technologies

MDPI

*Article*

# Analyzing CNN-Based Glaucoma Decision Criteria Using Adversarial Examples

**Shinichiro Ishikawa [1,*], Hiyori Sakemi [2,*], Koki Hirose [2], Tahsina Nabiha Khan [2], Kenshin Mizoe [2], Ikki Osaka [2], Osamu Fukuda [2], Nobuhiko Yamaguchi [2], Masateru Kawakubo [2] and Hiroshi Okumura [2]**

[1] Department of Medical Informatics and Ophthalmology, Saga University Hospital, 5-1-1 Nabeshima, Saga 849-8501, Japan
[2] Department of Science and Engineering, Saga University, 1 Honjo-machi, Saga 840-8502, Japan; 25726024@edu.cc.saga-u.ac.jp (K.H.); 23238108@edu.cc.saga-u.ac.jp (T.N.K.); 23238420@edu.cc.saga-u.ac.jp (K.M.); 24726005@edu.cc.saga-u.ac.jp (I.O.); fukudao@cc.saga-u.ac.jp (O.F.); yamag@cc.saga-u.ac.jp (N.Y.); kwkbmstr@edu.cc.saga-u.ac.jp (M.K.); okumurah@edu.cc.saga-u.ac.jp (H.O.)
* Correspondence: c1441@edu.cc.saga-u.ac.jp (S.I.); 25726015@edu.cc.saga-u.ac.jp (H.S.)

**Abstract**

Glaucoma is a leading cause of blindness, and early detection is critical. Convolutional neural networks (CNNs) have shown impressive performance in glaucoma diagnosis, but their black-box nature remains a barrier to clinical use. Existing explainable AI (XAI) methods such as Grad-CAM have limitations in identifying and quantifying subtle regional features. In this study, we propose a method to clarify what CNNs focus on by analyzing how model performance changes under localized adversarial noise. Using VGG16 for glaucoma classification, we applied noise generated by the Fast Gradient Sign Method (FGSM) to the whole fundus image and to specific subregions, then compared the impact on classification performance. Results showed that perturbations to the optic disc, especially its outer margin, had the greatest effect on model performance. This suggests that the CNN captures fine anatomical features such as optic disc cupping and neuroretinal rim thinning, which aligns with what ophthalmologists typically look for. At the same time, perturbations in the macula and perivascular regions also affected performance, indicating gaps between current clinical diagnostic criteria and the CNN's decision-making process. This approach can help establish the clinical reliability of CNNs and may also reveal features that have not been recognized in conventional clinical practice.

**Keywords:** deep learning; convolutional neural network (CNN); glaucoma; explainable artificial intelligence (XAI); adversarial examples; fundus images; decision-making; optic disc; macula

Check for updates

Academic Editor: Pedro Antonio Gutiérrez

Received: 7 June 2026
Revised: 9 July 2026
Accepted: 14 July 2026
Published: 16 July 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Glaucoma is an irreversible progressive optic neuropathy and one of the leading causes of blindness worldwide [1,2]. The number of people with glaucoma worldwide is estimated to have increased from approximately 4.07 million in 1990 to approximately 7.59 million in 2021, and is projected to reach approximately 9.42 million by 2050 [3]. Since vision loss from glaucoma cannot be recovered once it progresses, early detection and timely intervention are essential.

In clinical practice, glaucoma diagnosis is made by specialists in clinics and ophthalmology departments, based on a comprehensive assessment that combines structural and functional evaluations, such as fundus photography, optical coherence tomography

(OCT), visual field testing, and intraocular pressure measurement. Among these, fundus images provide important information for evaluating structural changes in the optic disc and the retinal nerve fiber layer. Key findings center on the evaluation of morphological changes, including pathological enlargement of the optic disc cupping (assessed by the vertical cup-to-disc ratio), thinning of the neuroretinal rim, and retinal nerve fiber layer defects (RNFLDs) [4]. However, accurately identifying these subtle structural changes requires extensive expertise and experience. Moreover, diagnostic outcomes are influenced by individual judgment, leading to variability in agreement between clinicians and institutions. Indeed, a study examining the concordance of definitive diagnoses among glaucoma specialists reported a low Fleiss' kappa of 0.149, with complete disagreement among three specialists in approximately 10.8% of cases [5]. In addition, as noted above, the number of patients requiring diagnosis and management continues to grow worldwide, and it is becoming increasingly difficult to meet this demand through visual assessment by a limited number of specialists alone. Because AI and deep learning (DL) models generally depend on large volumes of training data to achieve high performance, diseases such as glaucoma—characterized by a large and growing patient population—are considered particularly well suited to AI-based analysis. Against this background, research on AI- and DL-based diagnostic methods that support clinical assessment and enable objective, high-precision screening has been actively pursued in recent years.

To date, numerous fundus image diagnostic algorithms using deep learning models, including convolutional neural networks (CNNs), have been proposed. Studies using large datasets have shown that optimized deep learning models can achieve diagnostic accuracy comparable to ophthalmologists in terms of sensitivity and specificity [6,7]. In recent years, more advanced deep learning approaches extending these foundational CNN architectures have also been proposed. For example, a method that jointly models spatial position and channel-wise features while representing inter-regional relationships through graph structures across multiple resolution scales has demonstrated its effectiveness in fundus image diagnosis tasks [8]. In other domains, approaches such as representation learning guided by reversible transformation structures that preserve semantic consistency across modalities [9], and knowledge distillation methods that improve model generalization by cyclically exchanging knowledge among multiple learners [10], have been proposed, and their applicability to medical image analysis is also anticipated. Furthermore, in the field of medical image analysis, the application of physics-informed neural networks (PINNs), which incorporate physical constraints into image recognition models, has been progressing. Such approaches have been reported across a range of tasks, including image registration, image generation, disease classification, and image reconstruction, and are considered a means of addressing certain limitations of purely data-driven approaches by grounding predictions in physical or anatomical principles [11].

Despite their high classification performance, CNN-based models operate as black boxes, with opaque decision-making processes. This lack of interpretability is a major barrier to clinical adoption, raising concerns about reliability and validity in medical applications [12–14].

To address this, explainable AI (XAI) methods have been introduced to make model decisions more transparent [15]. In fundus image analysis, methods that visualize a model's attention as heatmaps have received particular attention. Grad-CAM (Gradient-weighted Class Activation Mapping) [16], one of the most widely used approaches, uses gradients of the prediction score to highlight which regions of an image the model focused on [14,17].

However, Grad-CAM and similar visualization methods have notable limitations [18]. First, because the final convolutional layer has already lost spatial detail through pooling operations, the resulting heatmaps are spatially coarse. Upsampling these low-resolution

maps back to the original image size produces blurry, imprecise visualizations that cannot pinpoint the fine-grained changes relevant to diagnosis [18]. Second, quantifying how much each region contributes to the model's prediction is difficult. A prior study (HIVE) also reported that even when a model's prediction is wrong, anatomically plausible heatmaps can lead users to over-trust the AI's judgment [19]. In other words, a visually reasonable map may not reflect the model's true decision logic. Furthermore, cases have been reported where identical heatmaps are produced for both correct and incorrect predictions, raising further doubts about their reliability. While Grad-CAM is useful for identifying which regions a model attends to, it does not allow objective quantitative assessment of how much each region influences the prediction.

On the other hand, XAI methods based on feature selection have been used in other classification tasks to address this need for quantitative contribution assessment. Kent et al. built a machine learning model that predicts preference for window views from evaluation scores based on environmental criteria (place, time, weather, nature, and people), and quantitatively assessed the importance of each criterion [20]. This made it possible to express, in numerical terms rather than through subjective interpretation alone, how much each variable contributes to a model's prediction. In this way, approaches that quantify the contribution of each region or variable are useful for making a model's decision-making process more interpretable and for identifying the most influential variables, thereby providing deeper insight into the model's underlying complexity and decision-making mechanisms.

Against this background, we explore a novel approach based on the idea that, by resolving the spatial resolution limitations of existing visualization methods while explicitly identifying the contributing regions on the image and quantitatively evaluating their contribution, more effective interpretation would become possible. We draw on the concept of adversarial examples [21], originally developed to examine the vulnerabilities and robustness of neural networks. Systematically analyzing how a model behaves when small perturbations are applied to input images is a well-established method for evaluating model robustness [22,23]. This approach has recently been applied to medical image analysis to assess the reliability and stability of medical AI models [24,25]. In addition, in real-world clinical settings, unforeseen variations such as individual differences among imaging devices, noise, and partial loss of image data are unavoidable. In other engineering fields, such as control engineering, fault-tolerant methods aimed at maintaining system performance and ensuring reliability even in the presence of such disturbances or failures have been widely studied [26,27]. These methods aim to maintain stable performance under uncertain conditions through the design of control systems that anticipate failures or disturbances, as well as through adaptive compensation. In this way, a design philosophy that accounts for system robustness in advance also offers an important perspective for improving the reliability and stability of medical AI models in real-world operational environments.

In this study, we apply this robustness analysis framework as a means of evaluating what a model relies on for its predictions. Our hypothesis is that if applying a small perturbation to a specific region causes a significant drop in diagnostic accuracy, the model heavily relies on information from that region. Based on this idea, we use adversarial perturbations generated by the Fast Gradient Sign Method (FGSM) combined with binary spatial masks to control which region is perturbed, and measure how much each local region affects classification performance. This allows us to quantitatively assess each region's contribution to the model's decision-making. We focus on the anatomical structures most relevant to clinical glaucoma diagnosis: the optic disc and the macula. We apply circular masks centered on these structures and systematically vary the radius to perturb regions of varied sizes. We further divide the area surrounding the optic disc into sectors based on TSNIT (Temporal–Superior–Nasal–Inferior–Temporal) and clock-hour classifications and

evaluate the effect of perturbations applied to each sector with varying radii. Finally, we compare the regional influence on CNN decision-making with actual clinical diagnostic criteria, and clarify the similarities and differences between the two.

## 2. Materials and Methods

We conducted experiments to quantitatively compare which regions affect CNN classification performance in glaucoma diagnosis. The overall workflow, including model training, perturbation generation, and stepwise AUC (area under the curve) evaluation, is shown in Figure 1. This study used both whole images and cropped images. When using cropped images of the optic disc or macula, interference from other regions is eliminated, allowing a more direct assessment of how local noise affects model performance.

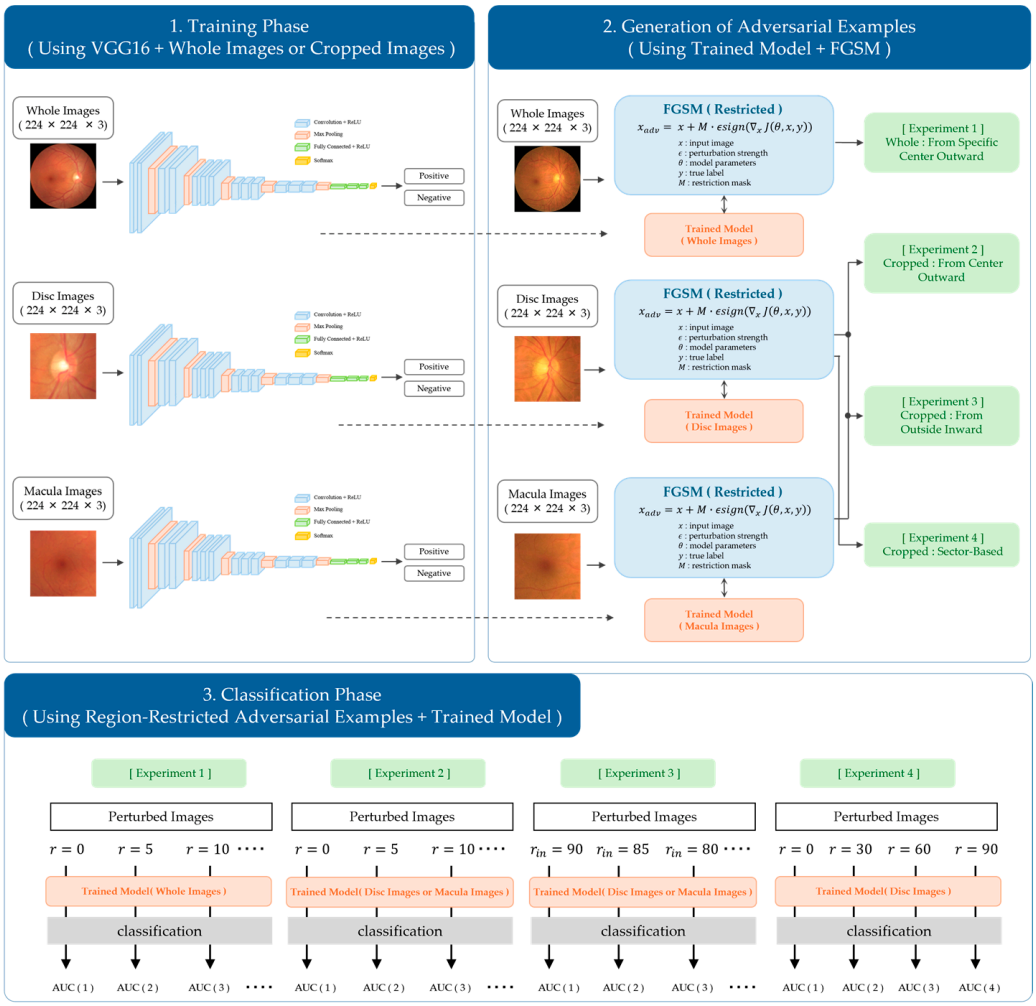



**Figure 1.** Overall experimental workflow.

### *2.1. Dataset, Preprocessing, and Extraction*

In this study, we used the Standardized Multi-Channel Dataset for Glaucoma (SMDG-19), Version 19 [28,29]. SMDG-19 is a large-scale collection that aggregates and standardizes publicly available fundus image datasets, organized into 19 partitions that correspond to 15 unique source datasets (several datasets are divided into separate training and test partitions). Each image is standardized by cropping the out-of-field region as much as possible, centering the fundus, padding missing regions, and resizing the image to 512 × 512 pixels while preserving the aspect ratio. We adopted this dataset because the public data are heterogeneous in their features—comprising a diverse set of parameters

(e.g., acquisition device, target population, and labeling criteria) suitable for training a prediction model—while being provided in a common format, thereby offering greater diversity than any single dataset could provide. The version used in this study (Version 19) contains 12,449 fundus images, each labeled as normal (label = 0), glaucoma (label = 1), or suspect (label = −1) (7499, 4817, and 133 images, respectively). The per-source image counts before and after preprocessing are summarized in Table 1.

