

# Opening the black box of deep learning: Validating the statistically significant association between class activation map (CAM) and clinical domain knowledge in fundus image-based glaucoma diagnosis

Han Yuan<sup>1</sup>

YUAN.HAN@U.DUKE.NUS.EDU

Lican Kang<sup>2</sup>

KANGLICAN@WHU.EDU.CN

Yong Li<sup>1</sup>

LIYONG@U.DUKE.NUS.EDU

Chenkui Miao<sup>3</sup>

MIAOCHENKUI@NJMU.EDU.CN

Yuan Luo<sup>2</sup>

YUANLUO@WHU.EDU.CN

Chang Zhu<sup>4</sup>

CHANGZHU@HUST.EDU.CN

<sup>1</sup> Duke-NUS Medical School, National University of Singapore

<sup>2</sup> School of Mathematics and Statistics, Wuhan University

<sup>3</sup> Jiangsu Provincial People's Hospital, Nanjing Medical University

<sup>4</sup> Tongji Medical College, Huazhong University of Science and Technology

## Abstract

While deep learning has exhibited remarkable predictive capabilities in various medical image tasks, its inherent black-box nature has hindered its widespread implementation in real-world healthcare settings. Our objective is to unveil the decision-making processes of deep learning models in the context of glaucoma classification by employing several Class Activation Map (CAM) techniques to generate model focus regions and comparing them with clinical domain knowledge of the anatomical area (optic cup, optic disk, and blood vessels). Four deep neural networks, including VGG-11, ResNet-18, DeiT-Tiny, and Swin Transformer-Tiny, were developed using binary diagnostic labels of glaucoma and five CAM methods (Grad-CAM, XGrad-CAM, Score-CAM, Eigen-CAM, and Layer-CAM) were employed to highlight the model focus area. We applied the paired-sample t-test to compare the percentage of anatomies in the model focus area to the proportion of anatomies in the entire image. After that, Pearson's and Spearman's correlation tests were implemented to examine the relationship between model predictive ability and the percentage of anatomical structures in the model focus area. On five public glaucoma datasets, all deep learning models consistently displayed statistically significantly higher percentages of anatomical structures in the focus area than the proportions of anatomical structures in the entire image. Also, we validated the positive relationship between the percentage of anatomical structures in the focus area and model predictive performance. Our study provides evidence of the convergence of decision logic between deep neural networks and human clinicians through rigorous statistical tests. We anticipate that it can help alleviate clinicians' concerns regarding the trustworthiness of deep learning in healthcare. For reproducibility, the code and dataset have been released at GitHub<sup>1</sup>.

**Keywords:** Glaucoma diagnosis, Optic cup and disk, Blood vessels, Explainable machine learning, Class activation map, Convolutional neural networks, Vision transformer

---

1. <https://github.com/Han-Yuan-Med/the-statistical-association-between-cam-and-domain-knowledge>

## 1. Introduction

In the last decade, deep learning has reshaped various ophthalmological disease diagnoses (Ting et al., 2019). Although deep neural networks feature high-fidelity accuracy on retrospective classification, localization, and segmentation tasks, clinicians still cannot fully trust their decisions on prospective medical practice due to their black-box characteristics (Van der Velden et al., 2022; Xie et al., 2022). To open the black box of deep learning and address the interpretability issue, various methods have been proposed to explain model inference logic on different healthcare data. In the realm of medical image analysis, a commonly used set of methods, known as the Class Activation Map (CAM) family, is developed to map the final diagnostic decision back onto the input images by highlighting the important pixels (regions) within the images. Such an assisted heatmap will be provided to clinicians to conduct further evaluation on whether the model focuses on the clinically relevant region of interest (ROI) (Zhou et al., 2016). If a deep learning model consistently performs well on a specialized task and puts its attention on the lesion area causing the disease, clinicians' concern about the model reliability would be alleviated (Yuan et al., 2023b).

Glaucoma is a neuro-degenerative disease caused by the increase of intra-ocular pressure and will progress into complete blindness without early medical intervention (Varma et al., 1992; Liang et al., 1997; Garway-Heath et al., 1998; Morgan et al., 2012). Conventional clinical diagnosis was based on the clinicians manual evaluation of optic cup, optic disk, blood vessels, intraocular pressure, and visual field (Varma et al., 1992; Liang et al., 1997; Garway-Heath et al., 1998; Morgan et al., 2012; Xue et al., 2022; Morano et al., 2021). Recently, multiple deep learning-based systems were proposed to automate this process (Thompson et al., 2020; Mirzania et al., 2021). For example, Shinde introduced a comprehensive four-stage pipeline for glaucoma classification, including the delineation of the optic disk area, the segmentation of the optic cup and disk, the extraction of relevant clinical features, and finally, the classification of glaucoma based on these previously extracted features (Shinde, 2021). However, prior research efforts in this field have primarily focused on enhancing the diagnostic accuracy of glaucoma (Mitra et al., 2018; Zhao et al., 2023; Liu et al., 2022; Zhao et al., 2019). While some studies have offered qualitative visualizations illustrating that the decision logic of deep neural networks aligns with clinical domain knowledge concerning the optic cup, optic disk, and blood vessels (Akter et al., 2022; Thakoor et al., 2019; Li et al., 2019; Liao et al., 2019), there remains a need for quantitative and systematic evaluations on model decision logic to further bolster the confidence of ophthalmologists in the deployment of deep neural networks.

To address the quantification issue above, we implemented a paired-sample t-test to measure the association between model focus region and domain knowledge-based anatomical area in the context of glaucoma diagnosis. Compared with the clinically irrelevant area in the input image, the model statistically significantly concentrated more on the anatomical area (optic cup, optic disk, blood vessels) where clinicians make their diagnosis, showing the convergence of data-driven solutions and human knowledge-based strategies (Varma et al., 1992; Liang et al., 1997; Garway-Heath et al., 1998). We further demonstrated that the model diagnostic performance was positively correlated with its attention ratio on the important lesion area, indicating that future model developers should take the

clinical knowledge into account rather than treat medical image tasks as general computer vision missions using a fully end-to-end approach. Compared with the previous work on comparing model explanation and domain knowledge (Liao et al., 2019), we performed rigorous statistical tests on both convolutional neural networks (CNN) and transformers. In summary, we presented that an image-level diagnostic annotations-based model correctly summarized reasonable medical knowledge towards clinicians, underscoring the potentiality of deep learning in distilling latent knowledge in the future machine intelligence-based machine.

