference DR grades were available for 60 of the 81 images. These were 37 images of Grade 2 and 14

images of Grade 3 and 9 images of Grade 4. No grade was inferred for the remaining images. All 81 images were used to evaluate the explanations of both models.

3. Image Preprocessing

All images were cropped to the retinal field of view and resized while keeping the aspect ratio. Symmetric padding then gave a fixed size of 512 by 512 pixels. Images were normalized with the ImageNet channel mean and standard deviation. Training images were augmented with horizontal flipping and rotation and changes in brightness and contrast.

4. Classification Models And Training

Both models use the ResNet-50 implementation of torchvision with ImageNet pretrained weights. The classification head was replaced to give five outputs for the ICDR grades 0 to 4 and a dropout rate of 0.3 was used. Training used class weighted cross entropy loss with weights inversely proportional to class frequency.

Model A was trained on the APTOS training split with the AdamW optimizer. The initial learning rate was 1e-4 and the weight decay was 1e-4. A cosine annealing schedule was used with batch size 8 on an NVIDIA Tesla T4 GPU with 16 GB memory. Early stopping monitored the validation quadratic weighted kappa with a patience of 5 epochs and a minimum improvement of 0.001. Training ran for 13 epochs and the checkpoint from epoch 8 was kept because it gave the highest validation kappa. The random seed was 42.

Model B was initialized from the weights of Model A and trained further on the 284 IDRiD training images. AdamW was used with a learning rate of 5e-5 and a weight decay of 1e-4 together with class weighted cross entropy. Training was limited to 25 epochs with a patience of 8 epochs and stopped after 18 epochs. The checkpoint from epoch 10 had the best validation kappa and was selected.

Three other variants were also tried. They changed the learning rate or froze layers or added label smoothing. None of them beat the validation kappa of Model B by the preset margin of 0.02 so Model B was kept. Both models were evaluated on the same 103 image IDRiD test set.

5. Classification Metrics

Accuracy was computed as the share of correct predictions.

6. Accuracy = Ncorrect / Ntotal (1)

Quadratic weighted kappa (QWK) was used because DR grades are ordered and larger grading errors should cost more. Within one grade accuracy is the share of predictions that differ from the reference grade by at most one.

$$WOGA = (1 / N) \Sigma I(| y_i - \hat{y}_i | \leq 1), \; i = 1 \text{ to } N \quad (2)$$

Here y is the reference grade and ŷ is the predicted grade. Ninety five percent confidence intervals were obtained with the percentile bootstrap. Differences between the models were also estimated by a paired bootstrap because both models were tested on the same images.

7. Grad-Cam Generation

Grad-CAM was generated with the pytorch grad-cam package for the predicted class. The target layer was the final convolutional stage layer4 which gives a map of 16 by 16 cells. Each map was upsampled to 512 by 512 pixels and normalized to the range 0 to 1. Binary regions were made at thresholds of 0.3 and 0.5 and 0.7. The maps are treated as attribution visualizations and not as lesion segmentations.

8. Lesion Localization

The composite lesion mask combined the MA and HE and EX and SE annotations. The optic disc was analysed separately because it is a normal structure. The retinal field of view was taken as the pixels with a summed colour intensity above 30. Localization was measured with Intersection over Union and the Dice coefficient.

$$IoU = | H \cap G | / | H \cup G | \quad (3)$$

$$Dice = 2 | H \cap G | / (| H | + | G |) \quad (4)$$

Here H is the thresholded Grad-CAM region and G is the lesion mask. In the pointing game a hit was counted when the point of maximum activation fell inside the lesion mask. A random control was made for every image. The Grad-CAM map was rotated by a random angle between 0 and 360 degrees and shifted by up to

120 pixels in each direction. This was repeated 100 times with random seed 42. A control was kept only when at least 90 percent of its activation mass lay inside the retinal field of view.

