Opening the black box of deep learning: Exploring the association between model explanations and human domain knowledge in fundus image-based glaucoma diagnosis

Han Yuan[1] YUAN.HAN@U.DUKE.NUS.EDU
Lican Kang[2] KANGLICAN@WHU.EDU.CN
Yong Li[1] LIYONG@U.DUKE.NUS.EDU
Chenkui Miao[3] MIAOCHENKUI@NJMU.EDU.CN
Yuan Luo[2] YUANLUO@WHU.EDU.CN
Chang Zhu[4] CHANGZHU@HUST.EDU.CN

[1] *Duke-NUS Medical School, National University of Singapore*
[2] *School of Mathematics and Statistics, Wuhan University*
[3] *Jiangsu Provincial People's Hospital, Nanjing Medical University*
[4] *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 and further explore the correlation between model decisions and human domain knowledge of anatomical structures (optic cup, optic disk, and blood vessels) for glaucoma diagnosis. Four deep neural networks, including VGG-11, ResNet-18, DeiT-Tiny, and Swin Transformer-Tiny, are developed using binary diagnostic labels of glaucoma. Five CAM methods (Grad-CAM, XGrad-CAM, Score-CAM, Eigen-CAM, and Layer-CAM) are employed to highlight the model focus area. Then paired-sample t-test is applied to compare the percentage of anatomical structures in the model focus area to the proportion of anatomical structures in the entire image. After that, Pearson's and Spearman's correlation tests are implemented to examine the relationship between model predictive ability and the percentage of anatomical structures in the model focus area. All deep learning models consistently displayed statistically significantly higher percentages of anatomical structure 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 structure 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, as demonstrated through various statistical tests. We anticipate that our work can help alleviate clinicians' concerns regarding the reliability of deep learning in healthcare. For reproducibility, the code and dataset have been released at GitHub[1].

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/Exploring-the-association-between-model-explanations-and-human-knowledge
https://github.com/TheBeastCoding/standardized-multichannel-dataset-glaucoma

© H. Yuan, L. Kang, Y. Li, C. Miao, Y. Luo & C. Zhu.

1. Introduction

In the last decade, deep learning has reshaped various medical diagnoses in Ophthalmology (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 address the interpretability issue and open the black box of deep learning, various methods have been proposed to explain model performance 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 employed 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 region of interest (ROI) or the clinically irrelevant area (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, the clinicians concern about the model reliability would be alleviated (Yuan et al., 2023b).

Glaucoma is a neuro-degenerative ophthalmological disease caused by the increase of intra-ocular pressure and will progress into complete blindness without early medical intervention (Varma et al., 1992; 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; 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 concentrates 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. 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 used a dataset named Standardized Multi-channel Dataset for the Glaucoma (SMDG-19) (Kiefer et al., 2023), which is a comprehensive compilation and standardization of 19 public full-fundus glaucoma images. This dataset comprises a total of 12,449 fundus images, each with a resolution of 512512 pixels and a binary diagnostic label indicating the presence or absence of glaucoma. Notably, two subsets of this extensive dataset are enriched with pixel-level annotations for anatomical structures, specifically the optic cup and disk, as well as blood vessels. The subset with optic cup and disk annotations encompasses 665 cases of glaucoma and 2,140 healthy controls, while the subset with blood vessels annotations includes 186 glaucoma cases and 276 healthy controls. We combined the two subsets to develop a glaucoma classifier, implement model explanations, and explore the association between model focus area and domain knowledge-based anatomical structures. To fully exploit the limited samples with pixel-level annotations of the optic cup, optic disk, and blood vessels, we implement double cross-validation. This involved initially dividing the dataset into three independent subsets, followed by the allocation of training, validation, and test datasets within each of these subsets. Ultimately, this process generated A_3^3: 6 distinct scenarios to comprehensively evaluate glaucoma classification by various deep neural networks and model focus areas generated by different CAM methods (Burzykowski et al., 2023). The concrete details of the data split are shown in Table 1.

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 binary diagnostic label Y_i^{train} and additional 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 on I_i^{test} and Y_i^{test}. After the model devel-

opment, 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.

