

Development of a risk assessment tool for early detection of glaucoma based on machine learning methods

Eduard Schiller¹, Sabina Rakhmetulayeva², Anastassiya Tyunina²

¹Department of Machine Learning and Data Science, Institute of Automation and Information Technologies, Satbayev University, Almaty, Kazakhstan

²Department of Information Systems, Institute of Automation and Information Technologies, Satbayev University, Almaty, Kazakhstan

Article Info

Article history:

Received Oct 21, 2025

Revised Apr 8, 2026

Accepted Sep 4, 2026

Keywords:

Cycle-consistent generative
adversarial network
Fundus photography
Glaucoma screening
ResNet-50
Synthetic data augmentation

ABSTRACT

Glaucoma is a progressive optic neuropathy and a major cause of irreversible blindness, making reliable early screening particularly important. This study proposes an automated glaucoma risk assessment framework based on color fundus photographs, combining a ResNet-50 classifier with cycle-consistent generative adversarial network (CycleGAN)-based synthetic data augmentation. The model was evaluated on the retinal image database for optic nerve evaluation for deep learning (RIM-ONE DL) dataset under limited and imbalanced data conditions. Training on real images alone resulted in an accuracy of 76.8%, recall of 74.2%, and area under the receiver operating characteristic curve (AUC-ROC) of 0.825. After adding 500 synthetic normal and 500 synthetic glaucoma images to the training set, accuracy increased to 91.4%, recall to 89.7%, and AUC-ROC to 0.943. Synthetic image quality was additionally evaluated using Fréchet inception distance (FID), kernel inception distance (KID), and learned perceptual image patch similarity (LPIPS) metrics and expert visual assessment. The results demonstrate that CycleGAN-based augmentation can substantially improve glaucoma classification performance under limited-data conditions while preserving clinically relevant retinal structures. The proposed approach may therefore serve as a supplementary decision-support tool for automated glaucoma screening.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Sabina Rakhmetulayeva

Department of Information Systems, Institute of Automation and Information Technologies

Satbayev University

Almaty, Kazakhstan

Email: ssrakhmetulayeva@gmail.com

1. INTRODUCTION

Glaucoma is a chronic and progressive optic neuropathy and remains one of the leading causes of irreversible blindness worldwide. The disease is characterized by the gradual loss of retinal ganglion cells and structural damage to the optic nerve head, ultimately resulting in permanent visual field defects. The global burden of glaucoma continues to increase, primarily due to population aging and increased life expectancy [1]. Figure 1 illustrates the projected growth in the number of affected individuals.

Despite recent advances in automated retinal image analysis, early glaucoma detection remains challenging, particularly when deep learning models are trained on limited and imbalanced annotated datasets. Convolutional neural networks (CNNs) are capable of automatically extracting complex structural patterns

associated with glaucomatous optic neuropathy and therefore provide a scalable and objective alternative to conventional manual assessment. However, their diagnostic performance strongly depends on the availability, diversity, and quality of labeled medical images, which remain limited in many ophthalmological applications.

To address these limitations, this study combines a ResNet-50 classifier with cycle-consistent generative adversarial network (CycleGAN)-based synthetic data augmentation. In contrast to approaches that evaluate augmentation primarily through changes in classification accuracy, the proposed framework assesses synthetic data at several complementary levels, including classification performance, synthetic image quality using Frechet inception distance (FID), kernel inception distance (KID), and learned perceptual image patch similarity (LPIPS), probability calibration, statistical significance testing, and expert visual assessment. This multi-level evaluation makes it possible to determine not only whether synthetic images improve model performance, but also whether they remain sufficiently similar to real fundus photographs and preserve clinically meaningful anatomical characteristics.

The main contribution of this study is therefore a comprehensive evaluation of CycleGAN-based augmentation under limited-data conditions. The proposed framework integrates generative modeling, quantitative assessment of synthetic image quality, probability calibration, statistical validation, and expert-based clinical assessment within a unified glaucoma screening pipeline. Furthermore, the study explicitly evaluates the effect of synthetic augmentation on model sensitivity, robustness, and generalization while ensuring that synthetic samples are included only in the training data and not in the validation or test sets.

Experimental results demonstrate that the addition of synthetic images substantially improves the performance of the ResNet-50 classifier. Recall increased from 74.2.% to 89.7%, while area under the receiver operating characteristic curve (AUC-ROC) improved from 0.825 to 0.943. These findings indicate that carefully controlled synthetic augmentation can reduce the adverse effects of limited data and class imbalance and may provide a practical strategy for improving automated glaucoma screening.