To judge how much of the low overlap could come from the coarse map, an upper bound was computed. For each image the lesion mask was assigned to a regular grid of 16 by 16, 32 by 32 or 64 by 64 cells. Every cell that contained a lesion pixel was marked, and the IoU between this ideal blocky map and the true mask was calculated. The values were averaged over the 81 images. This bound is an oracle limit and not an attainable score.

9. Decision Faithfulness

Faithfulness was assessed with a deletion test. Pixels were ranked by Grad-CAM activation and the highest ranked pixels were removed. The deleted share of the retinal area was 5% and 10% and 20% and 30% and 50%. After each step the model confidence for its originally predicted class was recorded and divided by the original confidence. In the primary procedure the removed pixels were set to zero in the normalized input. This value equals the ImageNet mean colour and is not raw black. Random control regions had the same area as the Grad-CAM regions and were produced by the rotation and shift procedure described above.

The relative confidence over the deletion levels gives a curve. The area under this curve (AUC) was computed by the trapezoidal rule over the deletion levels 0%, 5%, 10%, 20%, 30% and 50% and divided by the largest deletion fraction (50%), so that AUC lies between 0 and 1 for a curve that starts at 1. A relative confidence of 1 at every level would therefore give an AUC of 1. A lower AUC means that confidence fell faster. The faithfulness gap is the random AUC minus the Grad-CAM AUC.

. Gapfaith = Aucrandom − Aucgrad-Cam (5)

A positive gap means that removing the Grad-CAM regions lowered confidence more than removing random regions. Two robustness checks were added. In the first the removed pixels were replaced by a Gaussian blurred copy of the image with a kernel of 51 pixels and a sigma of 15. In the second the activation near the edge of the retinal field was suppressed before the regions were ranked.

12. Statistical Analysis

Paired comparisons of localization and faithfulness values used the Wilcoxon signed rank test. Confidence intervals for mean differences came from bootstrap resampling with 2000 resamples. The pointing game is reported as a hit rate. Because each image gives a hit or a miss, the hit rate of each model was compared with the mean random-control hit rate of that model by a one-sided exact binomial test. The Wilcoxon test was not used for this measure. The difference in hit rate between the two models was tested with the paired Wilcoxon signed-rank test. The Spearman rank correlation between IoU and the faithfulness gap was computed separately for each model. A result that was not significant was not taken as proof that no association exists. No correction for multiple testing was applied.

IV. RESULTS

1. Classification Performance

On its APTOS validation set Model A reached an accuracy of 82.26% and a QWK of 0.9076 with a macro F1 score of 0.6854. On the IDRiD test set of 103 images its accuracy fell to 34.95% and its QWK to 0.563. Model B reached an accuracy of 61.17% (63 of 103 images) and a QWK of 0.746 on the same images. The gain in accuracy was 26.2 percentage points with a 95% confidence interval of 16.5 to 36.9. The gain in QWK was 0.183 with an interval of 0.067 to 0.292. Table I gives the details.

TABLE I. CLASSIFICATION PERFORMANCE ON THE Idrid TEST SET (103 IMAGES)

Metric	Model A	Model B
Accuracy (95% CI)	0.350 (0.252 to 0.447)	0.612 (0.515 to 0.699)
QWK (95% CI)	0.563 (0.433 to 0.681)	0.746 (0.641 to 0.830)
Within one grade accuracy (95% CI)	0.806 (0.728 to 0.883)	0.835 (0.767 to 0.913)

Model B improved recall for Grades 0 and 2 and 3 and 4 while recall for Grade 1 fell. Grade 4 recall rose from 8% to 46%. Grade 1 recall fell from 40% to 20% but this grade has only five test images so one prediction changes the value by 20 points.

TABLE II. CLASS WISE RECALL ON THE IDRiD TEST SET

DR grade	Test images	Model A	Model B
Grade 0	34	32%	76%
Grade 1	5	40%	20%
Grade 2	32	44%	50%
Grade 3	19	42%	74%
Grade 4	13	8%	46%

2. Lesion Localization

Grad-CAM beat the random control for IoU and Dice at every threshold for both models with Wilcoxon p below 0.001. The absolute overlap was still small. At the threshold of 0.5 the mean IoU was 0.0432 for Model A and 0.0451 for Model B. The differences between the models were not significant for IoU ($p = 0.354$) or Dice ($p = 0.318$). Table III lists all thresholds.