Deep learning classifier training 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_θ parameterized by θ. Taking the input of I_i^{train}, f_θ 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 avoid overfitting of f_θ, the validation dataset D^{val} is applied to early stop the optimization procedure: If $l(f_\theta(I_i^{val}), Y_i^{val})$ has not decreased for a pre-defined epoch number N_{epoch}, the iteration of θ will be terminated and the last θ showing a decrease on $l(\theta; D^{val})$ will be saved as the optimal parameter θ^*. After the model training, we evaluate the classification performance of the trained f_{θ^*} 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 evaluation metrics, 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). Apart from the mean values on the test dataset, the standard error (SE) is reported based on the bootstrapping of samples in the test dataset (Efron, 1987).

Considering the middle-scale characteristic of the used dataset, we developed four glaucoma classifiers using 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 employ Stochastic Gradient Descent (SGD) (Rumelhart et al., 1986) with a learning rate of 0.001 and a momentum of 0.9. Each deep learning model undergoes training for 100 epochs, and early-stopping with a patience parameter of 10 is implemented based on model performance on the validation dataset. It's worth noting that all experiments are conducted across the six scenarios mentioned earlier, and the reported performance metrics represent the averaged results from these scenarios.

2.3. CAM explanation for deep learning classifiers

The developed modelf_{θ^*} 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 $R(I_i^{test})$ is outlined in I_i^{test} using a popular paradigm named CAM. CAM family calculates each pixels importance $E_{(w,h)}(I_i^{test})$ towards $f_{\theta^*}(I_i^{test})$ and 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 pixel-level importance $E_{(w,h)}(I_i^{test})$. Then the top 5% important pixels are extracted as the model focus area $R(I_i^{test})$. After that, we compare the two percentages of anatomical structures in the model focus area $(R(I_i^{test}) \bigcap A_i^{test})/R(I_i^{test})$ and the whole

image A_i^{test}/I_i^{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 two percentages. On the other hand, if the model prioritizes anatomical structures, the first percentage would be noticeably higher than the second.

To establish a robust analysis of the relationship between model decision logic (focus area) and medical domain knowledge (anatomical area of the optic cup, optic disk, and blood vessels), a paired-sample t-test (Student, 1908) is employed to compare the two percentages. 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 and the proportion of anatomical areas within its focus region. Code and dataset for reproduction are publicly available at GitHub[1].

3. Results

First, we quantitatively showed the model classification performance on glaucoma diagnosis in Table 2. VGG-11 achieved best performance across all classification evaluation metrics: an AUROC of 0.787 (SE: 0.015), an AUPRC of 0.569 (SE: 0.030), an accuracy of 0.683 (SE: 0.013), a sensitivity of 0.763 (SE: 0.025), a specificity of 0.655 (SE: 0.016), a PPV of 0.440 (SE: 0.019), and an NPV of 0.889 (SE: 0.013). ResNet-18, DeiT-Tiny, and Swin-Tiny also delivered a commendable performance in classifying glaucoma, demonstrating the feasibility of downstream explanation of the models' decision logic.

Table 3,4,5,6 presents the glaucoma explanation performance of different deep learning models based on various CAM methods. From the P values of the paired-sample t-test, we observed that anatomical structures of the optic cup, optic disk, and blood vessels play a significant role in fundus images-based glaucoma classification. A further Pearson's and Spearman's correlation study was conducted between model classification performance and the percentage of anatomical structures in their focus area. The results of the correlation analyses are summarized in Table 7. Notably, there exists a statistically significantly positive relationship between model predictive performance and their focus on anatomical structures, which aligns harmoniously with medical knowledge of glaucoma diagnosis.

To provide a comparative visualization with anatomical structures of the optic cup, optic disk, and blood vessels, Figure 1 portrays the model explanations of VGG-11 generated by five distinct CAM methods. The striking resemblance between the model's focus areas and the anatomical regions underscores a compelling point: despite being trained solely with image-level labels and lacking specific information regarding anatomical structures, deep neural networks rely on similar regions to classify glaucoma, akin to the decision-making process of human clinicians. This observation also demonstrated the capability of data-driven deep learning models in domain knowledge distillation.