**Table 1.** Per-source-dataset image counts before and after preprocessing. "Original" is the number of images SMDG-19 provides for each source; "Final" is the number remaining after all exclusions and preprocessing (described in Section 2.1). Sources with a Final count of 0 were entirely excluded by the narrow-field-of-view or quality/coordinate criteria.

| Source Dataset | Original Images | Final Images |
|---|---|---|
| OIA-ODIR | 4494 | 1421 |
| EyePACS | 3269 | 849 |
| REFUGE1 | 800 | 365 |
| sjchoi86-HRF | 401 | 292 |
| ORIGA | 650 | 278 |
| G1020 | 1020 | 274 |
| BEH | 634 | 153 |
| FIVES | 400 | 114 |
| HRF | 30 | 18 |
| JSIEC-1000 | 51 | 8 |
| DR-HAGIS | 10 | 1 |
| CRFO-v4 | 79 | 0 |
| PAPILA | 488 | 0 |
| DRISHTI-GS1 | 101 | 0 |
| LES-AV | 22 | 0 |
| total | 12,449 | 3773 |

These labels were inherited from the labels originally assigned by each constituent dataset and do not constitute a single ground truth created for SMDG-19 under a unified procedure. Across the 11 source datasets used in this study, the provenance of the labels is heterogeneous and includes: (i) labels based on clinical diagnoses that may draw on information beyond the fundus photograph (e.g., optical coherence tomography, visual fields, intraocular pressure, or clinical records); (ii) labels based on expert grading of fundus photographs or challenge-style adjudication; and (iii) labels from datasets originally intended for purposes other than glaucoma classification (e.g., vessel segmentation or multi-disease recognition), for which a centralized glaucoma-grading procedure is not described in the public documentation (Appendix A, Table A1). Accordingly, we do not claim that these labels constitute a ground truth established by a comprehensive ophthalmologist review of every case under a unified protocol; rather, we treat them as inheriting the heterogeneous reference standards of the constituent datasets.

First, images labeled as suspect (label = −1) were excluded. In addition, three of the 15 source datasets—PAPILA [30], DRISHTI-GS1 [31], and LES-AV [32]—were excluded. These datasets were captured with a narrow field of view of approximately 30 degrees centered on the optic disc, meaning the macula is not included in the frame. Since our analysis requires both the optic disc and macular regions, these datasets were not suitable. After these exclusions, 11,773 images remained.

Next, image quality was assessed using the quality prediction model of the fundus_image_toolbox (version 0.1.1) [33], which is an ensemble of multiple models and outputs a confidence score between 0 and 1 for each image, indicating image gradeability.

Rather than applying stringent quality filtering, we used the default threshold of confidence $\geq 0.5$ only to exclude images clearly unsuitable for analysis. This criterion retained 6344 images and excluded 5429.

We then automatically detected the center coordinates of the fovea and the optic disc in each image using the fovea–optic disc localization model (a multitask EfficientNet) of the same toolbox. This model was trained on publicly available datasets (ADAM, REFUGE, and IDRID) that include eyes with age-related macular degeneration, glaucoma, diabetic retinopathy, and healthy eyes. The developers reported a mean localization error for the fovea and optic disc of 0.88% of the image size, corresponding to a mean distance of 3.08 pixels for the 350 × 350 pixel images used for training and testing [33]. However, this value was derived from the model's own training and evaluation data, and localization errors may be larger for our dataset, which differs in acquisition conditions. So, we visually inspected every image with the detected centers overlaid and excluded those in which a detected center clearly deviated from the anatomical position of the optic disc or fovea (20 images). At this stage, 6324 images remained.

Using the detected coordinates, we also unified all images to a right-eye orientation (optic disc on the right, macula on the left) based on the relative positions of the fovea and optic disc. Horizontal flipping was applied to left-eye images, and center coordinates were transformed accordingly.

From these standardized 512 × 512 images, we prepared three types of input image, each 224 × 224 pixels:

- Whole image: the original 512 × 512 image resized to 224 × 224 pixels;
- Optic disc crop: a 224 × 224 pixel region cropped from the 512 × 512 image, centered on the optic disc;
- Macula crop: a 224 × 224 pixel region cropped from the 512 × 512 image, centered on the fovea.

Local patches for the optic disc and macula were generated using the coordinates obtained in Section 2.1. The patch size of 224 × 224 pixels matches the input specification of the pretrained VGG16 [34]. During cropping, images were excluded when the center coordinate of the fovea or the optic disc lay within 112 pixels (half the patch size) of the image border, because the resulting patch would extend beyond the fundus image boundary. This criterion was applied to both the fovea and the optic disc, and images that failed the criterion for either center were excluded, so that both regions could be compared fairly with the same image. This exclusion removed 2551 images, leaving 3773 images in the final dataset.

### *2.2. Baseline Models Training*

#### 2.2.1. Network and Training Configuration

We used VGG16 [34], a standard CNN architecture pretrained on ImageNet, for binary glaucoma classification (positive or negative). To adapt the architecture to this two-class task, the final fully connected layer was replaced with a two-output layer followed by a softmax activation function.

Three models were trained independently, each taking a different type of input image:

- Model 1: whole images;
- Model 2: cropped optic disc images;
- Model 3: cropped macula images.

All models used the same hyperparameters: an SGD optimizer with momentum 0.85 and weight decay 0.005, 200 epochs, a learning rate of 0.0003, and a batch size of 512. The model state with the lowest validation loss was retained for evaluation.

All three models were trained using a class-weighted cross-entropy loss. Because the training set was imbalanced between the normal and glaucoma classes, the class weights were computed from the training split in proportion to the inverse class frequencies, thereby increasing the contribution of the minority glaucoma class to the objective function. All models were implemented in Python 3.11.5 (Python Software Foundation, Beaverton, OR, USA) using PyTorch 2.1.1 and torchvision 0.16.1 with CUDA 12.1 (Meta AI, Menlo Park, CA, USA). Training and evaluation were performed on an NVIDIA RTX A4500 GPU (NVIDIA Corporation, Santa Clara, CA, USA) under Ubuntu Linux. Evaluation metrics, including ROC-AUC and bootstrap confidence intervals, were computed using scikit-learn 1.2.2, NumPy 1.26.2, and SciPy 1.11.4.

#### 2.2.2. Evaluation Regimes

Because SMDG-19 aggregates multiple public datasets, the way the data are partitioned can introduce leakage (patient- or eye-level overlap and near-duplicate images) and source confounding, both of which can inflate apparent performance. We therefore evaluated the models under two complementary regimes.

(a) Internal evaluation (leak-reduced split);

All sources were pooled, stratified by class, and split image-wise into training, validation, and test sets at a ratio of 7:1.5:1.5 (with a fixed random seed), yielding 2641 training images (822 positive/1819 negative), 566 validation images (176 positive/390 negative), and 566 test images (176 positive/390 negative). Where metadata permitted, images from the same patient were kept within a single split; for OIA-ODIR, images were grouped by patient using the left/right-eye information encoded in the file names. Because patient identifiers could not be recovered for most sources, we disclose the possibility of residual leakage as a limitation. This regime does not structurally eliminate leakage, but its discriminative signal is strong; we therefore use it as the substrate for analyzing what each model attends to and report its results as an in-distribution, best-case performance ceiling. The same split was shared across all three input conditions so that model comparisons were made on identical samples.

(b) External evaluation (leave-one-dataset-out, LODO);

One source dataset was held out entirely for testing while the model was trained on all remaining sources; this procedure was repeated for each eligible source in turn. Because the held-out source is never seen during training, leakage and source confounding are structurally removed regardless of whether patient identifiers are available. A source was deemed eligible for the held-out test set if it contained both classes and comprised at least 100 images; six sources met this criterion (REFUGE1, sjchoi86-HRF, ORIGA, G1020, BEH, and FIVES). Single-class or small sources (EyePACS-AIROGS, OIA-ODIR, HRF, JSIEC-1000, and DR-HAGIS) were always assigned to the training set. This regime measures generalization to datasets not seen during training.

The primary evaluation metric was the area under the receiver operating characteristic curve (ROC-AUC). Because the dataset is class-imbalanced (31.4% glaucoma), AUC is threshold-independent and robust to class imbalance, whereas accuracy can be misleading under such conditions. To assess the reliability of the AUC estimates, 95% confidence intervals were additionally computed using bootstrap resampling (1000 iterations). In addition, the false-negative rate (FNR), defined as the proportion of glaucomatous cases misclassified as non-glaucomatous, was reported.

#### 2.2.3. Baseline Classification Performance

Under internal evaluation (leak-reduced; test n = 566), the test AUCs of Models 1, 2, and 3 were 0.913 (whole), 0.894 (optic disc), and 0.838 (macula), respectively (Table 2).

Under external (LODO) evaluation, the macro-averaged AUC across held-out sources fell to 0.685 (whole), 0.712 (optic disc), and 0.562 (macula), with substantial variation across sources (Table 3).

**Table 2.** Internal-test classification performance (leak-reduced split; test n = 566).

| Metric | Whole | Optic Disc | Macula |
|---|---|---|---|
| AUC (95% CI) | 0.913 (0.885–0.937) | 0.894 (0.864–0.922) | 0.838 (0.802–0.870) |
| Accuracy | 0.853 | 0.834 | 0.763 |
| F1 (positive) | 0.782 | 0.755 | 0.663 |
| FNR | 0.153 | 0.176 | 0.250 |

**Table 3.** External (leave-one-dataset-out) classification performance. AUC (95% CI) for each held-out source and the macro mean.

| Hold-Out Source | Whole | Optic Disc | Macula |
|---|---|---|---|
| BEH | 0.648 (0.548–0.744) | 0.674 (0.573–0.775) | 0.672 (0.579–0.763) |
| FIVES | 0.719 (0.608–0.825) | 0.712 (0.593–0.820) | 0.591 (0.471–0.714) |
| G1020 | 0.538 (0.462–0.619) | 0.557 (0.483–0.634) | 0.446 (0.371–0.519) |
| ORIGA | 0.686 (0.598–0.763) | 0.763 (0.694–0.828) | 0.544 (0.461–0.624) |
| REFUGE1 | 0.855 (0.787–0.917) | 0.786 (0.685–0.876) | 0.597 (0.506–0.677) |
| sjchoi86-HRF | 0.666 (0.590–0.736) | 0.782 (0.715–0.842) | 0.524 (0.437–0.607) |
| Macro mean | 0.685 (0.538–0.855) | 0.712 (0.557–0.786) | 0.562 (0.446–0.672) |

The gap between internal and LODO performance is pronounced (whole 0.913 → 0.685; optic disc 0.894 → 0.712; macula 0.838 → 0.562), indicating that the high internal AUCs are optimistic estimates inflated by source confounding and residual leakage. The LODO level of approximately 0.7 is consistent with the cross-site generalization performance reported for glaucoma CNNs evaluated on the same SMDG collection without domain adaptation, where per-site AUCs span roughly 0.55–0.73 (mean ≈ 0.61–0.67) [35], and with the broader literature reporting substantial performance drops under external validation [36]. Therefore, the LODO performance observed in this dataset can be considered a realistic level consistent with previous findings. However, although the magnitude of performance degradation remains an important evaluation metric in this study, the obtained performance cannot yet be regarded as sufficiently high. Therefore, the principal aim of this study is to investigate the regional dependence of CNN predictions by observing region dependence in internally evaluated models and examining whether these tendencies generalize under LODO while considering the possibility that dataset-specific effects may still remain. In other words, this study approaches the problem from both internal and external (LODO) evaluation perspectives.

### *2.3. Region-Restricted Adversarial Examples*

#### 2.3.1. Perturbation Generation

We used a locally adapted version of the Fast Gradient Sign Method (FGSM) [21] to generate adversarial noise. Adversarial examples are images in which imperceptibly small perturbations are added to the input, causing a deep learning model to misclassify them. We exploit this property to identify regions where the model is most vulnerable and use this to infer each region's contribution to classification.

Let $x$ be the input image, $y$ its ground truth label, and $\theta$ the trained model parameters. We define a binary mask $M \in \{0, 1\}^{H \times W}$ to control the region where perturbation is applied. The locally perturbed adversarial image $x_{adv}$ is then defined as:

$$x_{adv} = x + M \cdot (\epsilon \cdot sign\ (\nabla_x\ J(\theta,\ x,\ y))) \quad (1)$$

The hyperparameter $\epsilon$ was set to 0.006 for Experiments 1–3 and 0.009 for Experiment 4. Perturbations were applied to the test set of 566 images.

#### 2.3.2. Mask Design for Each Experiment

We define four experimental configurations using mask $M$. Let $(i, j)$ denote any pixel coordinate in the image. The conditions for the perturbed region are defined as follows.

Experiment 1: Stepwise expansion of a circular region in whole images.

This experiment targets whole images. Mask $M$ defines a circular region centered on the optic disc center coordinates $(i_c, j_c)$ obtained in Section 2.1. The radius $r$ is increased from 0 to 90 pixels in steps of 5 pixels. The mask element $M(i, j)$ is defined as:

$$M(i,j) = \begin{cases} 1 & if \sqrt{(i-i_c)^2 + (j-j_c)^2} \leq r \\ 0 & otherwise \end{cases} \quad (2)$$

Using this mask $M$ and Model 1, locally perturbed adversarial examples were generated according to Equation (1). The change in the perturbed region as the radius increases is shown in Figure 2.