TABLE III. Grad-Cam Localization At Three Thresholds (81 Images)

Model	Threshold	Mean IoU	Mean Dice	Random IoU
Model A	0.3	0.0603	0.1090	0.0230
Model A	0.5	0.0432	0.0794	0.0118
Model A	0.7	0.0240	0.0454	0.0050
Model B	0.3	0.0625	0.1136	0.0210
Model B	0.5	0.0451	0.0833	0.0123
Model B	0.7	0.0213	0.0401	0.0053

In the pointing game Model A scored 13 hits in 81 images (16.0%) and Model B scored 8 hits (9.9%). The random control gave about 2.6% for Model A and 3.2% for Model B. The difference between the two models was not significant ($p = 0.132$). Both models exceeded their random controls (exact binomial test: Model A, 13 of 81 against 2.6%, $p < 0.001$; Model B, 8 of 81 against 3.2%, $p = 0.004$).

The lesion types differed strongly. Table IV shows that the overlap for microaneurysms was close to zero for both models. The mean lesion area was 3.17% of the image. The mean Grad-CAM area at threshold 0.5 was 3.16% for Model A and 2.68% for Model B.

TABLE IV. MEAN IoU BY LESION TYPE AT THRESHOLD 0.5

Model	MA	HE	EX	SE
Model A	0.0022	0.0334	0.0059	0.0393
Model B	0.0020	0.0311	0.0121	0.0412

A later evaluation of Model B restricted all masks and maps to the retinal field of view. It gave IoU and Dice of 0.0641 and 0.1161 at threshold 0.3 and 0.0457 and 0.0844 at 0.5 and 0.0215 and 0.0403 at 0.7. This correction was not repeated for Model A. These values are therefore not compared with the Model A values in Table III.

3. Decision Faithfulness

Deletion of the Grad-CAM regions lowered the confidence much faster than deletion of random regions. A random AUC could not be computed for one Model A image (IDRiD_12) so the paired comparison uses 80 images. For Model A the Grad-CAM AUC was 0.3507 and the random AUC was 0.7574 which gives a gap of 0.4067. For Model B on the same 80 images the values were 0.3682 and 0.8147 and the gap was 0.4465. On all 81 images Model B gave 0.3666 and 0.8149 and 0.4483. The gap was above zero for both models (Wilcoxon p below 0.001). Table V and Fig. 2 show the results.

TABLE V. Deletion Based Faithfulness (80 Paired Images, Zero Fill)

Model	Grad-CAM AUC	Random AUC	Faithfulness gap
Model A	0.3507	0.7574	0.4067
Model B	0.3682	0.8147	0.4465

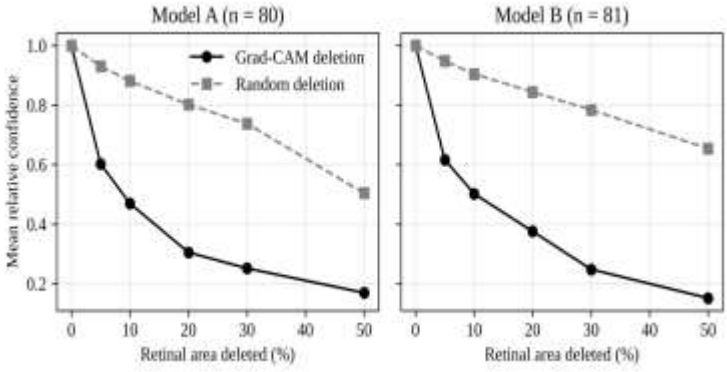

Fig. 2. Mean relative confidence during deletion of Grad-CAM regions and random regions.