4. Discussion

In this study, we developed four deep neural networks based on both CNN and Transformer architectures for glaucoma classification. We compared different 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 behind 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 focus on 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 Transformer-based models (DeiT-Tiny and Swin-Tiny) with more sophisticated structure design and inference logic. Such a result 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). Although Vision Transformer and its variants have achieved state-of-the-art performance on diverse general computer vision tasks (He et al., 2023; Shamshad et al., 2023), the experiments on glaucoma classification demonstrate that the conventional CNN model of VGG-11 is still capable even achieving the best performance in handling tasks with middle-scale samples as witnessed in Table 2. Additionally, VGG-11 holds relatively lower complexity than transformer models in terms of structure design. The inherent preservation of spatial information in VGG-11 leads to its focus region by various CAM methods overlaps more with optic cup, optic disk, and blood vessels than the transformer models, 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 classic CAM methods to pinpoint the significant sub-region towards model final decisions. This approach has been employed by prior researchers to segment areas of lesions or regions strongly associated with specific diseases, utilizing deep learning classifiers trained with limited 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 simply calculating overlap metrics such as Intersection over Union (IoU). In this study, we introduced a novel 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 various deep learning models explained by diverse CAM methods consistently and statistically significantly focus more on the anatomical structures of the optic cup, optic disk, and blood vessels that clinicians rely on for glaucoma diagnosis, which was further supported by the positive dependence 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 for CNN and Transformer models. 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 developed models on existing images. In future research, we intend to explore an alternative approach that involves the use of image generation methods to manipulate anatomical

structures (Zhou et al., 2023). 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 of glaucoma such as bleeding and notch (Liao et al., 2019). Lastly, this research exclusively explored glaucoma diagnosis and we plan to evaluate the association between model focus area and domain knowledge-based anatomical structure in a broader spectrum of ophthalmological tasks including but not limited to 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 the field of medicine. In this study, we validated the statistically significant alignment of decision-making logic between deep neural networks and clinical domain knowledge through several rigorous statistical tests. We hope that this research may mitigate clinicians' concerns regarding the reliability of deep learning in glaucoma diagnosis.

Table 1: 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	221	713
	Blood vessels	62	92
2	Optic cup & disk	222	713
	Blood vessels	62	92
3	Optic cup & disk	222	714
	Blood vessels	62	92

Table 2: Glaucoma classification performance of various deep learning models. Model performance were averaged across six-fold cross-validation.

Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	PPV	NPV
VGG-11	0.787 (0.015)	0.569 (0.030)	0.683 (0.013)	0.763 (0.025)	0.655 (0.016)	0.440 (0.019)	0.889 (0.013)
ResNet-18	0.673 (0.017)	0.417 (0.028)	0.587 (0.015)	0.685 (0.026)	0.553 (0.017)	0.353 (0.019)	0.834 (0.015)
DeiT-Tiny	0.729 (0.016)	0.498 (0.029)	0.632 (0.013)	0.729 (0.026)	0.597 (0.015)	0.389 (0.020)	0.865 (0.014)
Swin-Tiny	0.707 (0.017)	0.476 (0.028)	0.624 (0.013)	0.679 (0.025)	0.605 (0.015)	0.377 (0.019)	0.845 (0.015)

Table 3: Glaucoma explanation performance of various XAI methods integrated with VGG-11. Model performance were averaged across six-fold cross-validation. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomical structure	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
VGG-11	Grad-CAM	Optic cup	4.05 (0.08)	0.40 (0.00)	3.65 (0.08)	0.00E+00
		Optic disk	12.38 (0.18)	1.70 (0.00)	10.68 (0.18)	0.00E+00
		Blood vessels	2.90 (0.10)	1.32 (0.03)	1.58 (0.10)	1.76E-66
	XGrad-CAM	Optic cup	4.17 (0.08)	0.40 (0.00)	3.77 (0.08)	0.00E+00
		Optic disk	12.58 (0.18)	1.70 (0.00)	10.88 (0.18)	0.00E+00
		Blood vessels	2.90 (0.10)	1.32 (0.03)	1.58 (0.10)	1.01E-63
	Score-CAM	Optic cup	7.18 (0.08)	0.40 (0.00)	6.78 (0.08)	0.00E+00
		Optic disk	27.38 (0.15)	1.70 (0.00)	25.68 (0.15)	0.00E+00
		Blood vessels	5.69 (0.10)	1.32 (0.03)	4.37 (0.08)	6.83E-267
	Eigen-CAM	Optic cup	6.44 (0.08)	0.40 (0.00)	6.04 (0.08)	0.00E+00
		Optic disk	24.64 (0.20)	1.70 (0.00)	22.94 (0.20)	0.00E+00
		Blood vessels	6.18 (0.10)	1.32 (0.03)	4.86 (0.08)	1.02E-285
	Layer-CAM	Optic cup	6.94 (0.08)	0.40 (0.00)	6.54 (0.08)	0.00E+00
		Optic disk	25.49 (0.15)	1.70 (0.00)	23.79 (0.15)	0.00E+00
		Blood vessels	4.98 (0.10)	1.32 (0.03)	3.66 (0.08)	4.55E-220