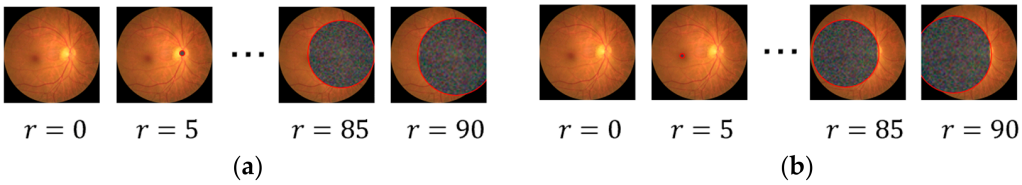



**Figure 2.** Examples of perturbed regions in Experiment 1. (**a**) Circular noise region centered on the optic disc, expanded in steps of 5 pixels; (**b**) Circular noise region centered on the macula, expanded in steps of 5 pixels.

Experiment 2: Stepwise expansion of a concentric circular region in cropped images.

This experiment targets cropped optic disc and macula images. Mask $M$ defines a circular region centered on the center of each cropped image $(i_c, j_c)$. The radius $r$ is increased from 0 to 90 pixels in steps of 5 pixels, with the condition defined as:

$$M(i,j) = \begin{cases} 1 & if \sqrt{(i-i_c)^2 + (j-j_c)^2} \leq r \\ 0 & otherwise \end{cases} \quad (3)$$

Using this mask $M$ and Models 2 and 3, locally perturbed adversarial examples were generated according to Equation (1). The change in the perturbed region is shown in Figure 3.

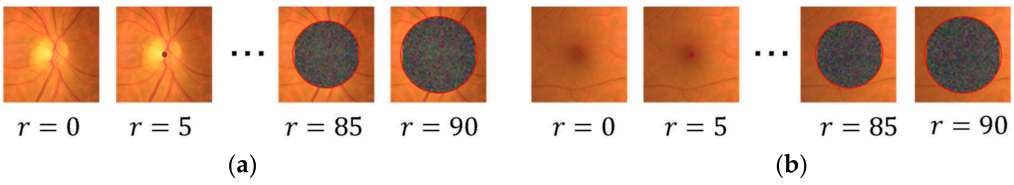



**Figure 3.** Examples of perturbed regions in Experiment 2. (**a**) Circular noise region centered on the optic disc, expanded in steps of 5 pixels; (**b**) Circular noise region centered on the macula, expanded in steps of 5 pixels.

Experiment 3: Expansion of an annular region in cropped images.

This experiment targets cropped optic disc and macula images. Mask $M$ defines an annular (donut-shaped) region centered on the image center $(i_c, j_c)$. The outer boundary is fixed at 90 pixels, and the inner radius $r_{in}$ is decreased from 90 to 0 pixels in steps of

5 pixels, progressively expanding the perturbed region from the periphery toward the center. The mask element is defined as:

$$M(i,j) = \begin{cases} 1 & if\ r_{in} < \sqrt{(i-i_c)^2 + (j-j_c)^2} \leq 90 \\ 0 & otherwise \end{cases} \quad (4)$$

Using this mask $M$ and Models 2 and 3, locally perturbed adversarial examples were generated according to Equation (1). The change in the annular region is shown in Figure 4.

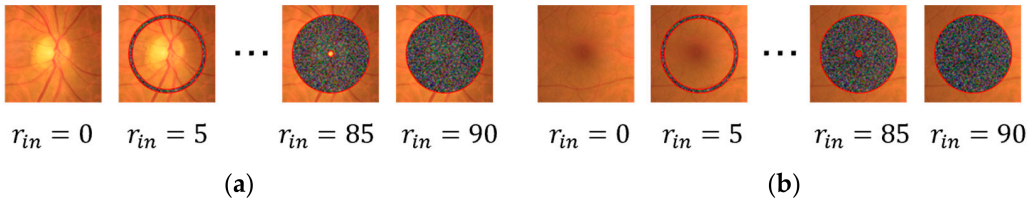



**Figure 4.** Examples of perturbed regions in Experiment 3. (**a**) Annular noise region centered on the optic disc, with inner radius decreased in steps of 5 pixels; (**b**) Annular noise region centered on the macula, with inner radius decreased in steps of 5 pixels.

Experiment 4: Stepwise radius increase in directional sectors of cropped images.

This experiment targets cropped optic disc images. Mask $M$ defines directional sectors based on TSNIT (4 sectors: Temporal, Superior, Nasal, Inferior) and Clock-hour (12 sectors) classifications, both of which are standard references in clinical glaucoma assessment.

TSNIT analysis evaluates retinal nerve fiber layer (RNFL) thickness around the optic disc and is a standard tool in glaucoma diagnosis. Based on its definition, the optic disc region is divided into four sectors: Temporal (T), Superior (S), Nasal (N), and Inferior (I). Clock-hour analysis divides the region around the optic disc into 12 sectors like a clock face and is used to assess structural changes in each direction. Both are important for detecting the superior and inferior nerve fiber damage characteristic of glaucoma. Sectors were mapped to the clock directions on the fundus image, numbered 1–12 clockwise starting from 12 o'clock.

For each TSNIT and Clock-hour sector, a sector-shaped perturbed region was defined, and the radius was increased stepwise while AUC was evaluated. The radius $r$ was increased from 0 to 90 pixels in steps of 30 pixels. For a sector with angular boundaries $[\varphi_{start}, \varphi_{end}]$, the mask element is defined as:

$$M(i,j) = \begin{cases} 1 & if\ \sqrt{(i-i_c)^2 + (j-j_c)^2} \leq r\ and\ \varphi_{start} \leq atan2(j-j_c, i-i_c) < \varphi_{end} \\ 0 & otherwise \end{cases} \quad (5)$$

Using this mask $M$ and Model 2, locally perturbed adversarial examples were generated according to Equation (1). The perturbed regions for each sector and radius are shown in Figure 5.

### *2.4. Classification Performance Evaluation*

Model classification performance was evaluated using the locally perturbed images generated in each experimental condition (Experiments 1–4). Model 1 was used for Experiment 1, Model 2 for Experiment 4, and for Experiments 2 and 3, Model 2 was used for optic disc crops and Model 3 for macula crops. For each experiment, perturbed images generated at each step of the control parameter $r$ (radius of the circular or sector region, or inner radius $r_{in}$ for the annular region) were fed into the corresponding model and classification performance was measured.

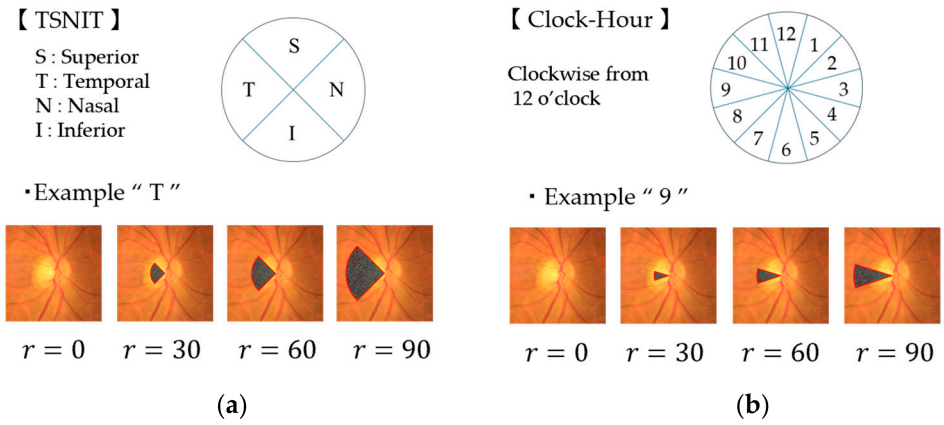



**Figure 5.** Sector definitions and examples of perturbed regions in Experiment 4. (**a**) TSNIT 4-sector division with radius expanded in steps of 30 pixels (Temporal sector shown as example); (**b**) Clock-hour 12-sector division with radius expanded in steps of 30 pixels (9 o'clock sector shown as example).

ROC-AUC, its bootstrap-based 95% confidence interval, and the false-negative rate (FNR) were used as the evaluation metrics for each condition (see Section 2.2.2 for details). AUC was computed at each step as $r$ was varied. The degree of performance drop as the perturbed area expanded was used to quantitatively compare each region's contribution to model predictions. Where necessary, differences in AUC between conditions were assessed using a permutation test (10,000 iterations) to obtain $p$-values, and statistical significance was adjusted to $p < 0.05$ after multiple comparison correction using the Holm–Bonferroni method.

### *2.5. Verification of Evidence Regions*

Since the performance degradation associated with localized adversarial perturbation (FGSM) could potentially be attributed not to genuine reliance on clinically relevant regions but rather to model-specific gradient sensitivity or local texture artifacts, occlusion sensitivity analysis was additionally performed as a complementary explainable AI (XAI) method.

In this analysis, a mask of fixed patch size (32 × 32 pixels) was slid across the input image with a stride of 16 pixels to locally occlude regions, and the degree of decrease in the model's predicted probability (Drop) at each position was calculated. This allowed the importance of each region to the model's prediction to be quantified.

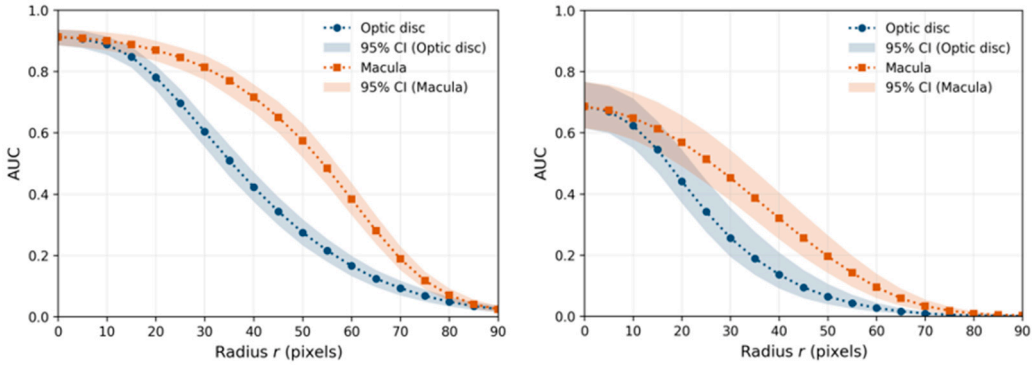
When aggregating and comparing the resulting occlusion heatmaps across the entire dataset, spatial alignment was performed to eliminate the influence of inter-image anatomical variability (i.e., differences in position and orientation). For the whole-field image (Whole image) model, the center coordinates of the optic disc and macula were detected in each image, and a rigid-body (affine) transformation—incorporating rotation and scale correction—was applied so that these positional relationships matched the mean coordinates across the entire dataset, thereby achieving alignment. For the optic disc and macula models, in which the region of interest had already been cropped, images were superimposed directly based on the crop center.

From these aligned heatmaps, statistical maps representing the mean (Mean map) values across the entire dataset were computed. This approach smooths out incidental responses to specific local noise, allowing the regions on which the model universally relies for its predictions to be identified and verified.

## 3. Results

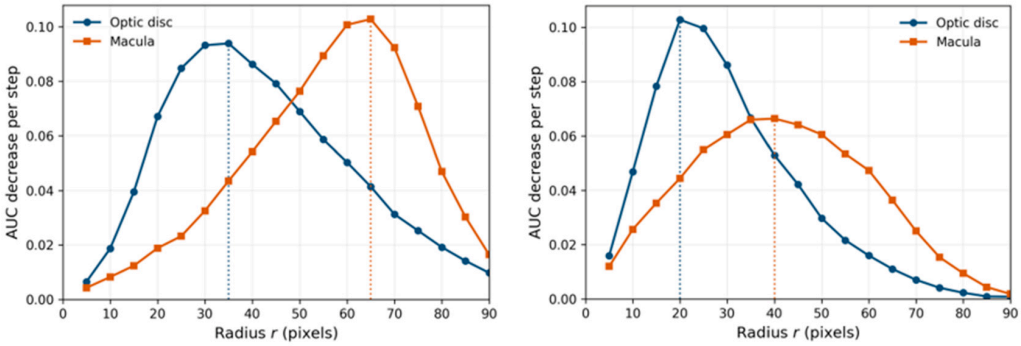
### *3.1. Regional Noise Impact on Whole Fundus Images: Experiment 1*

Experiment 1 examined how applying noise to the optic disc and macular regions affects classification performance in whole fundus images. AUC as a function of the noise radius r for each region is shown in Figure 6, the change in AUC between consecutive steps is shown in Figure 7, and the false negative rate (FNR) is shown in Figure 8. In each figure, panel (a) shows results using the leak-reduced dataset, and panel (b) shows results using the LODO dataset.




(**a**) (**b**)

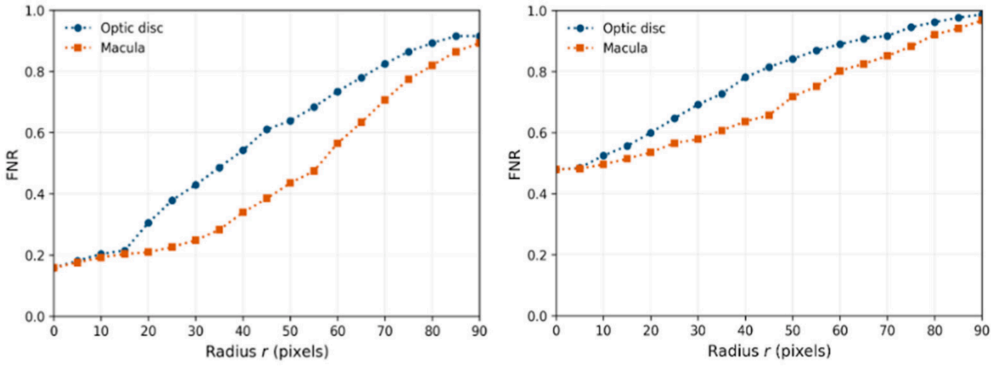
**Figure 6.** AUC as a function of the local noise radius r. Changes in AUC scores associated with noise applied to the optic disc and macular regions. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.