The models did not differ significantly in Grad-CAM AUC ($p = 0.331$) or in the gap ($p = 0.337$). The effect was not the same for every image. The gap was positive for 79 of 80 images of Model A and 76 of 81 images of Model B. The median gap was 0.348 for Model A and 0.482 for Model B. A few images also showed a

confidence above the original value after a small deletion The gap stayed positive when the zero fill was replaced by blur and when the edge band was suppressed. Table VI gives the values.

TABLE VI. Robustness Of The Deletion Result

Setting	Model	Grad-CAM AUC	Random AUC	Gap
Blur fill	Model A	0.4465	0.8229	0.3763
Blur fill	Model B	0.3896	0.7933	0.4036
Edge band removed	Model B	0.4179	0.8112	0.3932

4. Relation Between Localization And Faithfulness

Table VII gives the Spearman correlation between IoU and the faithfulness gap. For Model A the correlation was weak and positive at all three thresholds and it was significant. For Model B it was also positive but it was not significant at any threshold. Fig. 3 shows the values at threshold 0.5.

TABLE VII. Spearman Correlation Between Iou And Faithfulness Gap

Model	Threshold	n	Spearman rho	p value
Model A	0.3	80	0.285	0.010
Model A	0.5	80	0.280	0.012
Model A	0.7	80	0.256	0.022
Model B	0.3	81	0.150	0.182
Model B	0.5	81	0.205	0.066
Model B	0.7	81	0.109	0.331

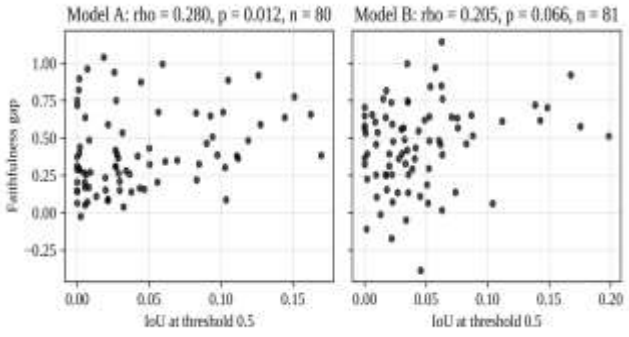

Fig. 3. IoU at threshold 0.5 against the faithfulness gap for each image.

5. Analysis By Dr Grade

Grade wise results are available for the 60 images with a known reference grade. The groups are small so this analysis is exploratory and no test across grades was run. The gap was positive in every group for both models. The pointing game hit rate rose with the reference grade in both models. The mean IoU stayed between 0.033 and 0.065 in all groups. For both models the gap for Grade 3 was larger than the gap for Grade 2.

TABLE VIII. Localization And Faithfulness By Reference Grade (Threshold 0.5)

Model	Grade	n	Mean IoU	Mean Dice	Pointing hit	Gap
Model A	2	37	0.0341	0.0634	10.8%	0.397
Model A	3	14	0.0535	0.0982	21.4%	0.437
Model A	4	9	0.0550	0.1004	33.3%	0.327
Model B	2	37	0.0383	0.0712	2.7%	0.409
Model B	3	14	0.0653	0.1193	7.1%	0.554
Model B	4	9	0.0334	0.0639	22.2%	0.507

6. Additional Explanation Analyses

Two other methods were tried on Model B. Grad-CAM++ used layer4 and LayerCAM used the finer layer3. The pointing game hit rate was highest for LayerCAM but its deletion AUC was the weakest of the three. The random AUC values in Table IX come from a separate run and differ slightly from those in Table V.

TABLE IX. Comparison Of Explanation Methods For Model B (81 Images)

Method	IoU at 0.5	Pointing hit	Deletion AUC	Random AUC
Grad-CAM (layer4)	0.0451	9.9% (8/81)	0.366	0.824
Grad-CAM++ (layer4)	0.0442	12.3% (10/81)	0.418	0.805

Method	IoU at 0.5	Pointing hit	Deletion AUC	Random AUC
LayerCAM (layer3)	0.0476	32.1% (26/81)	0.512	0.713

In Model B about 1.8% of the Grad-CAM energy fell inside the optic disc which covers about 2.6% of the image area. The peak of the map was never inside the optic disc in the 81 images. The activation was also not spread evenly over the retina. The central region held 25.9% of the energy and covers 25.0% of the area. The middle region held 54.4% for 47.2% of the area and the outer region held 19.7% for 27.8% of the area. These analyses are exploratory. The average heatmaps of both models show bands of high activation near the upper and lower borders of the retinal field, which are not present in the average lesion mask (Fig. 4).