## 2. Materials and Methods

### 2.1. Dataset

We conducted a comprehensive investigation into the association between deep learning inference logic and clinical domain knowledge by utilizing five public datasets: ORIGA (Zhang et al., 2010), HRF (Budai et al., 2013), LES-AV (Orlando et al., 2018), Drishti-GS (Sivaswamy et al., 2014), and FIVES (Jin et al., 2022). These datasets encompassed a total of 650, 30, 22, 101, and 300 fundus images, respectively. Table 1 provides an overview of the used datasets. All fundus images from the five datasets were resized to the resolution of  $224 \times 224$  pixels to meet the requirements of most pre-trained deep learning backbones. In addition to binary diagnostic labels indicating the presence or absence of glaucoma, ORIGA and Drishti-GS were enriched with pixel-level annotations for anatomical structures of the optic cup and disk while HRF, LES-AV, and FIVES contained delineations of blood vessels. ORIGA, HRF, and LES-AV were used to develop glaucoma classifiers, conduct model explanations, and validate the association between model focus area and domain knowledge-based anatomical structures. Drishti-GS and FIVES were introduced as external evaluation of model classification and CAM explanation. To maximize the utility of the limited samples with pixel-level annotations of blood vessels, we implemented double cross-validation in model development and internal evaluation. This involved initially dividing the dataset into three non-overlapped subsets, followed by the allocation of training, validation, and test datasets to each of these subsets. Ultimately, this process generated  $A_3^3$ : 6 distinct scenarios to comprehensively evaluate various glaucoma classifiers and the model focus areas generated by different CAM methods (Burzykowski et al., 2023). Table 2 shows the concrete details of the double cross-validation. For the external evaluation on Drishti-GS and FIVES, we randomly divided the dataset into non-overlapping validation and test sets with a 50:50 ratio.

### 2.2. Deep learning-based glaucoma classification

We first give necessary notations to facilitate the downstream elaboration on classifier training and explanation of model decision logic in terms of focus area. For the glaucoma classification task, we denote the training, validation, and test datasets as  $D^{train}$ ,  $D^{val}$ , and  $D^{test}$ , respectively. Take the training dataset  $D^{train}$  as an example, each fundus image  $I_i^{train}$  is paired with image-level diagnostic label  $Y_i^{train}$  and pixel-level annotation  $A_i^{train}$  of either optic cup and disk or blood vessels.  $I_i^{train}$  designates a two-dimensional image with a width of  $W_0$  and a height of  $H_0$  and  $p_{w,h}(I_i^{train})$  denotes a pixel in  $I_i^{train}$  whose coordinate of

width and height is  $(w, h)$ . For the generation of labels, clinicians screened each  $I_i^{train}$  and assigned the binary diagnostic label of  $Y_i^{train}$  in which 1 stands for glaucoma. Additionally, clinicians delineated the anatomical structure,  $A_i^{train}$ , marking the pixels belonging to this specific anatomical structure with 1. The glaucoma classifier is trained based on  $I_i^{train}$  and  $Y_i^{train}$ , validated using  $I_i^{val}$  and  $Y_i^{val}$ , and tested by  $I_i^{test}$  and  $Y_i^{test}$ . After the model development,  $A_i^{test}$  is adopted to quantify the association between clinical domain knowledge of anatomical structures and the model decision logic in terms of the focus area.

Specifically, the training process of deep neural networks is to find a set of parameters for a pre-defined architecture that minimizes the difference between model predictions and ground truth labels in the training set. Formally, with the training dataset  $D^{train}$ , we aim to optimize a deep learning model  $f_\theta$  parameterized by  $\theta$ . Taking the input of  $I_i^{train}$ ,  $f_\theta$  outputs  $f_\theta(I_i^{train})$  and the optimization target is to minimize the loss function  $l$  between  $f_\theta(I_i^{train})$  and sample labels  $Y_i^{train}$  across all samples in  $D^{train}$ . To facilitate fine-grained training of  $f_\theta$ , the validation dataset  $D^{val}$  is applied to schedule the learning rate decay: If  $l(f_\theta(I_i^{val}), Y_i^{val})$  has not decreased for a pre-defined epoch number  $N_{epoch}$ , the learning rate of  $\theta$  will be decreased and the  $\theta$  showing the best performance on  $l(\theta; D^{val})$  will be saved as the optimal parameter  $\theta^*$ . After the model training, we calculate the binarization threshold  $\tau$  based on the model's optimal performance on  $D^{val}$  and then evaluate the classification performance of the trained  $f_{\theta^*}$  on the unseen test dataset  $D^{test}$ . Various metrics are applied to quantify the model performance by comparing the model prediction  $f_{\theta^*}(I_i^{test})$  and the ground truth label  $Y_i^{test}$ . In this study, we utilize seven common metrics for glaucoma classification evaluation, including the area under the receiver operating characteristic curve (AUROC), the area under the precision recall curve (AUPRC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Also, the standard error (SE) is reported based on the bootstrapping of samples in the test dataset (Efron, 1987). Besides the internal evaluation above, we introduce the external unseen validation set  $D_{ext}^{val}$  and test set  $D_{ext}^{test}$  for generalizability evaluation. Without fine-tuning, the developed model  $f_{\theta^*}$  is deployed on  $D_{ext}^{val}$  and  $D_{ext}^{test}$  for binarization threshold determination and model performance evaluation.

Considering the middle-scale characteristic of the used dataset, we developed glaucoma classifiers using four lightweight deep neural networks. These classifiers encompass two CNN backbones: VGG (VGG-11) (Simonyan and Zisserman, 2015) and ResNet (ResNet-18) (He et al., 2016), as well as two Transformer architectures: Vision Transformer-Tiny (DeiT-Tiny) (Dosovitskiy et al., 2021; Touvron et al., 2021) and Swin Transformer-Tiny (Swin-Tiny) (Liu et al., 2021). For model training, we employed Stochastic Gradient Descent (SGD) (Rumelhart et al., 1986) with a learning rate of 0.001, a momentum of 0.9, and a decay of 0.9 with a patience parameter of 10. Each deep learning model underwent training for 100 epochs across the six-fold cross-validation scenarios mentioned above. We reported the model performance with SE on both internal and external test datasets.

### 2.3. CAM explanation for deep learning classifiers

The developed model  $f_{\theta^*}$  classifies an unseen image  $I_i^{test}$  from the test dataset  $D^{test}$  as  $f_{\theta^*}(I_i^{test})$ . To explain the model decision logic behind  $f_{\theta^*}(I_i^{test})$ , the model focus area  $R(I_i^{test})$  in  $I_i^{test}$  is extracted by various CAM methods to highlight the most important

pixels (regions) towards the model final decision  $f_{\theta^*}(I_i^{test})$ . CAM techniques calculate each pixel's importance towards  $f_{\theta^*}(I_i^{test})$  and  $E_{(w,h)}(I_i^{test})$  denoted the importance of a pixel with coordinate  $(w, h)$  in  $I_i^{test}$ . We further outline the focus area  $R(I_i^{test})$  consisting of the most significant pixels by selecting the pixels with top 5%  $E_{(w,h)}(I_i^{test})$ .

In this study, we utilize five mainstream explanation techniques in the CAM family: Grad-CAM (Selvaraju et al., 2017), XGrad-CAM (Fu et al., 2020), Score-CAM (Wang et al., 2020), Eigen-CAM (Muhammad and Yeasin, 2020), and Layer-CAM (Jiang et al., 2021) to generate the pixel-level importance  $E_{(w,h)}(I_i^{test})$ . The top 5% important pixels are extracted to formulate the model focus area  $R(I_i^{test})$ . After that, we compare the percentages of anatomical structures in the model focus area  $(R(I_i^{test}) \cap A_i^{test})/R(I_i^{test})$  and the whole image  $A_i^{test}/I_i^{test}$ . Similar to the external classification evaluation, we also implement external evaluation on model explanation using  $D_{ext}^{test}$ . Apparently, if deep learning models allocate no additional attention to anatomical structures compared with other regions, there would be no significant difference between the percentage of anatomies in the model focus area and the entire input image. However, if the model identifies glaucoma predominantly by relying on evidence from anatomical structures, the first percentage would be significantly higher than the second.