(**a**) (**b**)

**Figure 7.** Change in classification performance (AUC) between consecutive steps in whole fundus images. Difference in AUC scores (degree of performance degradation) as the perturbed area was expanded from the previous step's radius. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

Although the two datasets showed slight differences in the specific magnitude of performance degradation and in the timing at which the largest change occurred, the overall behavior of performance degradation followed a similar trend across both datasets. Furthermore, statistically significant performance degradation from baseline ($r = 0$) was confirmed for all noise conditions ($r > 0$) ($p < 0.05$). In addition, when directly comparing AUC between the optic disc and macular regions at the same noise radius r, statistically significant differences were observed at nearly all radii in both datasets, with the exception of the baseline ($r = 0$) ($p < 0.05$).




(**a**) (**b**)

**Figure 8.** False negative rate (FNR) as a function of the local noise radius r. Increase in the miss rate (false negative rate) associated with noise applied to the optic disc and macular regions. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

The pattern of performance degradation differed depending on the region perturbed. When noise was applied to the optic disc region, AUC began to drop sharply from the early stages ($r \geq 20$ for the leak-reduced dataset and $r \geq 10$ for the LODO dataset). Looking at the magnitude of performance change (Figure 7), the largest drop occurred in the earlier stages (around $r = 35$ for the leak-reduced dataset and around $r = 20$ for the LODO dataset), suggesting that the model's basis for prediction was substantially disrupted at this stage.

In contrast, perturbation centered on the macula produced only a slight drop in the initial stages, indicating greater robustness to noise compared with the optic disc region. Regarding the step-wise change in performance, a slight difference was observed between the two datasets: in the leak-reduced dataset, a sharp fluctuation was also observed in the macular region during the middle-to-later stages, whereas in the LODO dataset such abrupt changes were less pronounced, with the magnitude of change remaining relatively stable. Nevertheless, in both datasets, performance gradually declined as the radius increased, and a marked local drop in accuracy was observed at later stages ($r = 65$ for the leak-reduced dataset and $r = 40$ for the LODO dataset).

With respect to the false negative rate (FNR), both regions showed an increasing trend as the perturbed area expanded. Notably, in the intermediate range of radius, applying noise to the optic disc region resulted in a higher FNR than applying noise to the macular region, supporting the possibility that the model preferentially relies on information near the optic disc for glaucoma classification.

### *3.2. Central Circular Noise in Cropped Images: Experiment 2*

Experiment 2 used cropped images to isolate specific regions and examined how applying noise to local structural information affects classification performance. AUC as a function of the central circular noise radius r for each cropped image (optic disc and macula) is shown in Figure 9, the change in AUC between consecutive steps is shown in Figure 10, and the false negative rate (FNR) is shown in Figure 11. In each figure, panel (a) shows results using the leak-reduced dataset, and panel (b) shows results using the LODO dataset.

In the initial stages of noise application (radius up to 20 pixels), comparison between regions in both datasets showed that the optic disc model showed almost no performance degradation, whereas the macula model already showed a decline from the start. Subsequently, a substantial decline in performance also began in the optic disc region, with the

step-wise change in performance reaching its maximum during the middle stage (around r = 35 for the leak-reduced dataset and around r = 30 for the LODO dataset).

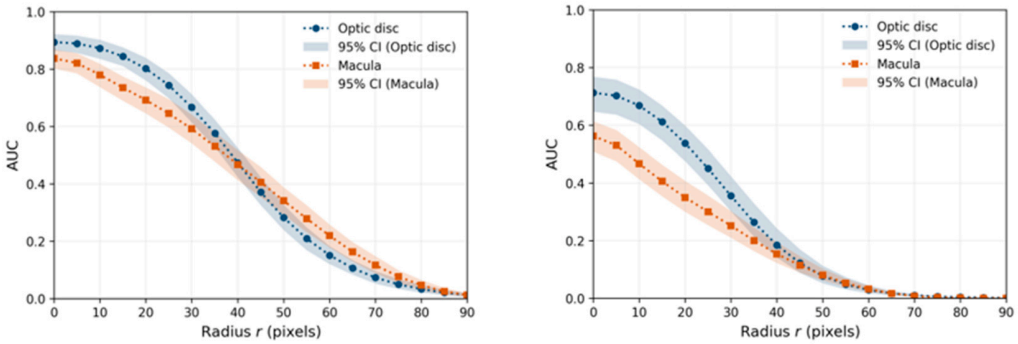



**Figure 9.** AUC as a function of the central circular noise radius r in cropped images. Changes in classification performance (AUC) associated with the noise radius r for the optic disc and macula crops. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

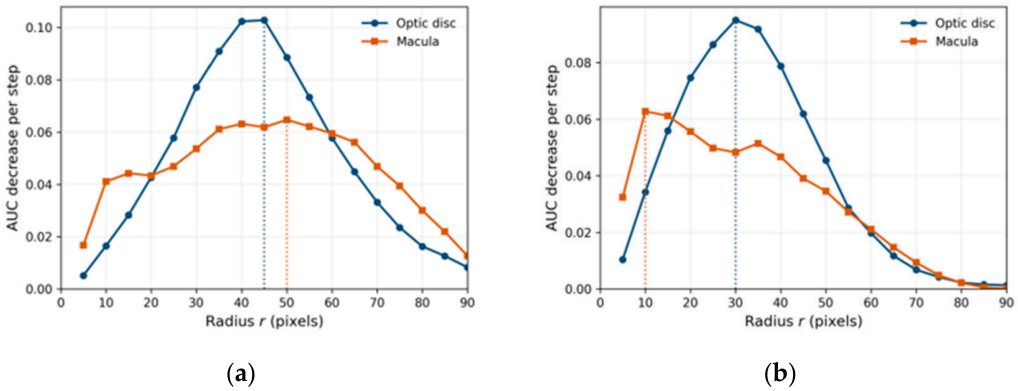



**Figure 10.** Change in classification performance (AUC) between consecutive steps in cropped images. Difference in AUC scores (degree of performance degradation) as the perturbed area was expanded from the previous step's radius. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

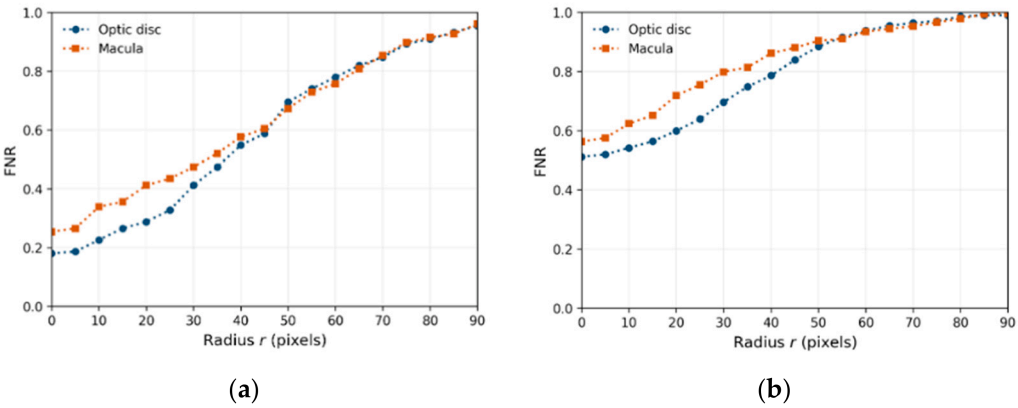



**Figure 11.** False negative rate (FNR) as a function of the central circular noise radius *r* in cropped images. Increase in the miss rate (false negative rate) associated with the noise radius *r* for the optic disc and macula crops. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

Comparing the degradation trends of the optic disc and macula regions across both datasets, similar behavior was observed up until approximately r = 40. Indeed, when directly comparing AUC between the two regions at the same radius, statistically significant

differences were consistently observed in both datasets from the initial stages up to the point of performance convergence (r < 40) ($p < 0.05$).

Beyond a radius of approximately 30 pixels, the optic disc model began to show a marked drop in AUC, with the largest step-to-step change occurring at a radius of 35–40 pixels (varying slightly between datasets). Notably, around 40–45 pixels the AUC curves of the two models crossed, with the optic disc model falling below the macula model. However, differences between the datasets emerged in the behavior observed after this crossover point (around r = 40–45). In the leak-reduced dataset, statistically significant differences between the two regions were again observed in the middle-to-later stages following the crossover (r = 45–75), whereas in the LODO dataset, no statistically significant differences were found at any radius beyond r = 40. The macula model, by contrast, showed no sharp drop at any radius, instead declining at a steady rate as the noise area expanded.

With respect to the false negative rate (FNR), a consistent increasing trend was observed with the expansion of the perturbed area in both regions and both datasets, confirming that the model's discriminative ability was progressively degraded.

### *3.3. Annular Noise in Cropped Images: Experiment 3*

Experiment 3 applied annular noise around the optic disc and macula to evaluate how peripheral anatomical structures affect classification performance. AUC as a function of the inner radius $r_{in}$ of the annular noise region (as the noise area was expanded from the outer periphery toward the center) for each cropped image is shown in Figure 12, the change in AUC between consecutive steps is shown in Figure 13, and the false negative rate (FNR) is shown in Figure 14. In each figure, panel (a) shows results using the leak-reduced dataset, and panel (b) shows results using the LODO dataset.

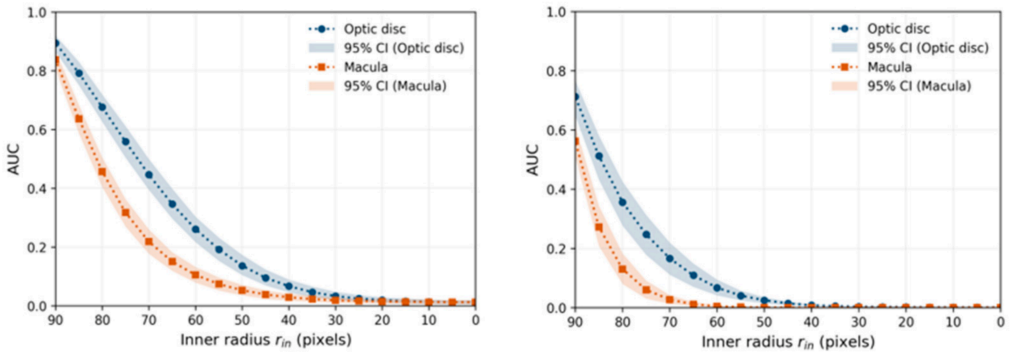
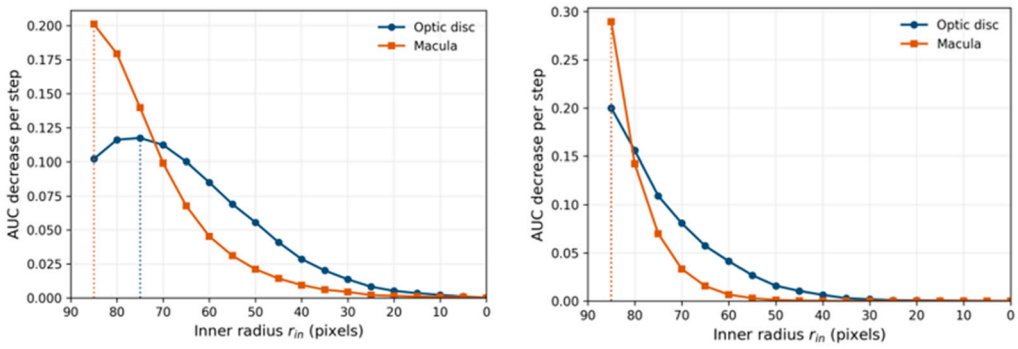



**Figure 12.** AUC as a function of the annular noise inner radius $r_{in}$ in cropped images. Changes in classification performance (AUC) associated with the annular noise inner radius $r_{in}$ for the optic disc and macula crops. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

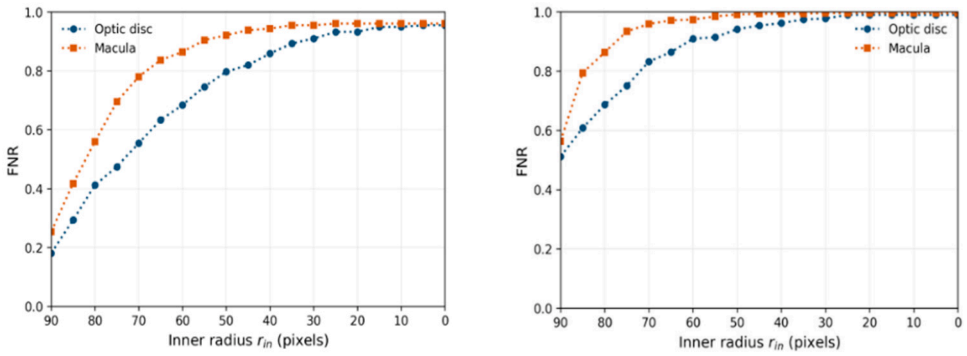
Both datasets showed a similar trend in the pattern of AUC degradation. For both the optic disc and macula, a notable drop in AUC was observed when noise was confined to the outer periphery (large $r_{in}$ values). Direct comparison of AUC between the optic disc and macula at the same inner radius $r_{in}$ revealed that the macula consistently showed a sharper and larger performance drop throughout the noise range.

The macula model showed a particularly large performance drop compared to the optic disc model. When $r_{in}$ was large, meaning noise was applied only to the periphery, the macula model showed a sharp fluctuation in AUC, with the largest single-step change occurring when $r_{in}$ decreased from 90 to 85 (i.e., immediately after noise was applied to the outermost band).




(**a**) (**b**)

**Figure 13.** Change in classification performance (AUC) between consecutive steps. Difference in AUC scores (degree of performance degradation) as the perturbed area was expanded inward from the previous step's inner radius. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.