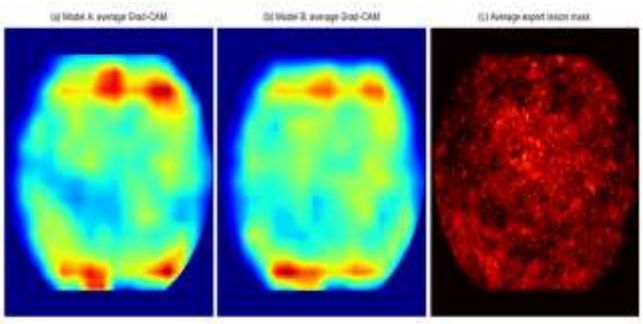

Fig. 4. Average Grad-CAM heatmaps of (a) Model A and (b) Model B, and (c) the average expert lesion mask over the 81 images.

7. Discussion

1. Principal Findings

Classification performance and explanation quality behaved as separate properties. Training on IDRiD raised the accuracy from 34.95% to 61.17%. This did not give a significant change in IoU or Dice or the deletion AUC or the faithfulness gap. A better classifier should therefore not be assumed to give better explanations. The strongest effect in the study was the faithfulness gap. Removing Grad-CAM regions cost far more confidence than removing random regions of the same size. The overlap with lesion annotations was low at the same time.

8. Cross Dataset Generalization

Model A reached 82.26% accuracy on its own APTOS validation set but only 34.95% on IDRiD. The size of this drop shows how hard it is to move a model from one retinal dataset to another. Bragança et al. studied cross dataset performance in DR screening and reported that external distribution shift remains a challenge even when results within one dataset are strong [30]. Adapting the model to IDRiD helped in our setting. The result does not show clinical readiness or generalization to other datasets. The fall in Grade 1 recall rests on five images and should not be read as a stable estimate.

9. Limited Overlap With Expert Annotations

Both models gave low IoU and Dice values. Even at the lowest threshold the best mean IoU was 0.0603 for Model A and 0.0625 for Model B. The random control scored lower every time so the maps do carry spatial information about the lesions. The maps should still not be read as lesion segmentations. One reason is the resolution of the target layer. The map has 16 by 16 cells and is only upsampled to 512 by 512 pixels. Upsampling does not restore detail that the feature map never held. A simple bound supports this view. A perfect map on a 16 by 16 grid would reach a mean IoU of only 0.198 on these 81 images. Grids of 32 by 32 and 64 by 64 would allow 0.302 and 0.425. The microaneurysm IoU of about 0.002 shows the problem for the smallest lesions. Shinde et al. proposed high resolution class activation maps for localizing pathology with CNNs [31]. A finer map is a reasonable direction for further work and we do not compare numbers with that study. Grad-CAM is also not trained to reproduce pixel level masks. It shows regions that are linked to a class prediction.

10. Decision Faithfulness And The Role Of Controls

Deleting Grad-CAM regions reduced confidence faster than deleting random regions for both models. The result held with blur replacement and with the edge band suppressed. This supports the view that the highlighted regions mattered to the predictions under these deletion procedures. The result still depends on the perturbation. Removing or replacing parts of an image changes its appearance and a fall in confidence does not show that the model relies on clinically meaningful lesions. The random control gives the reference that makes the result interpretable. The positive gap is evidence of relative sensitivity to the

highlighted regions and is not a proof of clinical validity. The random control is a rotated and shifted copy of the same map, so it matches the area and shape of the Grad-CAM region but not its position. A different random control, for example scattered patches, could give a different size of gap, so the gaps should be read relative to this control.