To establish a rigorous analysis of the relationship between model decision logic (focus area) and medical domain knowledge (the optic cup, optic disk, and blood vessels), the paired-sample t-test (Student, 1908) is employed to compare the percentages of anatomical structures in the model focus area and the whole image. Besides, we compute both Pearson's (Pearson, 1920) and Spearman's correlation coefficient (Fieller et al., 1957) to ascertain whether there exists a correlation between the model's classification performance of AUROC and the proportion of anatomical areas within its focus region. A correlation coefficient larger than 0 implies a positive correlation between the two variables. For reproduction, code and dataset are publicly available at GitHub<sup>1</sup>.

### 3. Results

First, we quantitatively demonstrated the internal classification performance of various models on glaucoma diagnosis in Table 3 and their external performance on Grishti-GS and FIVES in Table 4 and Table 5, respectively. In both internal and external evaluations, VGG-11 consistently outperformed other models, achieving the highest classification performance across all metrics, with the exception of sensitivity in the external Drishti-GS dataset. ResNet-18, DeiT-Tiny, and Swin-Tiny also delivered a commendable performance in internal and external glaucoma classification, demonstrating the feasibility of downstream analysis of the models' decision logic.

Table 6, 7, 8, and 9 presents the CAM explanation performance of different deep learning models on the internal test dataset and their external explanation performance on Drishti-GS and FIVES was provided in Table 10, 11, 12, and 13, and Table 14, 15, 16, and 17, respectively. From the  $P$  values of paired-sample t-tests, we observed that anatomical structures of the optic cup, optic disk, and blood vessels occupied statistically significant percentages in the focus regions of various glaucoma classifiers. Further, Table 18 summarizes Pearson's and Spearman's correlation tests between the model classification performance of AUROC and the anatomies' proportion in the focus area. Notably, there exists

a statistically significantly positive relationship between model predictive performance and their focus on anatomical structures, which aligns with the clinician’s diagnostic logic of glaucoma (Liao et al., 2019).

To provide a comparative visualization of anatomies and model decision logic, Figure 1 portrays a comparative visualization of anatomies and VGG-11 decision logic, and Figure 2 and 3 shows the corresponding results in the external test sets. The images from the first to the fifth column are original images, ground-truth anatomies, and focus area by Grad-CAM, XGrad-CAM, Score-CAM, Eigen-CAM, and Layer-CAM, respectively. The striking resemblance between the model’s focus areas and the anatomical regions underscores a compelling point: despite being trained with solely image-level labels, deep neural networks rely on reasonable evidence to classify glaucoma, akin to the decision-making process of human clinicians (Liao et al., 2019). This observation also demonstrated the practicability of knowledge distillation by data-driven deep neural networks (Qin et al., 2021).

#### 4. Discussion

In this study, we developed four deep neural networks based on both CNN and Vision Transformer architectures for glaucoma classification. We compared model focus regions generated by different CAM methods and anatomical structures annotated by clinicians to shed light on the extent of alignment between the decision-making logic of black-box deep learning models and the clinical domain knowledge that human experts exploit to make diagnoses. Based on the paired-sample t-test, we showed that the data-driven deep neural networks consistently paid more attention to the anatomical structures of the optic cup, optic disk, and blood vessels. This empirical evidence underscores the convergence of decision logic between our models and human experts in the context of glaucoma classification. Further, we implemented Pearson’s and Spearman’s correlation analysis and revealed the positive relationship between the model’s attentiveness to anatomical structures in the focus area and the model’s predictive performance.

In our experiments, VGG-11 outperformed both ResNet-18 and Vision Transformer-based DeiT-Tiny and Swin-Tiny with more sophisticated structure design. Such a phenomenon of “*The deeper is not the better*” has been reported in eye cancer classification (Santos-Bustos et al., 2022) and pneumonia detection (Ikechukwu et al., 2021) and is verified by our experimental results on glaucoma classification. We demonstrated that the conventional CNN model of VGG-11 is still capable even achieving the best performance in handling datasets with middle scale and resolution. In terms of AUROC and AUPRC, VGG-11 was 17.9% and 21.0% higher than ResNet-18, 9.2% and 15.3% higher than DeiT-Tiny, and 9.2% and 16.6% higher than Swin-Tiny in the internal test set and a similar substantial superiority can be observed in the external tests sets. Additionally, VGG-11 holds relatively lower complexity than Vision Transformers. The inherent preservation of spatial information in VGG-11 leads to its focus region by various CAM methods overlaps more with the anatomies than the Vision Transformers, making it better aligned with clinical knowledge and becoming more trustworthy from the perspective of healthcare professionals (Yuan et al., 2023a).

To uncover the model inference logic, we chose various CAM methods to pinpoint the focus regions toward model final decisions. This approach has been employed by prior

researchers to segment pixel-level lesion areas based on image-level labels (Yuan et al., 2023a; Zhang et al., 2020, 2021; Chan et al., 2021). However, most of these studies assessed whether the model focused on ROI by calculating overlap metrics such as Intersection over Union (IoU). In this study, we introduced another approach that conducts statistical tests on whether there is a significant difference between the percentage of ROI in the model focus area and the percentage of ROI in the entire image input. Therefore, if the first term is statistically significantly higher than the second term, it is ascertained that the model exhibits a preference for ROI over clinically irrelevant regions. In our case, we showed that the convergence of decision logic in various deep learning models and human clinicians (Yuan et al., 2021), which was further supported by the positive relationship between model predictive performance and the percentage of anatomies in the model focus area.

Based on this work, there are several limitations to be addressed in the future. First, the explanation methods utilized in this study were confined to the CAM family. Subsequent research should consider the incorporation of additional deep learning explanation methods such as Integrated Gradients (Sundararajan et al., 2017) to offer a more comprehensive analysis. Second, this work primarily focuses on explaining the existing fundus images. In future research, we intend to explore an alternative approach that uses image generation methods to manipulate anatomical structures (White et al., 2023; Wu et al., 2022). This approach will allow us to gain deeper insights into the behavior and the underlying decision logic of deep neural networks. Third, the statistical tests compared model focus area with anatomies and future work will consider additional clinical evidence such as bleeding and notch (Liao et al., 2019). Furthermore, our observations revealed a positive correlation between a model’s predictive performance and the percentage of its focus region over anatomical structures. This observation suggests a potential avenue for enhancing model performance, wherein the quantitative percentage of anatomies could be utilized as the rewards under the framework of reinforcement learning (Liu et al., 2022; Kang et al., 2023). Finally, this research exclusively explored glaucoma diagnosis and we plan to evaluate the association between model focus area and anatomical structure in a broader spectrum of ophthalmological tasks such as diabetic retinopathy, retinal vein occlusion, and fundus tumors (Cen et al., 2021).