Table 4: Glaucoma explanation performance of various XAI methods integrated with ResNet-18. Model performance were averaged across six-fold cross-validation. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomical structure	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
ResNet-18	Grad-CAM	Optic cup	1.86 (0.05)	0.40 (0.00)	1.46 (0.05)	7.18E-110
		Optic disk	5.18 (0.15)	1.70 (0.00)	3.48 (0.15)	5.68E-112
		Blood vessels	2.04 (0.08)	1.32 (0.03)	0.71 (0.08)	4.12E-15
	XGrad-CAM	Optic cup	1.07 (0.05)	0.40 (0.00)	0.67 (0.05)	1.12E-52
		Optic disk	4.04 (0.13)	1.70 (0.00)	2.35 (0.13)	8.46E-75
		Blood vessels	2.02 (0.08)	1.32 (0.03)	0.69 (0.08)	4.74E-19
	Score-CAM	Optic cup	1.43 (0.05)	0.40 (0.00)	1.03 (0.05)	8.27E-106
		Optic disk	5.46 (0.13)	1.70 (0.00)	3.76 (0.13)	3.26E-155
		Blood vessels	2.79 (0.08)	1.32 (0.03)	1.46 (0.08)	1.48E-64
	Eigen-CAM	Optic cup	4.27 (0.08)	0.40 (0.00)	3.87 (0.08)	0.00E+00
		Optic disk	16.36 (0.18)	1.70 (0.00)	14.66 (0.18)	0.00E+00
		Blood vessels	4.99 (0.08)	1.32 (0.03)	3.67 (0.08)	8.93E-200
	Layer-CAM	Optic cup	3.67 (0.08)	0.40 (0.00)	3.27 (0.08)	0.00E+00
		Optic disk	14.43 (0.15)	1.70 (0.00)	12.73 (0.15)	0.00E+00
		Blood vessels	4.53 (0.10)	1.32 (0.03)	3.21 (0.10)	2.44E-151

Table 5: Glaucoma explanation performance of various XAI methods integrated with DeiT-Tiny. Model performance were averaged across six-fold cross-validation. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomical structure	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
DeiT-Tiny	Grad-CAM	Optic cup	2.40 (0.08)	0.40 (0.00)	2.00 (0.08)	3.12E-183
		Optic disk	7.41 (0.15)	1.70 (0.00)	5.71 (0.15)	1.56E-228
		Blood vessels	1.64 (0.08)	1.32 (0.03)	0.32 (0.05)	1.75E-06
	XGrad-CAM	Optic cup	1.71 (0.05)	0.40 (0.00)	1.31 (0.05)	1.32E-166
		Optic disk	6.95 (0.15)	1.70 (0.00)	5.25 (0.15)	1.76E-251
		Blood vessels	3.02 (0.10)	1.32 (0.03)	1.70 (0.10)	1.89E-63
	Score-CAM	Optic cup	2.23 (0.05)	0.40 (0.00)	1.83 (0.05)	2.13E-156
		Optic disk	7.23 (0.15)	1.70 (0.00)	5.54 (0.15)	1.61E-189
		Blood vessels	1.40 (0.08)	1.32 (0.03)	0.08 (0.08)	2.62E-01
	Eigen-CAM	Optic cup	1.56 (0.05)	0.40 (0.00)	1.16 (0.05)	7.03E-107
		Optic disk	5.93 (0.18)	1.70 (0.00)	4.23 (0.18)	8.62E-145
		Blood vessels	2.41 (0.10)	1.32 (0.03)	1.09 (0.10)	4.35E-23
	Layer-CAM	Optic cup	4.58 (0.08)	0.40 (0.00)	4.18 (0.08)	0.00E+00
		Optic disk	15.31 (0.18)	1.70 (0.00)	13.61 (0.18)	0.00E+00
		Blood vessels	3.67 (0.08)	1.32 (0.03)	2.35 (0.08)	2.73E-117

Table 6: Glaucoma explanation performance of various XAI methods integrated with Swin-Tiny. Model performance were averaged across six-fold cross-validation. Top 5% activated pixels were selected as the model focus area.