11. Localization Does Not Imply Faithfulness

The relation between IoU and the faithfulness gap was weak. Model A showed a weak positive and significant correlation. Model B showed a positive correlation that did not reach significance. A region can overlap a lesion without being needed for the decision. A region can also drive the decision without matching an annotation. Akter et al. argued that localization gains obtained through explanation regularization should not be read as faithfulness gains [32]. Our results approach the same distinction from another direction. The heatmaps had very low overlap with the lesion masks and still caused a fast loss of confidence under deletion. A localization score alone therefore cannot stand in for a faithfulness test. These findings do not show that the two properties are unrelated. A larger cohort is needed to learn whether the weak association can be reproduced.

12. Comparison Of Explanation Methods

LayerCAM gave the highest pointing game hit rate at 32.1% but its deletion AUC was higher than that of Grad-CAM. Grad-CAM++ had a lower hit rate than LayerCAM. A method that does well on one property can do worse on the other. LayerCAM should not be called better than Grad-CAM on the basis of the pointing game alone. A method should be judged on the property it is meant to show.

13. Implications

A convincing heatmap or a high localization score is not enough to show that a model decides from clinically meaningful regions. For DR grading the evaluation should report classification performance and agreement with lesion annotations and faithfulness under controlled perturbation as three separate results. Better grading did not give better explanations here so explanation quality should be tested directly.

14. Limitations

The localization analysis used 81 images and grades were known for only 60 of them. The grade wise results are therefore exploratory. The Grad-CAM map came from a 16 by 16 feature map which limits the localization of small lesions. Only one architecture and one main target layer were used so the findings may not carry over to other models. The faithfulness results depend on the deletion procedure. Random controls and blur replacement were added but other perturbations could give other results.

The deletion results were not uniform. A few images showed no advantage for Grad-CAM deletion over the random control and some images showed a higher confidence after a small deletion. Mean values should be read together with the median and the spread over images. The Model A and Model B faithfulness comparison used 80 paired images because one random control could not be computed. The field of view correction was done for Model B only so its values should not be used to claim a gain over the Model A values in Table III. No Bonferroni or similar correction was applied to the correlation tests. Finally this study compared computed explanations with expert masks and deletion results. It did not test whether ophthalmologists decide better when they see the heatmaps.

The heatmaps also showed bright bands near the top and bottom borders of the retinal field when they were averaged over the 81 images (Fig. 4). The share of Grad-CAM energy in the innermost 10% of the retinal radius was positively related to the faithfulness gap (Spearman rho = 0.57 for Model A, n = 80; rho = 0.41 for Model B, n = 81; both $p < 0.001$) but not to IoU (rho = 0.04 and 0.06). Edge band removal reduced the Model B faithfulness gap from 0.458 to 0.393 (about 14%) and left it far above zero. Edge activity therefore explains part of the deletion effect but not all of it. The cause of the border pattern, for example a padding or border effect of the network, was not tested.

V. CONCLUSION

This study asked whether better lesion localization shows stronger decision faithfulness in Grad-CAM explanations for DR grading. Two ResNet-50 models were evaluated with classification metrics and expert

lesion masks and deletion tests. Training on IDRiD raised accuracy from 34.95% to 61.17% but did not change the localization and faithfulness metrics significantly. Grad-CAM beat random controls on both properties and its overlap with lesion annotations stayed low. The correlation between localization and faithfulness was weak and it was not consistent across the two models. Classification performance and localization and decision faithfulness should therefore be evaluated as separate properties of an explainable DR grading system.

1. Future Work

Future work should test higher resolution explanation methods for small lesions such as microaneurysms. The same protocol should be run on larger and more varied datasets and on other architectures. Other perturbation strategies and more explanation methods and multi layer visualization should be examined. Involving ophthalmologists in the evaluation would show whether the explanations help clinical interpretation.

2. Data Availability

The study used the public APTOS 2019 and IDRiD datasets. No new patient data were collected.