## 5. Conclusion

The black-box nature of deep learning has long hindered its application in healthcare. In this study, we validated the statistically significant alignment of clinical domain knowledge and the decision-making logic of deep neural networks through several rigorous statistical tests. We hope that this research mitigates domain experts’ concerns regarding the trustworthiness of deep learning in glaucoma diagnosis.

Table 1: An overview of the used dataset. ORIGA, HRF, and LES-AV were used for model development and internal evaluation. Drishti-GS and FIVES were introduced for external evaluation.

Dataset	Function	Anatomical annotation	Case number	Control number
ORIGA		Optic cup & disk	168	482
HRF	Model development & Internal evaluation	Blood vessels	15	15
LES-AV		Blood vessels	11	11
Drishti-GS		Optic cup & disk	70	31
FIVES	External evaluation	Blood vessels	150	150

Table 2: Data split details for double cross-validation. The split index of (1, 2, 3) denotes a scenario where subset 1 is the training dataset, subset 2 is the validation dataset, and subset 3 is the test dataset. Therefore, we have 6 different scenarios for a comprehensive evaluation of deep learning classification and explanation.

Data split index	Anatomical annotation	Case number	Control number
1	Optic cup & disk	56	161
	Blood vessels	9	9
2	Optic cup & disk	56	161
	Blood vessels	9	9
3	Optic cup & disk	56	160
	Blood vessels	8	8

Table 3: Glaucoma classification performance of various deep learning models on the internal test dataset.

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
VGG-11	0.746 (0.033)	0.524 (0.062)	0.699 (0.029)	0.657 (0.058)	0.715 (0.034)	0.471 (0.050)	0.848 (0.026)
ResNet-18	0.633 (0.040)	0.414 (0.054)	0.622 (0.030)	0.526 (0.058)	0.659 (0.035)	0.374 (0.049)	0.786 (0.031)
DeiT-Tiny	0.683 (0.037)	0.444 (0.055)	0.606 (0.031)	0.630 (0.054)	0.597 (0.035)	0.376 (0.045)	0.816 (0.030)
Swin-Tiny	0.683 (0.039)	0.437 (0.057)	0.623 (0.030)	0.678 (0.055)	0.602 (0.036)	0.396 (0.045)	0.833 (0.031)

Table 4: Glaucoma classification performance of various deep learning models on the external test dataset of Drishti-GS.

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
VGG-11	0.789 (0.068)	0.890 (0.048)	0.686 (0.060)	0.629 (0.081)	0.812 (0.091)	0.881 (0.063)	0.501 (0.094)
ResNet-18	0.687 (0.081)	0.808 (0.073)	0.526 (0.066)	0.419 (0.076)	0.760 (0.095)	0.793 (0.099)	0.385 (0.083)
DeiT-Tiny	0.697 (0.075)	0.841 (0.061)	0.611 (0.067)	0.595 (0.076)	0.646 (0.105)	0.801 (0.079)	0.440 (0.107)
Swin-Tiny	0.784 (0.072)	0.875 (0.054)	0.673 (0.059)	0.657 (0.074)	0.708 (0.105)	0.839 (0.068)	0.497 (0.100)

Table 5: Glaucoma classification performance of various deep learning models on the external test dataset of FIVES.

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
VGG-11	0.733 (0.040)	0.771 (0.036)	0.691 (0.039)	0.596 (0.055)	0.787 (0.044)	0.743 (0.051)	0.664 (0.048)
ResNet-18	0.661 (0.044)	0.662 (0.052)	0.605 (0.038)	0.571 (0.052)	0.638 (0.051)	0.614 (0.054)	0.599 (0.049)
DeiT-Tiny	0.629 (0.045)	0.644 (0.051)	0.586 (0.039)	0.484 (0.051)	0.686 (0.046)	0.627 (0.058)	0.576 (0.048)
Swin-Tiny	0.570 (0.046)	0.611 (0.052)	0.566 (0.042)	0.482 (0.058)	0.649 (0.051)	0.583 (0.058)	0.556 (0.051)

Table 6: Glaucoma explanation performance of various XAI methods integrated with VGG-11 on the internal test dataset. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Grad-CAM	Optic cup	4.74 (0.18)	0.71 (0.00)	4.03 (0.18)	4.19E-77	
	Optic disk	10.36 (0.36)	1.80 (0.00)	8.56 (0.36)	9.12E-85	
	Blood vessels	22.95 (0.94)	18.96 (0.28)	3.98 (0.84)	1.66E-04	
XGrad-CAM	Optic cup	5.68 (0.20)	0.71 (0.00)	4.97 (0.18)	4.11E-100	
	Optic disk	12.33 (0.41)	1.80 (0.00)	10.53 (0.41)	5.63E-112	
	Blood vessels	23.15 (0.92)	18.96 (0.28)	4.18 (1.05)	1.67E-04	
VGG-11	Score-CAM	Optic cup	10.22 (0.18)	0.71 (0.00)	9.51 (0.18)	3.48E-259
	Optic disk	24.07 (0.38)	1.80 (0.00)	22.27 (0.38)	2.26E-304	
	Blood vessels	33.23 (0.94)	18.96 (0.28)	14.26 (0.92)	1.93E-25	
Eigen-CAM	Optic cup	11.92 (0.18)	0.71 (0.00)	11.21 (0.18)	0.00E+00	
	Optic disk	29.21 (0.33)	1.80 (0.00)	27.41 (0.33)	0.00E+00	
	Blood vessels	41.33 (0.92)	18.96 (0.28)	22.37 (0.84)	8.39E-48	
Layer-CAM	Optic cup	12.34 (0.18)	0.71 (0.00)	11.63 (0.18)	0.00E+00	
	Optic disk	29.20 (0.31)	1.80 (0.00)	27.40 (0.31)	0.00E+00	
	Blood vessels	33.66 (0.94)	18.96 (0.28)	14.70 (1.07)	1.61E-23	

Table 7: Glaucoma explanation performance of various XAI methods integrated with ResNet-18 on the internal test dataset. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Grad-CAM	Optic cup	3.28 (0.18)	0.71 (0.00)	2.57 (0.18)	2.50E-46	
	Optic disk	6.97 (0.36)	1.80 (0.00)	5.17 (0.33)	3.41E-48	
	Blood vessels	22.43 (1.68)	18.96 (0.28)	3.47 (1.73)	4.83E-02	
XGrad-CAM	Optic cup	1.07 (0.08)	0.71 (0.00)	0.36 (0.08)	1.48E-04	
	Optic disk	2.56 (0.13)	1.80 (0.00)	0.76 (0.13)	1.06E-04	
	Blood vessels	23.02 (1.25)	18.96 (0.28)	4.06 (1.20)	1.23E-03	
ResNet-18	Optic cup	2.89 (0.18)	0.71 (0.00)	2.17 (0.18)	4.39E-41	
	Optic disk	6.63 (0.38)	1.80 (0.00)	4.83 (0.38)	4.35E-45	
	Blood vessels	31.08 (1.12)	18.96 (0.28)	12.12 (1.10)	4.31E-20	
Eigen-CAM	Optic cup	4.68 (0.18)	0.71 (0.00)	3.96 (0.18)	7.22E-86	
	Optic disk	10.75 (0.36)	1.80 (0.00)	8.95 (0.36)	6.72E-104	
	Blood vessels	33.88 (0.84)	18.96 (0.28)	14.92 (0.87)	7.76E-32	
Layer-CAM	Optic cup	5.66 (0.20)	0.71 (0.00)	4.95 (0.18)	8.76E-128	
	Optic disk	12.89 (0.38)	1.80 (0.00)	11.10 (0.38)	5.64E-154	
	Blood vessels	34.41 (0.66)	18.96 (0.28)	15.45 (0.69)	1.26E-35	