Model	XAI	Anatomical structure	Activation ratio (%)	Structure ratio (%)	Ratio difference (%)	P value
Swin-Tiny	Grad-CAM	Optic cup	2.04 (0.05)	0.40 (0.00)	1.63 (0.05)	2.30E-136
		Optic disk	6.63 (0.18)	1.70 (0.00)	4.93 (0.15)	1.23E-168
		Blood vessels	2.31 (0.10)	1.32 (0.03)	0.99 (0.10)	1.05E-25
	XGrad-CAM	Optic cup	1.63 (0.05)	0.40 (0.00)	1.23 (0.05)	8.89E-107
		Optic disk	5.68 (0.15)	1.70 (0.00)	3.99 (0.15)	1.78E-139
		Blood vessels	2.27 (0.08)	1.32 (0.03)	0.95 (0.08)	4.39E-26
	Score-CAM	Optic cup	2.38 (0.05)	0.40 (0.00)	1.98 (0.05)	5.22E-171
		Optic disk	8.51 (0.18)	1.70 (0.00)	6.81 (0.18)	1.91E-224
		Blood vessels	2.31 (0.10)	1.32 (0.03)	0.98 (0.10)	7.41E-23
	Eigen-CAM	Optic cup	3.10 (0.08)	0.40 (0.00)	2.70 (0.08)	1.21E-273
		Optic disk	11.31 (0.18)	1.70 (0.00)	9.61 (0.18)	0.00E+00
		Blood vessels	4.51 (0.10)	1.32 (0.03)	3.19 (0.08)	1.38E-143
	Layer-CAM	Optic cup	2.89 (0.05)	0.40 (0.00)	2.49 (0.05)	8.34E-246
		Optic disk	10.08 (0.18)	1.70 (0.00)	8.38 (0.18)	0.00E+00
		Blood vessels	3.08 (0.10)	1.32 (0.03)	1.76 (0.10)	6.24E-60

Table 7: Pearson's correlation coefficient between model predictive performance (AUROC) and model focus area (Activation ratio) on anatomical structures. Top 5% activated pixels were selected as the model focus area.

XAI	Anatomies	Pearson's Correlation	P value	Spearman's Correlation	P value
Grad-CAM	Optic cup	0.63	1.03E-03	0.60	2.13E-03
	Optic disk	0.55	5.71E-03	0.56	4.81E-03
	Blood vessels	0.48	1.65E-02	0.51	1.16E-02
XGrad-CAM	Optic cup	0.74	3.92E-05	0.84	3.18E-07
	Optic disk	0.74	3.40E-05	0.82	9.31E-07
	Blood vessels	0.66	5.07E-04	0.59	2.36E-03
Score-CAM	Optic cup	0.80	3.37E-06	0.86	7.10E-08
	Optic disk	0.76	1.39E-05	0.83	4.76E-07
	Blood vessels	0.59	2.57E-03	0.61	1.50E-03
Eigen-CAM	Optic cup	0.51	1.11E-02	0.58	2.73E-03
	Optic disk	0.50	1.39E-02	0.56	4.56E-03
	Blood vessels	0.35	9.62E-02	0.53	8.13E-03
Layer-CAM	Optic cup	0.75	2.77E-05	0.77	9.71E-06
	Optic disk	0.70	1.51E-04	0.72	6.53E-05
	Blood vessels	0.54	6.83E-03	0.37	7.73E-02