REFERENCES

1. P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Mériaudeau, "Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research," Data, vol. 3, no. 3, Art. no. 25, 2018, doi: 10.3390/data3030025.
2. P. Porwal et al., "IDRiD: Diabetic Retinopathy: Segmentation and Grading Challenge," Medical Image Analysis, vol. 59, Art. no. 101561, 2020, doi: 10.1016/j.media.2019.101561.
3. APTOS 2019 Blindness Detection, Kaggle, 2019. [Online]. Available: https://www.kaggle.com/competitions/aptos2019-blindness-detection
4. J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
6. V. Lakshminarayanan, H. Kheradfallah, A. Sarkar, and J. Jothi Balaji, "Automated Detection and Diagnosis of Diabetic Retinopathy: A Comprehensive Survey," Journal of Imaging, vol. 7, no. 9, Art. no. 165, 2021, doi: 10.3390/jimaging7090165.
7. C. Vanaja, R. Harikanth, A. Sabarinath, and S. Naveenkumar, "A Survey on Artificial Intelligence and Deep Learning Techniques for Diabetic Retinopathy Detection and Classification," in Proc. Int. Conf. Artificial Intelligence and Secure Data Analytics (ICAISDA 2025), Atlantis Press, 2026, pp. 212–222, doi: 10.2991/978-94-6239-616-6_18.
8. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
9. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. 34th Int. Conf. Machine Learning, vol. 70, 2017, pp. 3319–3328.
10. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the Visualization of What a Deep Neural Network Has Learned," IEEE Trans. Neural Networks and Learning Systems, vol. 28, no. 11, pp. 2660–2673, 2017, doi: 10.1109/TNNLS.2016.2599820.
11. V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-Box Models," in Proc. British Machine Vision Conf. (BMVC), 2018, Art. no. 151.
12. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in Advances in Neural Information Processing Systems, vol. 31, 2018.
13. C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (In)fidelity and Sensitivity of Explanations," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.
14. T. Gomez, T. Fréour, and H. Mouchère, "Metrics for Saliency Map Evaluation of Deep Learning

Explanation Methods," arXiv:2201.13291, 2022, doi: 10.48550/arXiv.2201.13291.