Table 8: Glaucoma explanation performance of various XAI methods integrated with DeiT-Tiny on the internal test dataset. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Grad-CAM	Optic cup	3.52 (0.13)	0.71 (0.00)	2.81 (0.15)	4.90E-53	
	Optic disk	7.25 (0.28)	1.80 (0.00)	5.45 (0.28)	4.38E-52	
	Blood vessels	22.57 (1.28)	18.96 (0.28)	3.61 (1.20)	3.63E-03	
XGrad-CAM	Optic cup	3.51 (0.15)	0.71 (0.00)	2.79 (0.15)	5.84E-63	
	Optic disk	7.99 (0.31)	1.80 (0.00)	6.19 (0.31)	7.93E-69	
	Blood vessels	19.09 (1.15)	18.96 (0.28)	0.13 (1.12)	9.12E-01	
DeiT-Tiny	Optic cup	2.84 (0.18)	0.71 (0.00)	2.13 (0.18)	1.69E-34	
	Optic disk	5.76 (0.36)	1.80 (0.00)	3.96 (0.36)	2.07E-31	
	Blood vessels	19.71 (1.28)	18.96 (0.28)	0.75 (1.20)	5.42E-01	
Eigen-CAM	Optic cup	1.81 (0.13)	0.71 (0.00)	1.10 (0.13)	1.90E-16	
	Optic disk	4.31 (0.23)	1.80 (0.00)	2.51 (0.23)	8.91E-18	
	Blood vessels	24.67 (0.97)	18.96 (0.28)	5.70 (0.97)	1.76E-07	
Layer-CAM	Optic cup	7.43 (0.20)	0.71 (0.00)	6.72 (0.20)	7.85E-174	
	Optic disk	16.18 (0.38)	1.80 (0.00)	14.38 (0.38)	4.41E-195	
	Blood vessels	25.32 (1.02)	18.96 (0.28)	6.36 (1.10)	6.46E-07	

Table 9: Glaucoma explanation performance of various XAI methods integrated with Swin-Tiny on the internal test dataset. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Grad-CAM	Optic cup	5.10 (0.20)	0.71 (0.00)	4.39 (0.20)	4.97E-77	
	Optic disk	11.33 (0.46)	1.80 (0.00)	9.53 (0.46)	5.91E-84	
	Blood vessels	31.42 (1.53)	18.96 (0.28)	12.46 (1.38)	1.98E-15	
XGrad-CAM	Optic cup	2.90 (0.15)	0.71 (0.00)	2.18 (0.15)	8.19E-38	
	Optic disk	6.68 (0.36)	1.80 (0.00)	4.88 (0.36)	1.03E-41	
	Blood vessels	25.52 (1.28)	18.96 (0.28)	6.56 (1.25)	7.32E-07	
Swin-Tiny	Optic cup	9.24 (0.23)	0.71 (0.00)	8.53 (0.20)	3.47E-204	
	Optic disk	21.65 (0.46)	1.80 (0.00)	19.85 (0.46)	1.05E-233	
	Blood vessels	32.84 (1.35)	18.96 (0.28)	13.88 (1.38)	1.52E-15	
Eigen-CAM	Optic cup	6.27 (0.18)	0.71 (0.00)	5.56 (0.18)	1.52E-125	
	Optic disk	13.53 (0.36)	1.80 (0.00)	11.73 (0.36)	2.49E-130	
	Blood vessels	35.76 (1.43)	18.96 (0.28)	16.80 (1.35)	2.14E-24	
Layer-CAM	Optic cup	6.88 (0.18)	0.71 (0.00)	6.17 (0.18)	2.21E-145	
	Optic disk	15.79 (0.38)	1.80 (0.00)	13.99 (0.38)	6.39E-164	
	Blood vessels	30.07 (1.07)	18.96 (0.28)	11.11 (1.15)	2.95E-15	

Table 10: Glaucoma explanation performance of various XAI methods integrated with VGG-11 on the external test dataset of Drishti-GS. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
XGrad-CAM	Optic cup	10.23 (0.84)	2.02 (0.05)	8.22 (0.84)	5.53E-18	
	Optic disk	14.57 (1.17)	3.44 (0.03)	11.13 (1.20)	6.77E-18	
	Optic cup	13.07 (0.82)	2.02 (0.05)	11.05 (0.79)	7.59E-26	
VGG-11	Optic disk	18.64 (1.12)	3.44 (0.03)	15.19 (1.10)	1.55E-26	
	Optic cup	27.72 (0.99)	2.02 (0.05)	25.70 (0.94)	4.14E-65	
	Optic disk	41.05 (1.43)	3.44 (0.03)	37.60 (1.43)	6.44E-71	
Eigen-CAM	Optic cup	33.12 (0.84)	2.02 (0.05)	31.10 (0.82)	2.53E-97	
	Optic disk	52.27 (1.17)	3.44 (0.03)	48.83 (1.17)	1.34E-125	
	Optic cup	34.60 (0.87)	2.02 (0.05)	32.59 (0.84)	1.66E-105	
Layer-CAM	Optic disk	52.58 (1.10)	3.44 (0.03)	49.13 (1.10)	5.34E-127	

Table 11: Glaucoma explanation performance of various XAI methods integrated with ResNet-18 on the external test dataset od Drishti-GS. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
XGrad-CAM	Optic cup	15.78 (1.17)	2.02 (0.05)	13.77 (1.15)	4.55E-28	
	Optic disk	22.05 (1.53)	3.44 (0.03)	18.60 (1.53)	4.79E-28	
	Optic cup	5.43 (0.61)	2.02 (0.05)	3.41 (0.59)	4.40E-07	
ResNet-18	Optic disk	8.17 (0.84)	3.44 (0.03)	4.73 (0.84)	4.63E-07	
	Optic cup	16.25 (0.92)	2.02 (0.05)	14.23 (0.92)	6.15E-33	
	Optic disk	23.50 (1.28)	3.44 (0.03)	20.06 (1.28)	8.22E-34	
Eigen-CAM	Optic cup	20.34 (1.02)	2.02 (0.05)	18.32 (1.02)	4.05E-41	
	Optic disk	30.52 (1.51)	3.44 (0.03)	27.08 (1.51)	3.19E-45	
	Optic cup	17.18 (1.12)	2.02 (0.05)	15.17 (1.12)	4.56E-35	
Layer-CAM	Optic disk	52.58 (1.10)	3.44 (0.03)	49.13 (1.10)	5.34E-127	