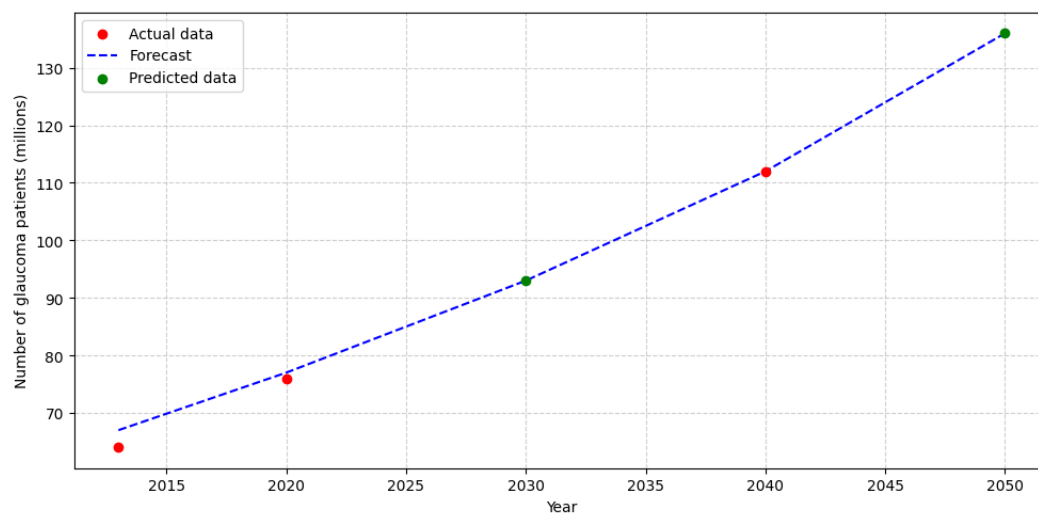


Figure 1. Forecast of an increase in the number of glaucoma cases

2. LITERATURE REVIEW

2.1. Theoretical aspects of glaucoma diagnosis and the role of machine learning in ophthalmology

Glaucoma is a progressive optic neuropathy and a major cause of irreversible blindness worldwide [2]. Its global prevalence is expected to increase substantially, with approximately 111.8 million people projected to be affected by 2040 [1]. Because conventional diagnosis remains resource-intensive and dependent on specialist expertise, AI-based fundus image analysis has emerged as a promising approach for scalable screening [3]–[5].

Residual neural network architectures have been widely adopted in medical image analysis because their skip connections enable stable optimization of deep models and alleviate vanishing-gradient problems [6]. ResNet-based architectures have demonstrated strong performance in fundus image classification and glaucoma detection, particularly when transfer learning is applied to relatively small ophthalmological datasets [7].

Nevertheless, publicly available ophthalmological datasets frequently contain substantially fewer pathological samples than healthy images. This imbalance may cause a classifier to become biased toward the majority class and may reduce sensitivity to glaucomatous cases [8], which is particularly problematic in screening applications where false-negative predictions can delay further clinical examination. Consequently, improving class balance and increasing training-set diversity remain important methodological objectives in automated glaucoma detection.

To mitigate these limitations, recent studies have explored synthetic data augmentation using generative adversarial networks. The GAN framework enables the generation of realistic synthetic images through adversarial training [9]. In ophthalmology, GAN-based methods have been applied to retinal image synthesis and disease-specific data augmentation, demonstrating improvements in classification robustness [10]. CycleGAN, in particular, enables unpaired image-to-image translation and has been successfully applied in medical imaging scenarios where paired samples are unavailable [11].

Overall, the reviewed literature indicates that the integration of deep residual architectures with synthetic data generation represents a promising strategy for improving the robustness and clinical applicability of automated glaucoma screening systems [5].

2.2. Analysis of existing methods and models for automated glaucoma detection

Automated glaucoma detection has attracted increasing research interest due to the growing prevalence of the disease and the widespread availability of retinal imaging technologies. Structural changes associated with glaucomatous optic neuropathy, particularly alterations of the optic disc and optic cup, can be assessed using color fundus photographs and optical coherence tomography (OCT). Recent advances in machine learning, especially deep learning, have enabled the development of automated systems capable of extracting diagnostically relevant features directly from ophthalmic images [4], [5].