(**a**) (**b**)

**Figure 14.** False negative rate (FNR) as a function of the annular noise inner radius $r_{in}$ in cropped images. Increase in the miss rate (false negative rate) associated with the annular noise inner radius $r_{in}$ for the optic disc and macula crops. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

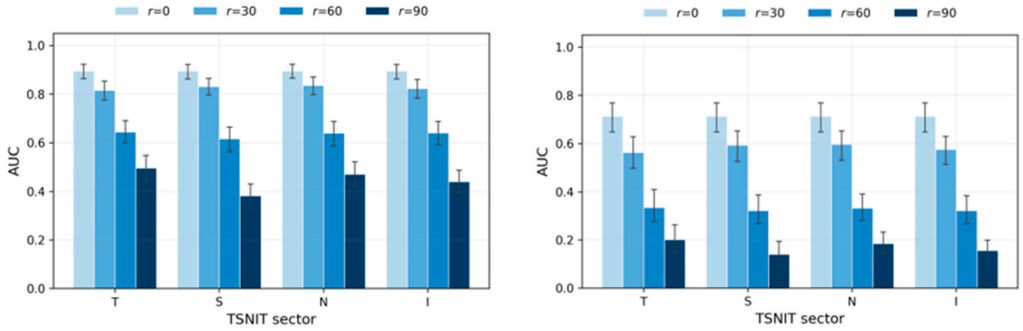
As $r_{in}$ decreased and the noise region approached the center (the fovea or the optic disc center), the rate of AUC decline gradually slowed for both regions, eventually converging to a low value. Along with this convergence in behavior, statistically significant differences between the two regions were observed from the initial through the middle stages ($p < 0.05$); however, at the final stage ($r_{in} < 25$ for the leak-reduced dataset and $r_{in} < 35$ for the LODO dataset), these differences were no longer significant ($p \geq 0.05$), indicating that the performance gap between the two regions became comparable.

With respect to the false negative rate (FNR) shown in Figure 14, both regions showed an increasing trend as the noise area expanded (i.e., as $r_{in}$ decreased). This increase was relatively rapid during the initial stage of noise application to the peripheral annular region ($r_{in}$ = 90 to approximately 70), after which the curve gradually converged toward a high value. Notably, up through the middle stage, the macula consistently showed higher FNR values than the optic disc, indicating that perturbation of the outermost periphery elevated misclassification in the macula model at an earlier stage.

### *3.4. TSNIT and Clock-Hour Sector Analysis: Experiment 4*

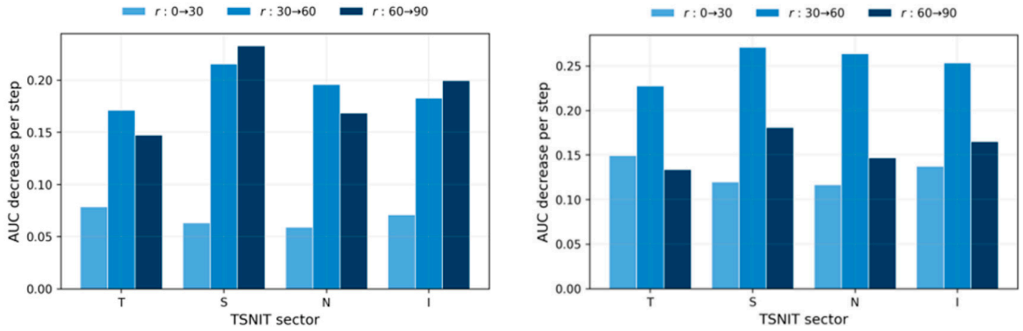
Experiment 4 divided the noise region according to TSNIT (4 sectors) and Clock-hour (12 sectors) classifications used in clinical diagnosis, and evaluated model vulnerability for each sector. AUC as a function of the noise radius r for each sector (S: Superior, I: Inferior,

N: Nasal, T: Temporal) is shown in Figure 15, the change in AUC between consecutive steps is shown in Figure 16, and the false negative rate (FNR) is shown in Figure 17. In each figure, panel (a) shows results using the leak-reduced dataset, and panel (b) shows results using the LODO dataset.



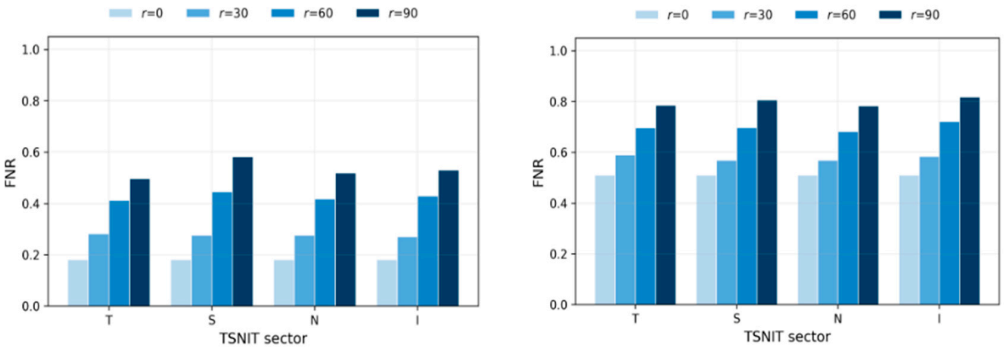

(**a**) (**b**)

**Figure 15.** AUC as a function of the noise radius *r* for each TSNIT sector in cropped optic disc images. Changes in classification performance (AUC) associated with the noise radius *r* for each TSNIT sector. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.




(**a**) (**b**)

**Figure 16.** Change in classification performance (AUC) between consecutive steps. Difference in AUC scores (degree of performance degradation) as the noise region for each sector was expanded from the previous step's radius. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.




(**a**) (**b**)

**Figure 17.** False negative rate (FNR) as a function of the noise radius *r* for each TSNIT sector in cropped optic disc images. Increase in the miss rate (false negative rate) associated with the noise radius *r* for each TSNIT sector. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

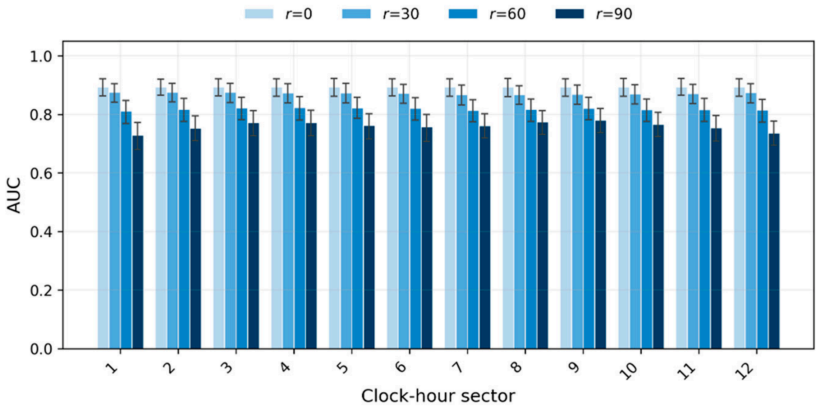
The TSNIT results showed clear differences in performance degradation depending on the direction of noise. In particular, comparing results at the maximum radius (r = 90 pixels),

the Superior (S) sector showed the most pronounced AUC drop in both datasets, followed by moderate drops in the Inferior (I) and Nasal (N) sectors. The Temporal (T) sector showed the smallest AUC drop and relatively maintained classification performance. Indeed, when comparing AUC across sectors at r = 90 pixels, statistically significant differences were observed among all sectors in both datasets ($p < 0.05$).

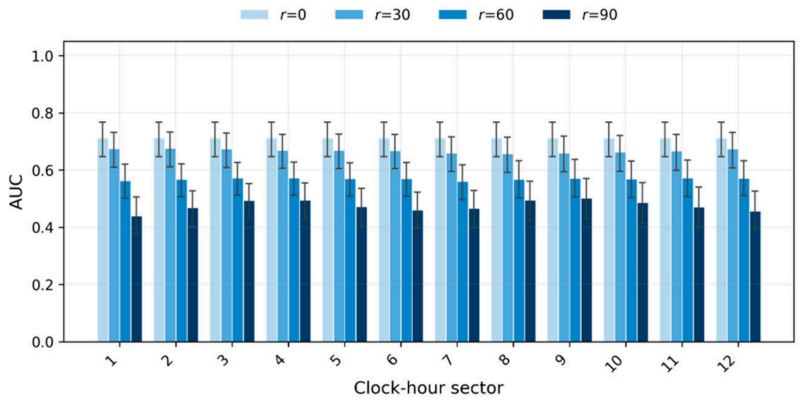
Focusing on the step-wise change in performance, the leak-reduced dataset showed large drops in both the r = 30–60 and r = 60–90 intervals, whereas the LODO dataset showed the largest change in the r = 30–60 interval. Taken together, these results across both datasets indicate that perturbation in the r = 30–60 range is particularly disruptive to the model's decision-making.

The false negative rate (FNR) also showed a consistent increasing trend across all TSNIT sectors as the noise radius expanded.

AUC as a function of the noise radius r for each sector (1–12 o'clock) is shown in Figure 18, the change in AUC between consecutive steps is shown in Figure 19, and the false negative rate (FNR) is shown in Figure 20. In each figure, panel (a) shows results using the leak-reduced dataset, and panel (b) shows results using the LODO dataset.

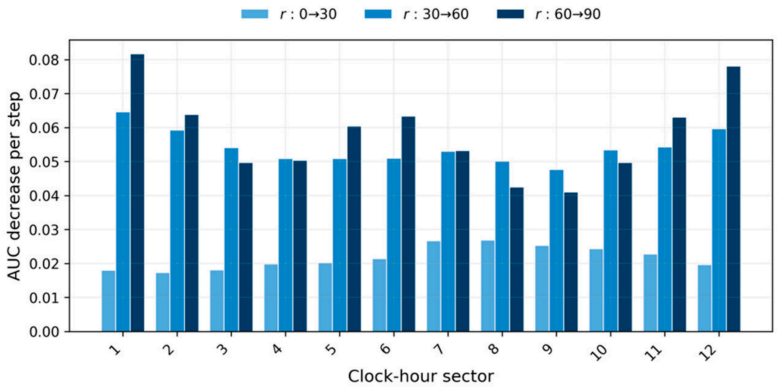



(**a**)

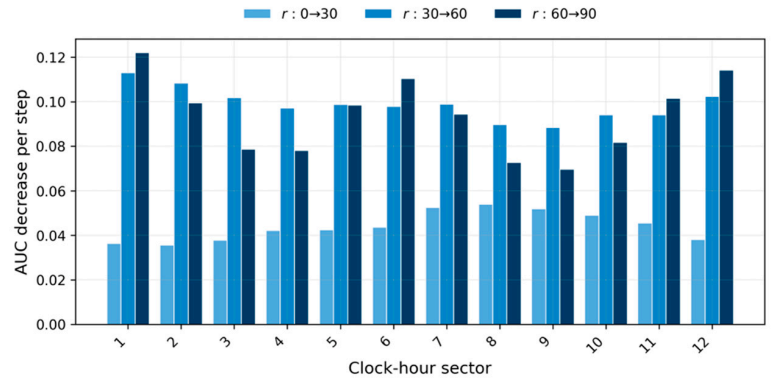



(**b**)

**Figure 18.** AUC as a function of the noise radius *r* for each Clock-hour sector in cropped optic disc images. Changes in classification performance (AUC) associated with the noise radius *r* for each Clock-hour sector (1–12 o'clock). (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.




(**a**)




(**b**)

**Figure 19.** Change in classification performance (AUC) between consecutive steps. Difference in AUC scores (degree of performance degradation) as the noise region for each sector was expanded from the previous step's radius. (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

Consistent with the TSNIT results, the Clock-hour analysis also showed larger AUC changes in the peripheral range (30–90 pixels) than in the central range (0–30 pixels) across all sectors in both datasets.

Differences in the degree of performance degradation were also observed between sectors, and this pattern was consistent across both datasets. The most pronounced performance drops at the maximum radius (r = 90 pixels) were in the superior region, particularly at the 1 o'clock direction (leak-reduced: AUC = 0.7296, LODO: AUC = 0.4408) and 12 o'clock (leak-reduced: AUC = 0.7365, LODO: AUC = 0.4574), followed by 2 o'clock (leak-reduced: AUC = 0.7534, LODO: AUC = 0.4686) and 11 o'clock (leak-reduced: AUC = 0.7538, LODO: AUC = 0.4712). Comparing AUC among these sectors, no statistically significant difference was found between the 2 o'clock and 11 o'clock directions in either dataset, whereas statistically significant differences were observed between all other sector pairs ($p < 0.05$).

The smallest drops were seen in the horizontal directions, including 9 o'clock (leak-reduced: AUC = 0.7799, LODO: AUC = 0.5023), 8 o'clock (leak-reduced: AUC = 0.7744, LODO: AUC = 0.4956), 4 o'clock (leak-reduced: AUC = 0.7728, LODO: AUC = 0.4948), and 3 o'clock (leak-reduced: AUC = 0.7721, LODO: AUC = 0.4938), with the 9 o'clock (temporal) direction being the least affected. Comparing AUC among these sectors, no statistically significant differences were found between the 8 o'clock and 4 o'clock pair, or between the 4 o'clock and 3 o'clock pair, in either dataset, whereas statistically significant differences were observed between all other sector pairs ($p < 0.05$).

The false negative rate (FNR) also showed a consistent increasing trend across all Clock-hour sectors as the noise radius expanded.