Table 12: Glaucoma explanation performance of various XAI methods integrated with DeiT-Tiny on the external test dataset of Drishti-GS. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Grad-CAM	Optic cup	11.95 (0.87)	2.02 (0.05)	9.94 (0.87)	2.35E-20	
	Optic disk	17.58 (1.28)	3.44 (0.03)	14.13 (1.28)	2.21E-20	
XGrad-CAM	Optic cup	10.42 (0.79)	2.02 (0.05)	8.41 (0.79)	1.29E-19	
	Optic disk	15.09 (1.10)	3.44 (0.03)	11.65 (1.10)	7.11E-20	
DeiT-Tiny	Optic cup	7.85 (0.82)	2.02 (0.05)	5.83 (0.77)	1.35E-10	
	Optic disk	11.73 (1.10)	3.44 (0.03)	8.29 (1.10)	2.82E-10	
Score-CAM	Optic cup	6.39 (0.87)	2.02 (0.05)	4.38 (0.87)	1.25E-07	
	Optic disk	10.48 (1.17)	3.44 (0.03)	7.04 (1.17)	6.94E-09	
Eigen-CAM	Optic cup	19.71 (0.69)	2.02 (0.05)	17.69 (0.69)	4.33E-58	
	Optic disk	28.62 (0.89)	3.44 (0.03)	25.17 (0.92)	1.01E-64	

Table 13: Glaucoma explanation performance of various XAI methods integrated with Swin-Tiny on the external test dataset of Drishti-GS. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Grad-CAM	Optic cup	19.66 (1.02)	2.02 (0.05)	17.65 (0.99)	3.03E-36	
	Optic disk	28.57 (1.43)	3.44 (0.03)	25.12 (1.45)	3.36E-37	
XGrad-CAM	Optic cup	12.97 (0.99)	2.02 (0.05)	10.96 (0.99)	2.92E-22	
	Optic disk	19.15 (1.33)	3.44 (0.03)	15.71 (1.33)	1.85E-23	
Swin-Tiny	Optic cup	28.56 (1.05)	2.02 (0.05)	26.55 (1.02)	1.30E-66	
	Optic disk	42.79 (1.58)	3.44 (0.03)	39.34 (1.58)	1.07E-72	
Score-CAM	Optic cup	25.04 (1.15)	2.02 (0.05)	23.02 (1.15)	5.47E-52	
	Optic disk	36.57 (1.61)	3.44 (0.03)	33.12 (1.61)	1.30E-54	
Eigen-CAM	Optic cup	17.44 (0.89)	2.02 (0.05)	15.43 (0.87)	5.27E-38	
	Optic disk	26.21 (1.15)	3.44 (0.03)	22.77 (1.12)	8.48E-42	

Table 14: Glaucoma explanation performance of various XAI methods integrated with VGG-11 on the external test dataset of FIVES. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
VGG-11	Grad-CAM	Blood vessels	17.48 (0.31)	14.29 (0.10)	3.19 (0.38)	3.46E-17
	XGrad-CAM	Blood vessels	17.02 (0.31)	14.29 (0.10)	2.73 (0.36)	1.01E-12
	Score-CAM	Blood vessels	24.11 (0.38)	14.29 (0.10)	9.82 (0.38)	6.02E-93
	Eigen-CAM	Blood vessels	30.67 (0.46)	14.29 (0.10)	16.38 (0.38)	3.39E-218
	Layer-CAM	Blood vessels	25.30 (0.33)	14.29 (0.10)	11.01 (0.36)	2.35E-134

Table 15: Glaucoma explanation performance of various XAI methods integrated with ResNet-18 on the external test dataset of FIVES. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
ResNet-18	Grad-CAM	Blood vessels	19.60 (0.46)	14.29 (0.10)	5.31 (0.48)	1.91E-25
	XGrad-CAM	Blood vessels	16.77 (0.31)	14.29 (0.10)	2.48 (0.33)	8.98E-14
	Score-CAM	Blood vessels	23.59 (0.31)	14.29 (0.10)	9.30 (0.28)	6.16E-116
	Eigen-CAM	Blood vessels	27.89 (0.38)	14.29 (0.10)	13.60 (0.33)	4.44E-196
	Layer-CAM	Blood vessels	26.56 (0.38)	14.29 (0.10)	12.27 (0.33)	3.14E-183

Table 16: Glaucoma explanation performance of various XAI methods integrated with DeiT-Tiny on the external test dataset of FIVES. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
DeiT-Tiny	Grad-CAM	Blood vessels	16.41 (0.36)	14.29 (0.10)	2.12 (0.36)	1.56E-08
	XGrad-CAM	Blood vessels	15.79 (0.36)	14.29 (0.10)	1.50 (0.33)	3.29E-06
	Score-CAM	Blood vessels	14.97 (0.41)	14.29 (0.10)	0.68 (0.41)	1.05E-01
	Eigen-CAM	Blood vessels	14.88 (0.31)	14.29 (0.10)	0.59 (0.28)	3.05E-02
	Layer-CAM	Blood vessels	20.78 (0.36)	14.29 (0.10)	6.49 (0.33)	1.73E-65

Table 17: Glaucoma explanation performance of various XAI methods integrated with Swin-Tiny on the external test dataset of FIVES. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomy	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Swin-Tiny	Grad-CAM	Blood vessels	21.40 (0.38)	14.29 (0.10)	7.11 (0.41)	8.20E-55
	XGrad-CAM	Blood vessels	18.96 (0.41)	14.29 (0.10)	4.67 (0.36)	3.50E-31
	Score-CAM	Blood vessels	26.53 (0.38)	14.29 (0.10)	12.24 (0.38)	9.14E-128
	Eigen-CAM	Blood vessels	20.61 (0.54)	14.29 (0.10)	6.32 (0.43)	2.87E-42
	Layer-CAM	Blood vessels	21.31 (0.41)	14.29 (0.10)	7.02 (0.38)	2.13E-62

Table 18: Pearson’s and Spearman’s correlation coefficient between model predictive performance measured by AUROC and model focus area of activation ratio on anatomical structures.

Evaluation	Anatomy	Pearson’s Correlation	P value	Spearman’s Correlation	P value
Internal	Optic cup	0.42	2.04E-06	0.42	1.79E-06
	Optic disk	0.43	1.26E-06	0.41	3.77E-06
	Blood vessels	0.13	1.57E-01	0.16	7.81E-02
	Optic cup	0.27	2.51E-03	0.29	1.19E-03
External	Optic disk	0.28	1.91E-03	0.29	1.29E-03
	Blood vessels	0.27	2.40E-03	0.28	1.77E-03