ResNet-50 has become a commonly adopted backbone due to its residual learning mechanism, which allows stable training of deep architectures by mitigating vanishing gradient effects [6]. This property is particularly important for medical imaging tasks, where datasets are often limited in size and subtle structural differences must be captured reliably.

Comparative evaluations against earlier architectures such as VGG16, Inception-V3, and DenseNet121 indicate that ResNet-50 often achieves superior classification performance, especially when transfer learning is employed [12], [13]. For clinical deployment, interpretability of automated systems is a critical requirement. Visualization techniques such as gradient-weighted class activation mapping (Grad-CAM) are commonly used to highlight image regions that contribute most strongly to model predictions [14].

Specialized deep learning architectures, including supervised learning-based segmentation network (SLS-Net) and supervised learning-based segmentation and regression network (SLSR-Net), have been proposed for joint optic disc and cup segmentation [15]. Machine learning models applied to glaucoma diagnosis can be broadly divided into classical approaches and deep learning-based methods. This taxonomy is illustrated in Figure 2, which presents a structured overview of models used for eye disease detection.

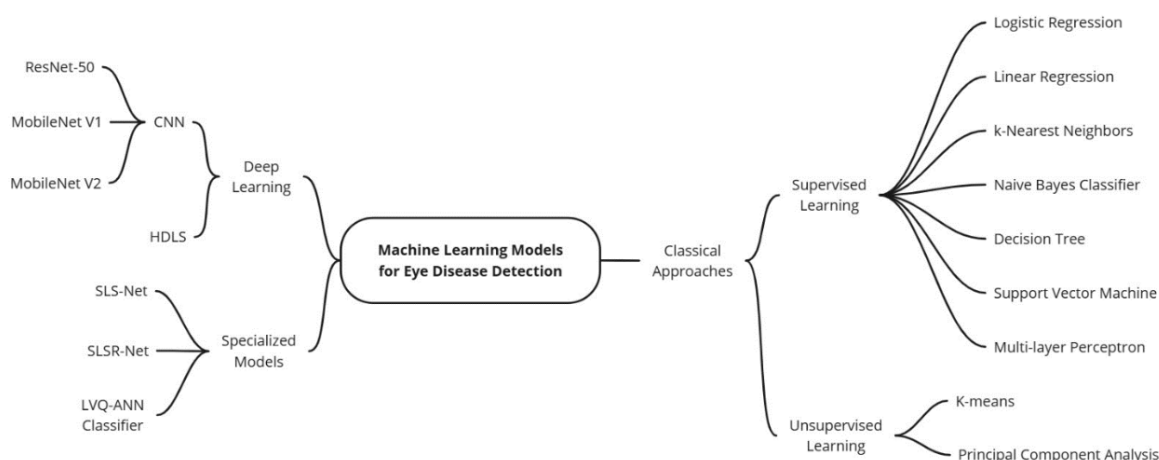


Figure 2. Taxonomy of machine learning models for eye disease detection

To provide a quantitative comparison of CNN architectures commonly used for glaucoma detection, Table 1 summarizes representative performance results reported in peer-reviewed studies. As shown in Table 1, ResNet-50 provides the strongest overall performance among the compared architectures. Its residual architecture and availability of ImageNet-pretrained weights motivated its selection as the primary classifier in this study.

Table 1. Comparison of CNN architectures for glaucoma detection based on fundus images

Architecture	Dataset	Task	Performance metrics
VGG16	G1020	Classification	ACC=92.1%; AUC=0.94
InceptionV3	G1020	Classification	ACC=94.3%; AUC=0.96
DenseNet121	REFUGE	Classification	ACC=96.0%; Sensitivity=95.1%
ResNet-50	G1020	Classification	ACC=98.48%; Sensitivity=99.30%; AUC=0.97

ACC — accuracy; AUC — area under the ROC curve

In addition to deep CNN architectures, a broad range of classical and specialized machine learning methods has been investigated for ophthalmological diagnosis. Classical models are generally applied to structured clinical variables or handcrafted features, whereas deep learning approaches are capable of learning hierarchical representations directly from fundus images. To provide a broader methodological context, Table 2 summarizes representative model categories and their applications in ophthalmological disease diagnosis.