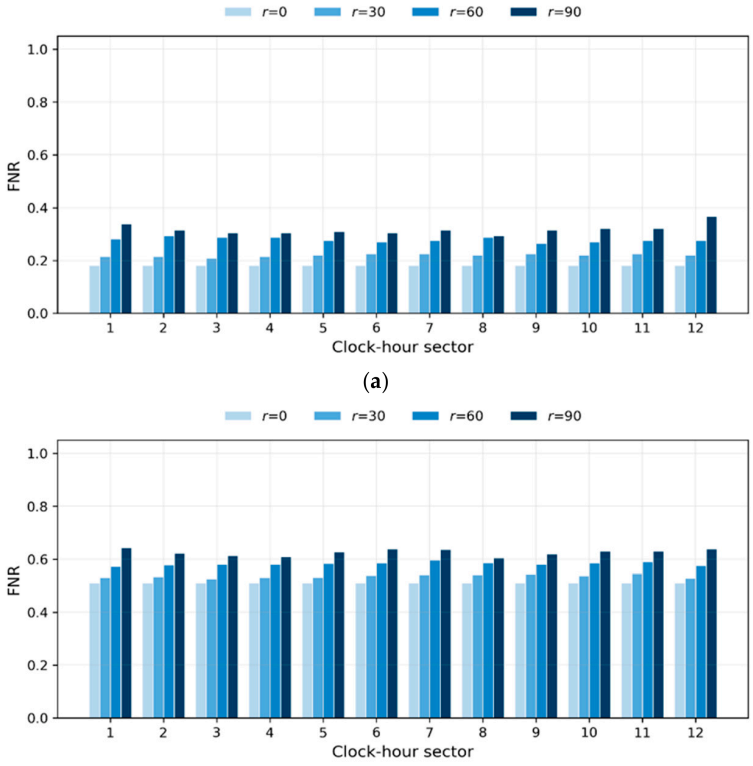



**Figure 20.** False negative rate (FNR) as a function of the noise radius *r* for each Clock-hour sector in cropped optic disc images. Increase in the miss rate (false negative rate) associated with the noise radius *r* for each Clock-hour sector (1–12 o'clock). (**a**) Results using the leak-reduced dataset; (**b**) results using the LODO dataset.

### *3.5. Complementary Verification via Occlusion Sensitivity Analysis*

To complementarily verify which anatomical structures the model actually relies on for its predictions—corresponding to the vulnerable regions identified in the preceding local noise perturbation experiments—occlusion sensitivity analysis was performed. The statistical heatmaps (Mean maps) aggregating the degree of decrease in predicted probability (Drop) for each model are shown in Figure 21 for the leak-reduced dataset and in Figure 22 for the LODO dataset.

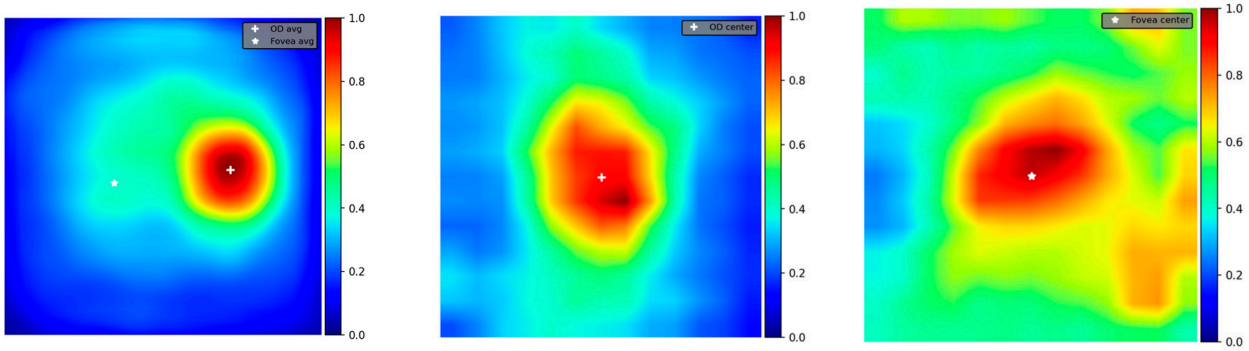
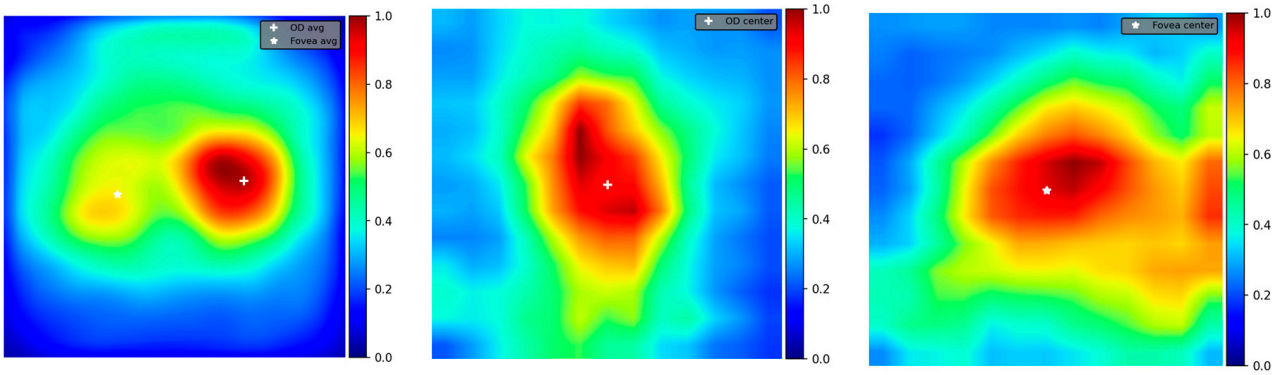



**Figure 21.** Aggregated occlusion sensitivity heatmaps for the leak-reduced dataset. Mean occlusion sensitivity maps averaged across the entire dataset for each experimental model. (**a**) Whole-image model (Model 1); (**b**) optic disc model (Model 2); (**c**) macula model (Model 3).

The output patterns of the heatmaps and the distribution of important regions were consistent between the two datasets. In the evaluation using the whole-image model (Model 1: whole), the optic disc region was identified as the area exerting the strongest influence on the model's classification predictions, while a clear increase in sensitivity was also observed around the macular region.




(**a**) (**b**) (**c**)

**Figure 22.** Aggregated occlusion sensitivity heatmaps for the LODO dataset. Mean occlusion sensitivity maps averaged across the entire dataset for each experimental model. (**a**) Whole-image model (Model 1); (**b**) optic disc model (Model 2); (**c**) macula model (Model 3).

In the evaluation of the models using locally cropped images, regions of interest strongly associated with the corresponding anatomical structures were also visualized. For the optic disc model (Model 2: optic disc), the resulting heatmap accurately captured the structure of the optic disc, with a distinctive pattern in which the high-importance region was distributed broadly along a vertically elongated shape. For the macula model (Model 3: macula), the heatmap clearly indicated a region centered on the macula, with the distribution of the region of interest extending more broadly into the surrounding area compared to the optic disc model.

## 4. Discussion

The series of localized adversarial perturbation analyses conducted in this study revealed both the regional dependence of CNN-based glaucoma diagnosis and points of convergence and divergence with the diagnostic criteria used by ophthalmologists.

Before proceeding with a detailed discussion based on these results, it is necessary to clarify what can and cannot be assessed through FGSM-based perturbation analysis. This method allows for the measurement of changes in model prediction performance (i.e., adversarial vulnerability) resulting from the addition of noise to specific local regions, thereby enabling a spatial understanding of which regions contain elements that influence the model's inference, and to what extent.

However, a decrease in performance resulting from this method alone cannot directly demonstrate the medical or anatomical meaning of the features captured by the model. A performance decline caused by perturbation in a given local region does not necessarily indicate that the model interprets and utilizes specific lesions or morphological changes in the same manner as a physician. Rather, such a decline may reflect mathematical or image-processing factors—such as local texture sensitivity, the frequency components of the image, segmentation size, or the strength of the attack applied to that region.

In light of this interpretive uncertainty inherent to perturbation analysis, the present study discusses the regions influencing the CNN's decisions in conjunction with the results of Occlusion Sensitivity (Section 3.5), an established attribution method. Although this

approach does not directly reveal the anatomical understanding embedded within the model, it allows for a comparison between the spatial distribution of regions that influence the model's inference and the diagnostic criteria employed by physicians, thereby providing a basis for considering the model's behavior.

Figure 23 shows the regions within the optic disc area in Experiment 2 that exhibited the largest changes in accuracy reduction. These regions correspond to areas located on the outer side of the optic cup, namely the disc rim and its surrounding whitish structures. This finding is consistent with the regions of clinical attention associated with the cup-to-disc (C/D) ratio and rim thinning—features that ophthalmologists emphasize in their diagnostic assessments. This suggests that the spatial distribution of regions influencing the CNN's inference is similar to the arrangement of structural features that serve as the basis for physicians' diagnostic judgments.



**Figure 23.** Regions with large AUC fluctuation at $r$ = 30, 40, and 50 pixels from the optic disc center (dashed lines) in Experiment 2. Red indicates high-fluctuation regions.

The sector-wise evaluation based on TSNIT and Clock-hour (Experiment 4) also showed that perturbations to the Superior and Inferior sectors caused substantial drops in performance. Figure 24 shows the sectors with the largest and smallest effects.

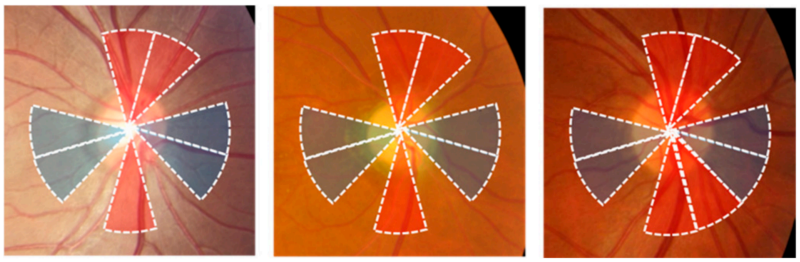


**Figure 24.** Comparison of AUC fluctuation by Clock-hour direction in the optic disc region (Experiment 4). Directions with the largest fluctuation at radii 60–90 pixels (12, 1, and 6 o'clock) are shown in red; directions with the smallest fluctuation (3, 4, 8, and 9 o'clock) are shown in blue. Dashed lines indicate the sector boundaries.

This suggests a possible similarity between the spatial arrangement of the model's inference-influencing regions and the clinical characteristic whereby glaucomatous changes tend to appear preferentially in the superior and inferior quadrants, owing to the vulnerability of the retinal nerve fiber layer (RNFL) in these regions (i.e., the anatomical predisposition for lesion manifestation). Furthermore, the smaller degree of variation observed in the horizontal sectors, where structural features are relatively sparse, suggests that the factors influencing the CNN's decisions may extend beyond the optic disc itself to include the surrounding vascular course and spatial configuration.

Furthermore, this influence observed around the optic disc, particularly in the superior–inferior direction, is consistent with the tendency shown by the disc image model (Model 2) in the occlusion sensitivity analysis presented in Section 3.5, in which the degree of influence was distributed in a vertically elongated pattern. The fact that these two methodologically

distinct approaches both identified an influence oriented toward the superior–inferior direction of the disc can be regarded as a complementary element supporting the results of the present perturbation analysis. Unlike FGSM, which generates noise based on gradients, occlusion analysis is a method that locally masks portions of the image, and is therefore considered less directly susceptible to mathematical effects arising from local texture sensitivity or high-frequency components (such as vascular branching or edges) around the disc. In addition, since occlusion analysis evaluates the entire image uniformly using a fixed patch size, it is thought that biases arising from differences in spatial mask size may be reduced. Moreover, its results are not influenced by the choice of a specific attack-strength parameter. Accordingly, the fact that a similar tendency was observed across these two distinct approaches reduces the likelihood that the present results represent an artifact dependent solely on specific attack conditions or particular image-geometric features, thereby lending additional support to the validity of the regions identified as influencing the model's decisions.

While such commonalities were observed, there were also regions influencing the CNN that did not align with current clinical diagnostic criteria.

In both the analysis of the optic disc in Experiment 1 and the experiment involving noise application from the outer region in Experiment 3, substantial changes were in some cases observed in regions located outside the optic disc itself (Figures 25 and 26). These regions include areas of vascular convergence and branching patterns, suggesting that the surrounding structural features in this periphery may constitute elements influencing the CNN's inference. This is a region that is not directly used as an indicator in current standard clinical diagnosis.

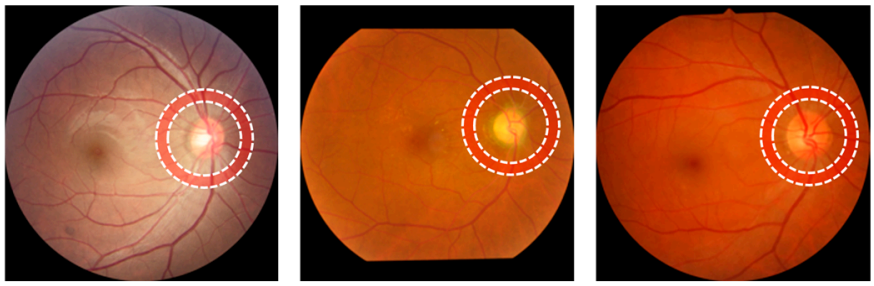


**Figure 25.** Regions with large AUC fluctuation at radii 30–40 pixels (dashed line) from the optic disc center in whole images (Experiment 1). Red indicates high-fluctuation regions.



**Figure 26.** Regions with large AUC fluctuation at $r_{in}$ = 70, 80, and 90 pixels (dashed lines) in the optic disc area (Experiment 3). Red indicates high-fluctuation regions.

Another notable discrepancy involves the substantial influence of the macula on CNN prediction performance. In clinical practice, the macula is not considered a primary diagnostic target and is typically used only as a supplementary reference alongside visual field testing. However, in Experiment 2 using cropped macula images, a clear performance drop was observed from the initial stages. In Experiment 3 (Figure 27), the AUC drop when

noise was applied to the outermost band (radii 85–90 pixels) of the macula region exceeded that of the corresponding outer margin of the optic disc.