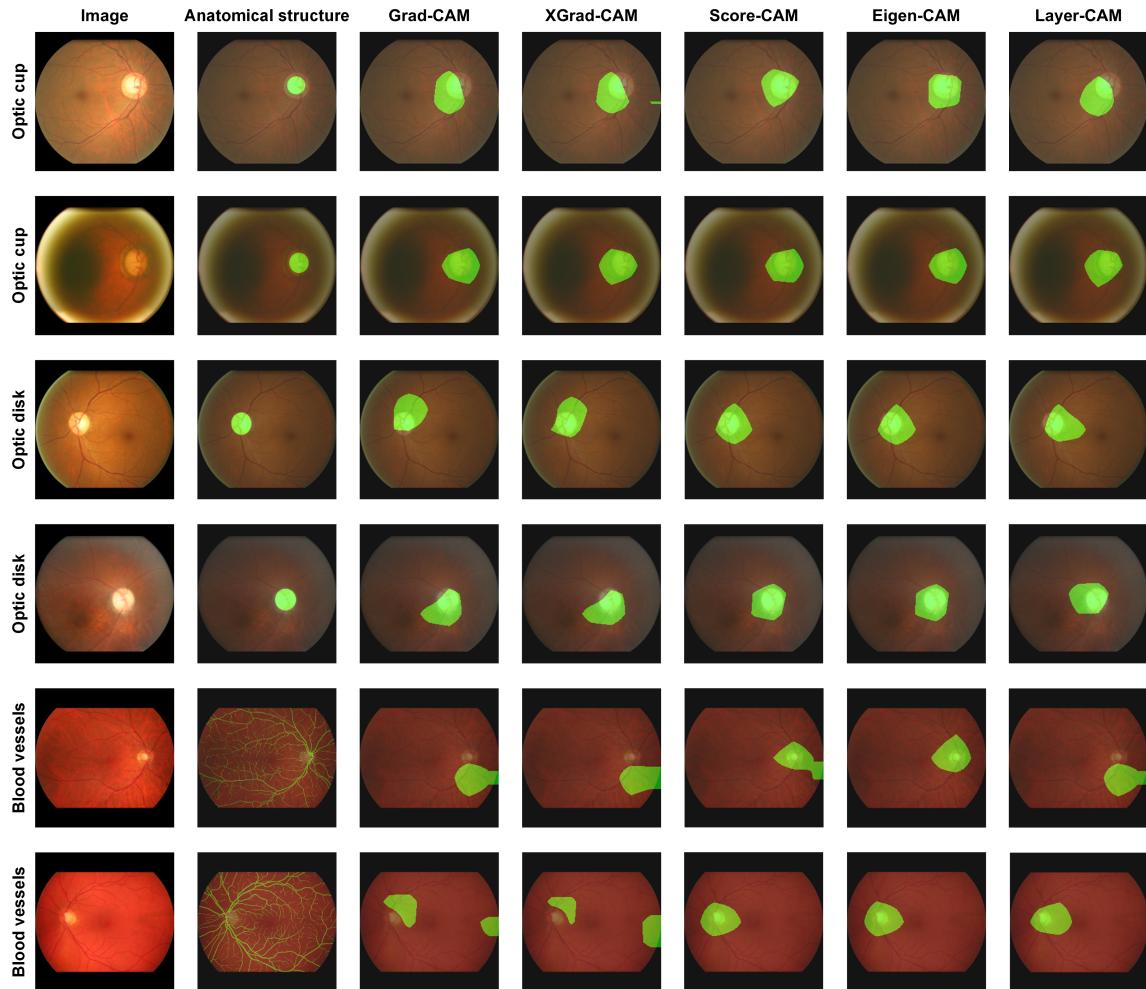


Figure 1: Visualization comparison of anatomical structures and VGG-11 explanations by different CAM methods on the internal test dataset

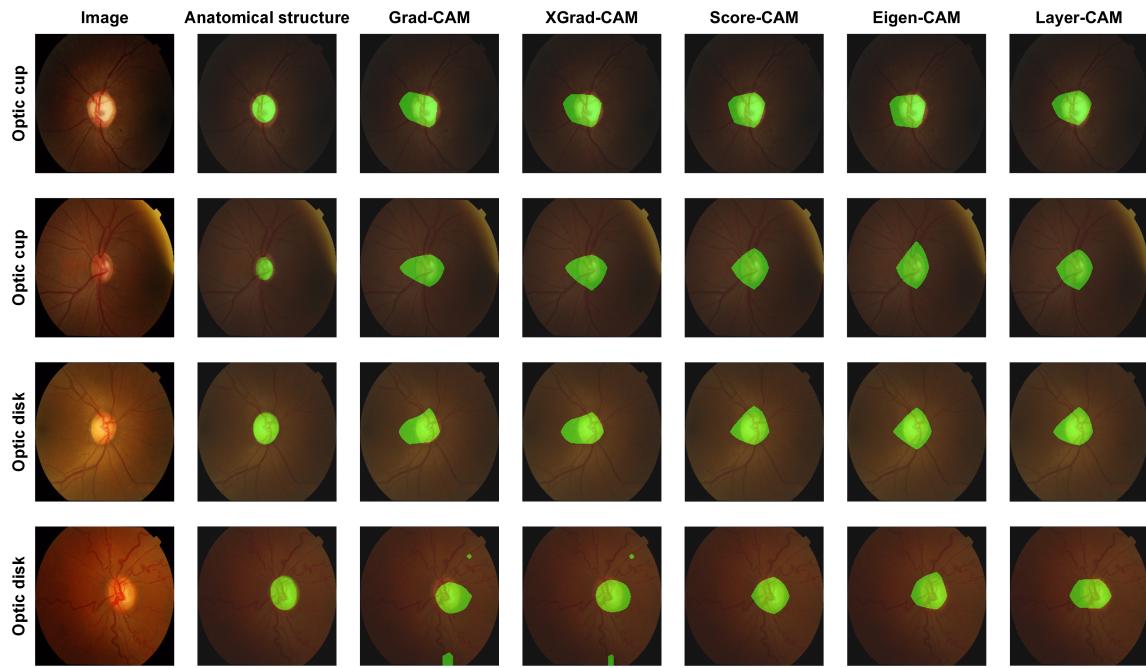


Figure 2: Visualization comparison of anatomical structures and VGG-11 explanations by different CAM methods on the external test dataset of Drishti-GS

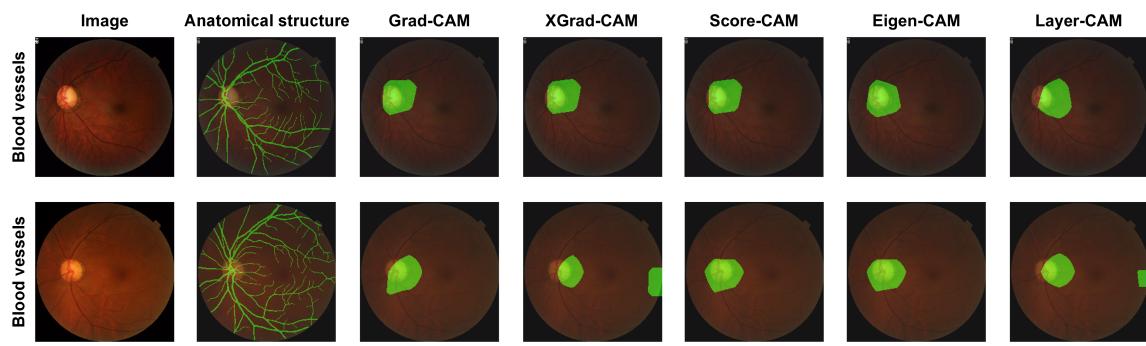


Figure 3: Visualization comparison of anatomical structures and VGG-11 explanations by different CAM methods on the external test dataset of FIVES