15. V. Lamprou, A. Kallipolitis, and I. Maglogiannis, "On the Evaluation of Deep Learning Interpretability Methods for Medical Images Under the Scope of Faithfulness," Computer Methods and Programs in Biomedicine, vol. 253, Art. no. 108238, 2024, doi: 10.1016/j.cmpb.2024.108238.
16. Y. Brima and M. Atemkeng, "Saliency-Driven Explainable Deep Learning in Medical Imaging: Bridging Visual Explainability and Statistical Quantitative Analysis," BioData Mining, vol. 17, Art. no. 18, 2024, doi: 10.1186/s13040-024-00370-4.
17. M. A. Lago, G. Zamzmi, B. Eich, and J. G. Delfino, "Evaluating Explainability: A Framework for Systematic Assessment of Explainable AI Features in Medical Imaging," Bioengineering, vol. 13, no. 1, Art. no. 111, 2026, doi: 10.3390/bioengineering13010111.
18. C. Ozer, A. Guler, A. Turkvatan Cansever, and I. Oksuz, "Consistent Explainable Image Quality Assessment for Medical Imaging," Health Information Science and Systems, vol. 14, Art. no. 31, 2026, doi: 10.1007/s13755-025-00411-0.
19. H. Jiang, J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma, and W. Qian, "A Multi-Label Deep Learning Model with Interpretable Grad-CAM for Diabetic Retinopathy Classification," in Proc. 42nd Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC), 2020, pp. 1560–1563, doi: 10.1109/EMBC44109.2020.9175884.
20. T. Van Craenendonck, B. Elen, N. Gerrits, and P. De Boever, "Systematic Comparison of Heatmapping Techniques in Deep Learning in the Context of Diabetic Retinopathy Lesion Detection," Translational Vision Science & Technology, vol. 9, no. 2, Art. no. 64, 2020, doi: 10.1167/tvst.9.2.64.
21. Q. Meng, Y. Hashimoto, and S. Satoh, "How to Extract More Information With Less Burden: Fundus Image Classification and Retinal Disease Localization With Ophthalmologist Intervention," IEEE J. Biomedical and Health Informatics, vol. 24, no. 12, pp. 3351–3361, Dec. 2020, doi: 10.1109/JBHI.2020.3011805.
22. M. Shorfuzzaman, M. S. Hossain, and A. El Saddik, "An Explainable Deep Learning Ensemble Model for Robust Diagnosis of Diabetic Retinopathy Grading," ACM Trans. Multimedia Computing, Communications, and Applications, vol. 17, no. 3s, Art. no. 113, 2021, doi: 10.1145/3469841.
23. K. K. Ghadai, S. K. Nayak, and B. R. Senapati, "A Clinically Interpretable Deep Learning Pipeline for Diabetic Retinopathy Classification Using EfficientNet, Advanced Data Augmentation and GradCAM," BMC Ophthalmology, vol. 26, Art. no. 406, 2026, doi: 10.1186/s12886-026-04897-4.
24. A. G. Persada, I. Ardiyanto, M. B. Sasongko, and H. A. Nugroho, "A Review of CAM-Based Visual Explanation on Diabetic Retinopathy," IEEE Access, vol. 14, pp. 13737–13767, 2026, doi: 10.1109/ACCESS.2026.3656888.
25. L. Cascone, P. Campiglia, M. Nappi, F. Narducci, and B. Simone, "Benchmarking Explainable AI Methods for Vision Transformer-Based Diabetic Retinopathy Analysis," IEEE Open J. Computer Society, vol. 7, pp. 947–957, 2026, doi: 10.1109/OJCS.2026.3688710.
26. F. K. Malerbi, L. F. Nakayama, P. Prado, F. Yamanaka, G. B. Melo, C. V. Regatieri, and J. A. Stuchi, "Heatmap Analysis for Artificial Intelligence Explainability in Diabetic Retinopathy Detection: Illuminating the Rationale of Deep Learning Decisions," Annals of Translational Medicine, vol. 12, no. 5, Art. no. 89, 2024, doi: 10.21037/atm-24-73.
27. S. Biswas, M. A. A. Khan, M. H. Ali, J. Rohdin, S. Pramanik, M. I. A. Khan, S. K. Chakravarty, and B. K. Pramanik, "Interpreting Deep Neural Networks in Diabetic Retinopathy Grading: A Comparison with Human Decision Criteria," Life, vol. 15, no. 9, Art. no. 1473, 2025, doi: 10.3390/life15091473.
28. Z. Zhang, H. Zhao, L. Dong, L. Luo, and H. Wang, "A Study on the Interpretability of Diabetic Retinopathy Diagnostic Models," Bioengineering, vol. 12, no. 11, Art. no. 1231, 2025, doi: 10.3390/bioengineering12111231.
29. J. Civit-Masot et al., "An Explainable Ensemble for Diabetic Retinopathy Grading with a Novel Confidence Quality Factor and Configurable Heatmaps," Medical & Biological Engineering & Computing, vol. 64, pp. 1231–1247, 2026, doi: 10.1007/s11517-026-03514-2.
30. H. Bragança, Í. P. Caliari, W. L. Vital, A. Fontenele, S. Cavalcante, and G. Messias, "Toward Reliable Diabetic Retinopathy Screening," Sensors, vol. 26, no. 14, Art. no. 4515, 2026, doi: 10.3390/s26144515.

31. S. Shinde, P. Tupe-Waghmare, T. Chougule, J. Saini, and M. Ingalhalikar, "Predictive and Discriminative Localization of Pathology Using High Resolution Class Activation Maps with CNNs," PeerJ Computer Science, vol. 7, Art. no. e622, 2021, doi: 10.7717/peerj-cs.622.
32. A. Akter, M. Dalla Mura, D. Mateus, and D. Al Chanti, "Localization Gains from Explanation Regularization Should Not Be Interpreted as Faithfulness Gains," arXiv:2610.26292, 2026.