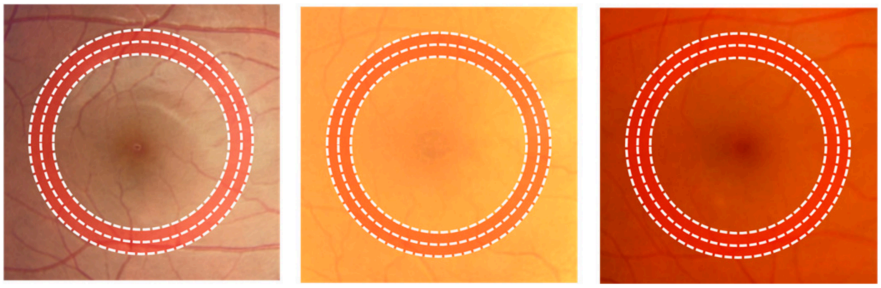


**Figure 27.** Regions with large AUC fluctuation at $r_{in}$ = 70, 80, and 90 pixels (dashed lines) in the macular area (Experiment 3). Red indicates high-fluctuation regions.

This sensitivity in regions beyond the optic disc is also consistent with the results of the occlusion sensitivity analysis presented in Section 3.5. In the heatmaps derived from the whole-image model (Model 1), while the optic disc was identified as the region exerting the strongest influence, a decrease in predicted probability (i.e., sensitivity) was also observed around the macula and the surrounding vascular pathways. The fact that these regions showed consistent influence across two methodologically distinct approaches can be regarded as an element that complements the validity of the present analytical results.

Thus, the observed high sensitivity in regions that diverge from current clinical diagnostic criteria (namely, the peridiscal vascular areas and the macula) can be attributed to two possible explanations.

The first possibility is that these findings reflect mathematical or geometric artifacts (biases) specific to the dataset. It cannot be ruled out that variations in illumination or image quality specific to particular camera models or institutions may have introduced biases during model training, which subsequently manifest as sensitivity. However, in the present study, Leave-One-Dataset-Out (LODO) validation was conducted using multiple distinct datasets held out in turn, in order to smooth out inter-institutional bias and domain-specific noise, with the macro-mean employed as the evaluation metric. Accordingly, although the influence of domain-specific noise cannot be entirely eliminated when evaluating across data derived from markedly different imaging environments, this approach is considered to reduce, to some extent, the likelihood that the present results are attributable solely to the characteristics of a particular institution or device.

The second possibility is that the model is capturing previously unrecognized features related to glaucoma that have not been directly identified in conventional clinical diagnosis. Recent advances in deep learning have demonstrated that AI can achieve tasks that even experienced ophthalmologists find difficult [37]. For example, Poplin et al. demonstrated that a deep learning algorithm could predict cardiovascular risk factors and demographic information from fundus images with high accuracy [38]. In particular, the successful prediction of age (mean absolute error 3.26 years), sex (AUC 0.97), smoking status (AUC 0.71), and systolic blood pressure (mean absolute error 11.23 mmHg) was surprising even to ophthalmologists with years of experience, as such associations would not have been anticipated based on conventional clinical knowledge. Accordingly, it is plausible that CNNs are capturing features that have not previously been recognized by humans.

In light of this background, the ophthalmologist's assessment of the results obtained in this study is as follows. First, regarding the perturbation effects centered on the optic disc observed across EX1–4, these showed a high degree of consistency with known pathological features and were judged to be clinically valid. In addition, with respect to the examination centered on the macula, the performance decline observed in EX1 spatially corresponded

to the distribution of pathological features associated with the thinning of retinal nerve fiber layer defects (RNFLD), and was therefore considered to represent plausible behavior. On the other hand, the results of EX3 were considered valid in that they correspond to pathological features in cases of advanced glaucoma involving central visual field loss. However, since the present study was not limited to advanced glaucoma cases, this analysis suggested that features influencing inference are also present around the macula even in the early stages of glaucoma. Although this finding is not inconsistent with known pathological features, it represents an element that has not been strongly recognized as a direct clinical indicator in current standard diagnostic practice.

It should be noted, however, that this series of clinical interpretations is based on the perspective of the authors of this study (ophthalmologists) and has not undergone objective verification, such as an independent blinded evaluation by external specialists. Therefore, while the possibility cannot be ruled out that the influence attributed to the macular and vascular regions captured by the model reflects clinically unexplored, previously unrecognized glaucoma-related features, it remains difficult at this stage to rigorously distinguish whether this represents a genuinely generalizable novel disease marker, or merely reflects the aforementioned dataset-specific bias, is difficult to rigorously distinguish at this stage. As such, there remains room for further discussion in disentangling medically meaningful interpretation from mathematical bias in the model's behavior, and this gives rise to several limitations of the present study, as outlined below.

The first limitation concerns the need for caution in rigorously verifying whether the observed tendency—namely, that the model is influenced by the macula and vascular regions—truly reflects medically or pathologically meaningful features. As noted in the ophthalmologist's assessment above, the high sensitivity observed around the macula is consistent with the pathological distribution associated with advanced glaucoma involving central visual field loss; however, the images used in this study were not restricted to a specific disease stage. Consequently, although the results suggest that macula-related features may also be present in the early stages of glaucoma, the dataset used in this study lacks detailed clinical metadata regarding disease severity or stage, making it difficult at this stage to rigorously stratify and verify the findings by stage. Furthermore, as previously noted, these interpretations remain based solely on the observations of the authors (ophthalmologists) and have not been obtained through an objective verification process such as rigorous blinded evaluation by independent experts. Therefore, in order to rigorously determine whether these findings represent dataset-specific biases or genuinely generalizable novel disease markers, future work should include detailed analyses stratified by glaucoma severity, large-scale blinded image review involving multiple external experts, and further scrutiny of the underlying medical evidence. These are identified as important directions for future research and constitute a key limitation of the present study. In addition, the evaluation is limited to a single network architecture (VGG16). Previous studies have reported that other architectures can achieve higher classification performance [36]. Future work should apply the same analysis to different model architectures to examine how architectural differences affect diagnostic reasoning and attention regions.

Second, there are limitations regarding the scale and variation in the datasets used for validation. In this study, two datasets—leak-reduced and LODO (Leave-One-Dataset-Out)—were used to minimize the influence of inter-institutional bias and data leakage as much as possible. However, particularly under the more stringent LODO condition, in which leakage was more thoroughly eliminated, the baseline classification accuracy without noise (r = 0) was itself somewhat limited and appears to plateau near this value given the current dataset scale and model configuration. To further improve the model's generalization performance and more robustly establish the universality of its diagnostic

basis, comprehensive validation—starting from the model training stage—using larger-scale external datasets collected from a more diverse range of institutions will be necessary. In addition, although the present study employed the macro-mean derived from LODO validation in an attempt to smooth out image-related biases (artifacts) such as those arising from specific camera models or illumination variation, it cannot be concluded that the influence of such mathematical or geometric biases has been fully eliminated at the current dataset scale, and the possibility remains that this mitigation is, at present, insufficient. Future work should therefore expand the validation framework by incorporating a larger number of datasets, in order to further reduce the potential influence of domain-specific noise. In doing so, as discussed in relation to the first limitation, it will also be necessary to utilize datasets that include information on disease stage or severity, and to examine in detail how the basis for diagnosis shifts according to disease progression.

Third, there remains room to extend this analysis by applying noise to a wider range of local regions. This study focused on the major anatomical landmarks of the optic disc and macula; however, further investigation is needed into whether other structural information across the entire fundus image may also influence the model's predictions.

Looking ahead, if the CNN is found to rely on clinically non-meaningful regions (such as those related to device characteristics or illumination conditions), the findings of this study would serve as foundational data for designing safer and more robust models. Conversely, if the identified regions correspond to a genuine glaucoma-related feature, this would provide evidence that AI is capable of capturing precursors of pathology that have not been recognized by humans, extending beyond current diagnostic criteria, and would point to the potential of such approaches to support the discovery of novel disease markers. Accordingly, the analysis of accuracy fluctuations induced by localized adversarial perturbations is expected to serve, in future work, as a foundational framework for evaluating the basis of inference in medical AI.

## 5. Conclusions

In this study, we quantitatively evaluated the regional influence on CNN decision-making in automated glaucoma diagnosis using locally applied adversarial perturbations. The results showed that perturbations near the neuroretinal rim and in the superior and inferior sectors had the greatest impact on model performance, suggesting that the CNN may rely on pathologically relevant structural features as a basis for its decisions. At the same time, the notable influence of the vascular and macular regions revealed a divergence from conventional clinical diagnostic criteria, suggesting that the CNN may also be utilizing features beyond the optic disc. In conclusion, this approach can serve as a means of quantitatively bridging the gap between clinical interpretation and the black-box nature of CNN decision-making, and holds potential to contribute not only to validating the reasoning of automated diagnostic models but also to the discovery of novel digital biomarkers in fundus images. Whether the extra-papillary features observed in this study carry genuine clinical significance remains to be verified in future work; nonetheless, this approach is expected to offer new insights for evaluating the reasoning basis of medical AI going forward.

**Author Contributions:** Conceptualization, H.S., I.O. and S.I.; methodology, H.S., K.H., I.O. and S.I.; software, K.H., I.O., T.N.K. and K.M.; validation, H.S., K.H., I.O., T.N.K. and K.M.; formal analysis, H.S. and K.H.; investigation, I.O., T.N.K. and K.M.; resources, S.I. and H.O.; data curation, H.S., K.H., I.O. and S.I.; writing—original draft preparation, H.S. and K.H.; writing—review and editing, H.S., S.I., H.O., O.F., N.Y. and M.K.; visualization, H.S., K.H. and I.O.; supervision, S.I., H.O., O.F., N.Y. and M.K.; project administration, H.S., I.O. and S.I.; funding acquisition, S.I. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Japan Society for the Promotion of Science (JSPS), grant number JP 23K11150.

**Institutional Review Board Statement:** This study did not involve any new data collection from human subjects; all experiments were performed on the publicly available SMDG-19 dataset, which aggregates de-identified fundus images from previously published datasets, each obtained under its original institutional ethical approvals.

**Informed Consent Statement:** This study used publicly available, de-identified fundus image datasets aggregated in SMDG-19; informed consent had been obtained by the original data providers.

**Data Availability Statement:** The SMDG-19 dataset used in this study is publicly available at https://www.kaggle.com/datasets/deathtrooper/multichannel-glaucoma-benchmark-dataset (accessed on 15 June 2025). The preprocessing code, trained models, and experimental scripts developed in this study are available from the corresponding author upon reasonable request.

**Acknowledgments:** During the course of this study (2025–2026), the authors used multiple generative AI tools—Claude (Anthropic, San Francisco, CA, USA), Gemini (Google, Mountain View, CA, USA), and ChatGPT (OpenAI, San Francisco, CA, USA)—to assist with literature exploration, drafting and refining text, and supporting the development of analysis code. The providers released successive model updates during this period; the most recent versions used were Claude Opus 4.8, Gemini 3.5 Flash, and GPT-5.5. All AI-generated outputs were reviewed, verified, and edited by the authors, who take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| CNN | Convolutional Neural Network |
| DL | Deep Learning |
| AI | Artificial Intelligence |
| XAI | Explainable Artificial Intelligence |
| FGSM | Fast Gradient Sign Method |
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| AUC | Area Under the Curve |
| RNFL | Retinal Nerve Fiber Layer |
| RNFLD | Retinal Nerve Fiber Layer Defect |
| C/D ratio | Cup-to-Disc Ratio |
| TSNIT | Temporal–Superior–Nasal–Inferior–Temporal |
| VGG16 | Visual Geometry Group 16 |
| SGD | Stochastic Gradient Descent |
| LODO | Leave-One-Domain-Out |

## Appendix A

**Table A1.** Provenance of the glaucoma labels for each source dataset used in this study. The labels in SMDG-19 are not a single ground truth created under a centralized, unified procedure for SMDG-19; they inherit the heterogeneous reference standards of the constituent datasets. This table summarizes only what is stated in each source's original paper or official page.

| Source Dataset | How the Glaucoma Label Was Determined | Graders (Number/Expertise) | Original Purpose | Provenance Group | Reference |
|---|---|---|---|---|---|
| REFUGE1 | Reference standard based not on fundus photographs alone but on health records; OCT, visual fields, and other clinical information were also referenced. | Existing clinical records (no new fundus-only grading described). | Optic disc/cup segmentation and glaucoma classification challenge. | A. Clinical/ multimodal | [39] |
| G1020 | A clinical diagnosis of glaucoma is assigned per patient; structural annotations were verified by specialists. | Number of glaucoma diagnosers not stated; structural annotations verified/corrected by one specialist and one ophthalmologist with >25 years' experience. | Glaucoma-detection benchmark. | A. Clinical/ multimodal | [40] |
| ORIGA (-light) | Annotated by trained professionals at SERI; source images from the Singapore Malay Eye Study (SiMES), where glaucoma was defined by ISGEO criteria using disc assessment, visual fields, gonioscopy, and IOP re-measurement; the ORIGA paper alone does not detail the final binary decision rule. | Trained professionals (number not stated), backed by the SiMES clinical workup. | Database for glaucoma analysis/research. | A. Clinical (background)/ under-documented | [41,42] |
| EyePACS-AIROGS | Referable/non-referable glaucoma from fundus photographs; each image graded twice, with disagreements adjudicated by glaucoma specialists; based on fundus findings suggestive of visual-field damage (not a multimodal diagnosis from clinical records). | Graders who passed an optic-disc assessment test (ophthalmologists, glaucoma specialists, residents, optometrists); two graders, disagreements adjudicated by one of two glaucoma specialists. | Glaucoma-screening challenge. | B. Image grading/ adjudication | [43,44] |
| JSIEC-1000 | 39 disease/finding classes assigned by an expert panel using consensus imaging reference standards (image-based criteria from EyeWiki and textbooks); image-based rather than a confirmed diagnosis from clinical records. | Per group, one senior retinal specialist (>7 yr) and one trained ophthalmologist (>3 yr); disagreements finalized by a panel of five senior retinal specialists. | Multi-disease (39-class) fundus detection. | B. Image grading/ adjudication | [45] |