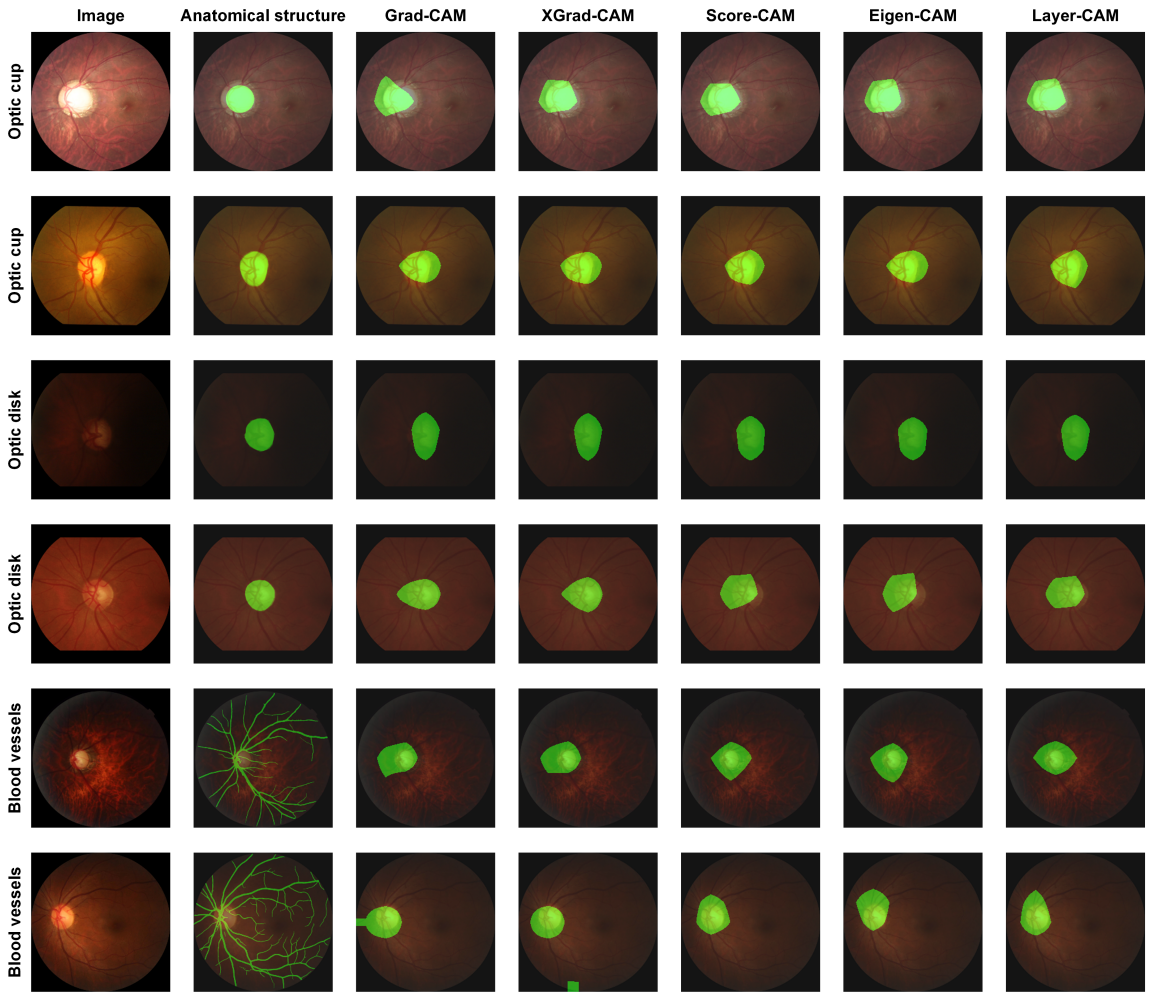

Figure 1: Visualization comparison of anatomical structures and VGG-11 explanations by different CAM methods

References

Nahida Akter, John Fletcher, Stuart Perry, et al. Glaucoma diagnosis using multi-feature analysis and a deep learning technique. *Scientific Reports*, 2022.

Tomasz Burzykowski, Melvin Geubbelmans, 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.

Kelei He, Chen Gan, Zhuoyuan Li, et al. Transformers in medical image analysis. *Intelligent Medicine*, 2023.

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.

Riley Kiefer, Muhammad Abid, Jessica Steen, et al. A catalog of public glaucoma datasets for machine learning applications: A detailed description and analysis of public glaucoma datasets available to machine learning engineers tackling glaucoma-related problems using retinal fundus images and oct images. In *Proceedings of the International Conference on Information System and Data Mining*, 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.

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.

Karl Pearson. Notes on the history of correlation. *Biometrika*, 1920.

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.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, et al. Transformers in medical imaging: A survey. *Medical Image Analysis*, 2023.

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.

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.

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.

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.

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.

Tao Zhou, Qi Li, Huiling Lu, et al. Gan review: Models and medical image fusion applications. *Information Fusion*, 2023.