## References

- Nahida Akter, John Fletcher, Stuart Perry, et al. Glaucoma diagnosis using multi-feature analysis and a deep learning technique. *Scientific Reports*, 2022.
- Attila Budai, Rüdiger Bock, Andreas Maier, et al. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, 2013.
- Tomasz Burzykowski, Melvin Geubbelsmans, Axel-Jan Rousseau, et al. Validation of machine learning algorithms. *American Journal of Orthodontics and Dentofacial Orthopedics*, 2023.
- Lingping Cen, Jie Ji, Jianwei Lin, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications*, 2021.
- Lyndon Chan, Mahdi S Hosseini, and Konstantinos N Plataniotis. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *International Journal of Computer Vision*, 2021.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.
- Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 1987.
- Edgar C Fieller, Herman O Hartley, and Egon S Pearson. Tests for rank correlation coefficients. *Biometrika*, 1957.
- Ruigang Fu, Qingyong Hu, Xiaohu Dong, et al. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *Proceedings of the British Machine Vision Conference*, 2020.
- David F Garway-Heath, Simon T Ruben, Ananth Viswanathan, et al. Vertical cup/disc ratio in relation to optic disc size: its value in the assessment of the glaucoma suspect. *British Journal of Ophthalmology*, 1998.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
- A Victor Ikechukwu, S Murali, R Deepu, and RC Shivamurthy. Resnet-50 vs vgg-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images. *Global Transitions Proceedings*, 2021.
- Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, et al. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021.
- Kai Jin, Xingru Huang, Jingxing Zhou, et al. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 2022.

- Lican Kang, Han Yuan, and Chang Zhu. Error analysis of fitted q-iteration with relu-activated deep neural networks. In *Proceedings of the International Conference on Learning Representations*, 2023.
- Liu Li, Mai Xu, Hanruo Liu, et al. A large-scale database and a cnn model for attention-based glaucoma detection. *IEEE Transactions on Medical Imaging*, 2019.
- Junzhong Liang, David R Williams, and Donald T Miller. Supernormal vision and high-resolution retinal imaging through adaptive optics. *Journal of the Optical Society of America A*, 1997.
- WangMin Liao, BeiJi Zou, RongChang Zhao, et al. Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 2019.
- Bingyan Liu, Daru Pan, Zhenbin Shuai, et al. Ecsd-net: A joint optic disc and cup segmentation and glaucoma classification network based on unsupervised domain adaptation. *Computer Methods and Programs in Biomedicine*, 2022.
- Ze Liu, Yutong Lin, Yue Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- Delaram Mirzania, Atalie C Thompson, and Kelly W Muir. Applications of deep learning in detection of glaucoma: a systematic review. *European Journal of Ophthalmology*, 2021.
- Anirban Mitra, Priya Shankar Banerjee, Sudipta Roy, et al. The region of interest localization for glaucoma analysis from retinal fundus image using deep learning. *Computer Methods and Programs in Biomedicine*, 2018.
- José Morano, Álvaro S Hervella, Jorge Novo, et al. Simultaneous segmentation and classification of the retinal arteries and veins from color fundus images. *Artificial Intelligence in Medicine*, 2021.
- James E Morgan, Ioanna Bourtsoukli, Kadaba N Rajkumar, et al. The accuracy of the inferior; superior; nasal; temporal neuroretinal rim area rule for diagnosing glaucomatous optic disc damage. *Ophthalmology*, 2012.
- Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *Proceedings of the International Joint Conference on Neural Networks*, 2020.
- José Ignacio Orlando, João Barbosa Breda, Karel Van Keer, et al. Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, 2018.
- Karl Pearson. Notes on the history of correlation. *Biometrika*, 1920.
- Dian Qin, Jia-Jun Bu, Zhe Liu, et al. Efficient medical image segmentation based on knowledge distillation. *IEEE Transactions on Medical Imaging*, 2021.

- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 1986.
- Daniel Fernando Santos-Bustos, Binh Minh Nguyen, and Helbert Eduardo Espitia. Towards automated eye cancer classification via vgg and resnet networks using transfer learning. *Engineering Science and Technology*, 2022.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- Rutuja Shinde. Glaucoma detection in retinal fundus images using u-net and supervised machine learning algorithms. *Intelligence-Based Medicine*, 2021.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.
- Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, et al. Drishti-gs: Retinal image dataset for optic nerve head (ohn) segmentation. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 2014.
- Student. The probable error of a mean. *Biometrika*, 1908.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, 2017.
- Kaveri A Thakoor, Xinhui Li, Emmanouil Tsamis, et al. Enhancing the accuracy of glaucoma detection from oct probability maps using convolutional neural networks. In *Proceedings of the International Conference of the Engineering in Medicine and Biology Society*, 2019.
- Atalie C Thompson, Alessandro A Jammal, and Felipe A Medeiros. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Translational vision science & technology*, 2020.
- Daniel Shu Wei Ting, Louis R Pasquale, et al. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 2019.
- Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning*, 2021.
- Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, et al. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 2022.
- Rohit Varma, William C Steinmann, and Ingrid U Scott. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*, 1992.

- Haofan Wang, Zifan Wang, Mengnan Du, et al. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- Adam White, Kwun Ho Ngan, James Phelan, et al. Contrastive counterfactual visual explanations with overdetermination. *Machine Learning*, 2023.
- Chulin Wu, Heye Zhang, Jiaqi Chen, et al. Vessel-gan: Angiographic reconstructions from myocardial ct perfusion with explainable generative adversarial networks. *Future Generation Computer Systems*, 2022.
- Feng Xie, Han Yuan, Yilin Ning, et al. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of Biomedical Informatics*, 2022.
- Ying Xue, Jiazhu Zhu, Xiaoling Huang, et al. A multi-feature deep learning system to enhance glaucoma severity diagnosis with high accuracy and fast speed. *Journal of Biomedical Informatics*, 2022.
- Han Yuan, Pengtao Jiang, and Gangming Zhao. Human-guided design to explain deep learning-based pneumothorax classifier. In *Proceedings of the International Conference on Medical Imaging with Deep Learning*, 2023a.
- Han Yuan, Mingxuan Liu, Lican Kang, et al. An empirical study of the effect of background data size on the stability of shapley additive explanations (shap) for deep learning models. In *Proceedings of the International Conference on Learning Representations*, 2023b.
- Xin Yuan, Lingxiao Zhou, Shuyang Yu, et al. A multi-scale convolutional neural network with context for joint segmentation of optic disc and cup. *Artificial Intelligence in Medicine*, 2021.
- Dingwen Zhang, Junwei Han, Gong Cheng, et al. Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- Man Zhang, Yong Zhou, Jiaqi Zhao, et al. A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 2020.
- Zhuo Zhang, Feng Shou Yin, Jiang Liu, et al. Orig-a-light: An online retinal fundus image database for glaucoma analysis and research. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010.
- Aidi Zhao, Hong Su, Chongyang She, et al. Joint optic disc and cup segmentation based on elliptical-like morphological feature and spatial geometry constraint. *Computers in Biology and Medicine*, 2023.
- Rongchang Zhao, Xuanlin Chen, Xiyao Liu, et al. Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. *IEEE Journal of Biomedical and Health Informatics*, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, et al. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.