**Table A1.** *Cont.*

| Source Dataset | How the Glaucoma Label Was Determined | Graders (Number/Expertise) | Original Purpose | Provenance Group | Reference |
|---|---|---|---|---|---|
| OIA-ODIR (ODIR-5K) | Glaucoma/normal inherited from multi-disease annotation (physician diagnostic keywords); the benchmark uses ≥2-of-3 agreement plus adjudication; no glaucoma-specific visual-field/OCT/IOP rule is stated in the public documentation. | Three annotators (>2 yr) and three adjudicating ophthalmologists (>10 yr). | 8-class multi-disease fundus classification. | B. Image grading/ adjudication | [46,47] |
| BEH (Bangladesh Eye Hospital) | Fundus images annotated by two ophthalmologists; an explicit diagnostic rule is not public. | Two specialists (a pediatric ophthalmologist and a glaucoma/refractive surgeon). | Glaucoma detection (private dataset). | B. Image-based expert annotation | [48] |
| FIVES | 200 glaucoma/200 normal; patients were reportedly diagnosed after comprehensive systemic and ophthalmic examination, but the individual glaucoma/non-glaucoma decision rule is not specified (the public description centers on vessel segmentation). | Graders of the disease labels not stated (vessel annotation by 3 senior and 24 junior graders). | Retinal vessel segmentation. | C. Glaucoma not the primary purpose | [49] |
| HRF | 15 healthy/15 diabetic retinopathy/ 15 glaucoma; the public gold standard is vessel segmentation, and the basis for the disease-category assignment is not described. | Assigners of the disease categories unknown. | High-resolution fundus vessel-segmentation database. | C. Glaucoma not primary/under-documented | [50] |
| DR-HAGIS | Primary purpose is retinal vessel extraction; images from UK NHS diabetic eye screening; the procedure for creating glaucoma labels is not detailed. | Unknown. | Vessel extraction from diabetic-screening images. | C. Glaucoma not primary/under-documented | [51] |
| sjchoi86-HRF | A GitHub repository with four categories (normal/cataract/glaucoma/retina-disease); the public README does not describe who assigned the glaucoma labels or how. | Unknown (no grader information in the README). | General-purpose 4-class retinal-disease repository. | C. Sparsely documented | [52] |

## References


1. World Health Organization (WHO). Blindness and Vision Impairment. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on 20 May 2026).
2. Glaucoma Research Foundation. Glaucoma Facts and Stats. Available online: https://glaucoma.org/articles/glaucoma-facts-and-stats (accessed on 19 May 2026).
3. Peng, Y.; Han, L. Global, regional and national burden of glaucoma from 1990 to 2021 and projections to 2050: A retrospective cross-sectional study. *BMJ Open* **2026**, *16*, e108975. [CrossRef] [PubMed]
4. Yuzawa, M.; Kawamura, A.; Mori, R. *Color Atlas of Fundus Lesions*, 7th ed.; Igaku-Shoin: Tokyo, Japan, 2021; pp. 556–569.
5. Wang, C.; Chen, D.-F.; Shang, X. Evaluating Diagnostic Concordance in Primary Open-Angle Glaucoma Among Academic Glaucoma Subspecialists. *Diagnostics* **2024**, *14*, 2460. [CrossRef] [PubMed]
6. Huang, X.; Islam, M.R. Artificial intelligence in glaucoma: Opportunities, challenges, and future directions. *BioMed. Eng. OnLine* **2023**, *22*, 126. [CrossRef] [PubMed]
7. Li, Z.; He, Y. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* **2018**, *125*, 1199–1206. [CrossRef] [PubMed]
8. Lenka, S.; Mayaluri, Z.L.; Panda, G. Glaucoma Detection from Retinal Fundus Images Using Graph Convolution Based Multi-Task Model. *E-Prime* **2025**, *11*, 100931. [CrossRef]
9. Pu, R.; Qin, Y.; Peng, D.; Song, X.; Zheng, H. Deep Reversible Consistency Learning for Cross-Modal Retrieval. *IEEE Trans. Multimed.* **2025**, *27*, 4095–4106. [CrossRef]
10. Chen, Y.; Lu, W.; Qin, X.; Wang, J.; Xie, X. MetaFed: Federated Learning Among Federations with Cyclic Knowledge Distillation for Personalized Healthcare. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 16671–16682. [CrossRef] [PubMed]
11. Banerjee, C.; Nguyen, K.; Salvado, O.; Tran, T.; Fookes, C. PINNs for Medical Image Analysis: A Survey. *arXiv* **2024**, arXiv:2408.01026.
12. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608. [CrossRef]
13. Ennab, M.; Mcheick, H. Designing an Interpretability-Based Model to Explain the Artificial Intelligence Algorithms in Healthcare. *Diagnostics* **2022**, *12*, 1557. [CrossRef] [PubMed]
14. Ennab, M.; Mcheick, H. Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 12. [CrossRef]
15. Ahmed, F.; Naz, N.S. Explainable artificial intelligence (XAI) in medical imaging: A systematic review of techniques, applications, and challenges. *BMC Med. Imaging* **2026**, *26*, 37. [CrossRef] [PubMed]
16. Selvaraju, R.R.; Cogswell, M. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
17. Zhang, H.; Ogasawara, K. Grad-CAM-Based Explainable Artificial Intelligence Related to Medical Text Processing. *Bioengineering* **2023**, *10*, 1070. [CrossRef] [PubMed]
18. Dusza, P.; Banzato, T. Comparative evaluation of CAM methods for enhancing explainability in veterinary radiography. *Sci. Rep.* **2025**, *15*, 29690. [CrossRef] [PubMed]
19. Kim, S.S.Y.; Meister, N. HIVE: Evaluating the Human Interpretability of Visual Explanations. *arXiv* **2021**, arXiv:2112.03184. [CrossRef]
20. Kent, M.G.; Schiavon, S. Predicting Window View Preferences Using the Environmental Information Criteria. *Leukos* **2022**, *19*, 190–209. [CrossRef]
21. Goodfellow, I.J.; Shlens, J. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**, arXiv:1412.6572. [CrossRef]
22. Margeloiu, A.; Simidjievski, N. Improving Interpretability in Medical Imaging Diagnosis Using Adversarial Training. *arXiv* **2020**, arXiv:2012.01166. [CrossRef]
23. Woods, W.; Chen, J.; Teuscher, C. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nat. Mach. Intell.* **2019**, *1*, 508–516. [CrossRef]
24. Paschali, M.; Conjeti, S. Generalizability vs. Robustness: Adversarial Examples for Medical Imaging. *arXiv* **2018**, arXiv:1804.00504. [CrossRef]
25. Ma, X.; Niu, Y. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit.* **2021**, *110*, 107332. [CrossRef]
26. Zhao, H.; Wang, B.; Fu, Y.; Li, N.; Gao, Z. Fixed-time adaptive fault-tolerant control of a multi-mode VTOL UAV with variable prescribed performance boundaries under random disturbances. *ISA Trans.* **2026**, *175*, 126–138. [CrossRef] [PubMed]
27. Zhao, H.; Wang, B.; Fu, Y.; Zhou, M.; Li, N.; Gao, Z. Fixed-Time Adaptive Smooth Control with Relaxed Performance Constraints for a Canard Rotor/Wing UAV Subject to Aerodynamic Interferences and Actuator Faults. *Aerosp. Sci. Technol.* **2026**, *178*, 112959. [CrossRef]

28. Kiefer, R.; Abid, M. A Catalog of Public Glaucoma Datasets for Machine Learning Applications. In *Proceedings of the 2023 7th International Conference on Information System and Data Mining (ICISDM 2023), Atlanta, GA, USA, 10–12 May 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 24–31. [CrossRef]
29. Kiefer, R.; Steen, J. Standardized Multi-Channel Dataset for Glaucoma (SMDG-19), Version 19 [Dataset]. Kaggle. 2023. Available online: https://www.kaggle.com/datasets/deathtrooper/multichannel-glaucoma-benchmark-dataset (accessed on 15 June 2025).
30. Kovalyk, O.; Morales-Sánchez, J. PAPILA: Dataset with Fundus Images and Clinical Data of Both Eyes of the Same Patient for Glaucoma Assessment. *Sci. Data* **2022**, *9*, 291. [CrossRef] [PubMed]
31. Sivaswamy, J.; Krishnadas, S.R. Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 29 April–2 May 2014; pp. 53–56. [CrossRef]
32. Orlando, J.I.; Barbosa Breda, J. Towards a Glaucoma Risk Index Based on Simulated Hemodynamics from Fundus Images. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI 2018)*; Springer: Cham, Switzerland, 2018; pp. 65–73. [CrossRef]
33. Gervelmeyer, J.; Müller, S. Fundus Image Toolbox: A Python Package for Fundus Image Processing. *J. Open Source Softw.* **2025**, *10*, 7101. [CrossRef]
34. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
35. Zeng, Q.; Zhang, L.; Liu, Y.; Zhu, C.; Zhang, F. FunOTTA: On-the-Fly Adaptation on Cross-Domain Fundus Image via Stable Test-time Training. *arXiv* **2024**, arXiv:2407.04396. [CrossRef]
36. Diaz-Pinto, A.; Morales, S. CNNs for automatic glaucoma assessment using fundus images: An extensive validation. *Biomed. Eng. OnLine* **2019**, *18*, 29. [CrossRef] [PubMed]
37. Korot, E.; Wagner, S.K. Will AI Replace Ophthalmologists? *Transl. Vis. Sci. Technol.* **2020**, *9*, 2. [CrossRef] [PubMed]
38. Poplin, R.; Varadarajan, A.V. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2018**, *2*, 158–164. [CrossRef] [PubMed]
39. Orlando, J.I.; Fu, H.; Barbosa Breda, J.; van Keer, K.; Bathula, D.R.; Diaz-Pinto, A.; Fang, R.; Heng, P.-A.; Kim, J.; Lee, J.; et al. REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs. *Med. Image Anal.* **2020**, *59*, 101570. [CrossRef] [PubMed]
40. Bajwa, M.N.; Singh, G.A.P.; Neumeier, W.; Malik, M.I.; Dengel, A.; Ahmed, S. G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–7. [CrossRef]
41. Zhang, Z.; Yin, F.S.; Liu, J.; Wong, W.K.; Tan, N.M.; Lee, B.H.; Cheng, J.; Wong, T.Y. ORIGA-light: An Online Retinal Fundus Image Database for Glaucoma Analysis and Research. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Buenos Aires, Argentina, 31 August–4 September 2010; pp. 3065–3068. [CrossRef] [PubMed]
42. Shen, S.Y.; Wong, T.Y.; Foster, P.J.; Loo, J.-L.; Rosman, M.; Loon, S.-C.; Wong, W.L.; Saw, S.-M.; Aung, T. The Prevalence and Types of Glaucoma in Malay People: The Singapore Malay Eye Study. *Invest. Ophthalmol. Vis. Sci.* **2008**, *49*, 3846–3851. [CrossRef] [PubMed]
43. de Vente, C.; Vermeer, K.A.; Jaccard, N.; Wang, H.; Sun, H.; Khader, F.; Truhn, D.; Aimyshev, T.; Zhanibekuly, Y.; Le, T.-D.; et al. AIROGS: Artificial Intelligence for Robust Glaucoma Screening Challenge. *IEEE Trans. Med. Imaging* **2024**, *43*, 542–557. [CrossRef] [PubMed]
44. Lemij, H.G.; de Vente, C.; Sánchez, C.I.; Vermeer, K.A. Characteristics of a Large, Labeled Data Set for the Training of Artificial Intelligence for Glaucoma Screening with Fundus Photographs. *Ophthalmol. Sci.* **2023**, *3*, 100300. [CrossRef] [PubMed]
45. Cen, L.-P.; Ji, J.; Lin, J.-W.; Ju, S.-T.; Lin, H.-J.; Li, T.-P.; Wang, Y.; Yang, J.-F.; Liu, Y.-F.; Tan, S.; et al. Automatic Detection of 39 Fundus Diseases and Conditions in Retinal Photographs Using Deep Neural Networks. *Nat. Commun.* **2021**, *12*, 4828. [CrossRef] [PubMed]
46. Li, N.; Li, T.; Hu, C.; Wang, K.; Kang, H. A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-Disease Detection. In *Benchmarking, Measuring, and Optimizing (Bench 2020); Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2021; Volume 12614, pp. 177–193. [CrossRef] [PubMed]
47. Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019). Available online: https://odir2019.grand-challenge.org/ (accessed on 28 June 2026).
48. Islam, M.T.; Mashfu, S.T.; Faisal, A.; Siam, S.C.; Naheen, I.T.; Khan, R. Deep Learning-Based Glaucoma Detection with Cropped Optic Cup and Disc and Blood Vessel Segmentation. *IEEE Access* **2022**, *10*, 2828–2841. [CrossRef]
49. Jin, K.; Huang, X.; Zhou, J.; Li, Y.; Yan, Y.; Sun, Y.; Zhang, Q.; Wang, Y.; Ye, J. FIVES: A Fundus Image Dataset for Artificial Intelligence Based Vessel Segmentation. *Sci. Data* **2022**, *9*, 475. [CrossRef] [PubMed]

50. Budai, A.; Bock, R.; Maier, A.; Hornegger, J.; Michelson, G. Robust Vessel Segmentation in Fundus Images. *Int. J. Biomed. Imaging* **2013**, *2013*, 154860. [CrossRef] [PubMed]
51. Holm, S.; Russell, G.; Nourrit, V.; McLoughlin, N. DR HAGIS—A Fundus Image Database for the Automatic Extraction of Retinal Surface Vessels from Diabetic Patients. *J. Med. Imaging* **2017**, *4*, 014503. [CrossRef] [PubMed]
52. GitHub Repository. yiweichen04. retina_dataset (sjchoi86-HRF). Available online: https://github.com/yiweichen04/retina_dataset (accessed on 28 June 2026).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.