Table 2. Machine learning models applied in ophthalmological disease diagnosis

Category	Model	Application area
Classical machine learning	Logistic regression	Clinical risk prediction
Classical machine learning	Support vector machine	Glaucoma classification
Classical machine learning	Random forest	EMR-based diagnosis
Deep learning	CNN	Fundus image classification
Deep learning	ResNet-50	Glaucoma screening
Deep learning	MobileNet V2	Lightweight screening models
Specialized models	SLS-Net	Optic disc and cup segmentation
Specialized models	SLSR-Net	Cup-to-disc ratio (CDR)-based glaucoma assessment

Limited availability of labeled medical images and class imbalance remain major challenges in glaucoma-related deep learning research. GANs have therefore been explored as a means of augmenting training datasets. Synthetic images were generated using CycleGAN as described in section 3 [11]. A comparative overview of generative approaches used for retinal image augmentation is shown in Figure 3.

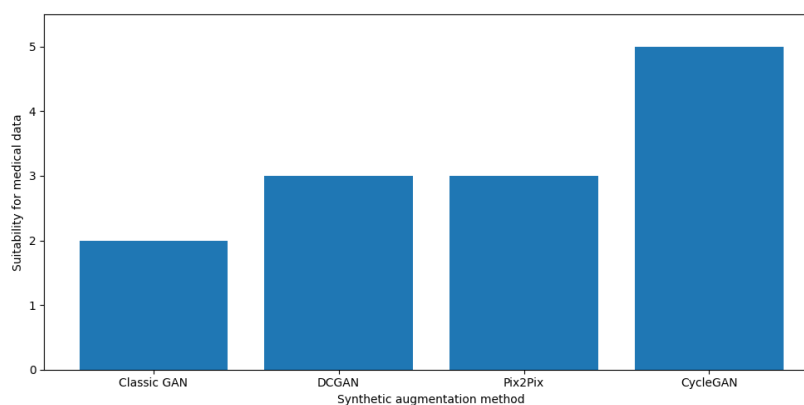


Figure 3. Comparison of GAN approaches for synthetic augmentation of medical images

Various generative modeling techniques have been proposed to mitigate data scarcity and class imbalance in medical imaging. Table 3 compares commonly used synthetic data generation methods in terms of data requirements, image quality, and computational cost.

Table 3. Comparison of synthetic data generation methods for medical image augmentation

Method	Paired data required	Image quality	Computational cost
DCGAN	Yes	Low	Low
CycleGAN	No	Medium–high	Moderate
CUT	No	High	High
Stable siffusion low-rank adaptation (LoRA)	No	Very high	Very high

CycleGAN provides a balanced trade-off between image quality and computational efficiency for medical applications

Recent studies published in the Bulletin of Electrical Engineering and Informatics also demonstrate the effectiveness of deep learning approaches for automated glaucoma diagnosis. Hesham *et al.* [16] proposed a dual deep learning framework for real-time glaucoma detection that combines retinal image classification with optic disc and optic cup analysis. Their results confirm that integrating complementary deep learning components can improve the reliability of glaucoma screening and support the development of practical decision-support systems for ophthalmic applications. More broadly, the development of reliable computational models requires careful consideration of numerical stability, boundary conditions, and the representation of complex spatial processes. Recent studies have demonstrated the importance of these aspects in multidimensional boundary-value problems, finite-element simulation of wave propagation in heterogeneous porous media, and stable difference schemes for complex dynamical systems [17]–[19]. These approaches illustrate the broader importance of mathematically robust and computationally stable modeling strategies.

2.3. Limitations of existing methods

Despite the considerable progress achieved in applying machine learning techniques to glaucoma diagnosis, existing approaches still suffer from several critical limitations that hinder their widespread adoption in real clinical practice. One of the most important challenges is the strong dependence of deep learning models, particularly CNNs, on the quality of input fundus images. Architectures such as ResNet-50, Inception-V3, and DenseNet demonstrate high sensitivity to image resolution, illumination conditions, contrast variations, and acquisition artifacts. In real-world screening scenarios, degradation of clinically relevant anatomical structures—such as the optic disc boundary, neuroretinal rim, and peripapillary vessel patterns—can significantly reduce feature extraction quality and increase the risk of missing early glaucomatous changes [20], [21].

Another major limitation is the requirement for large volumes of high-quality annotated data. The creation of reliable ophthalmic datasets requires specialized imaging equipment, standardized acquisition protocols, and expert labeling by trained ophthalmologists, which is both time-consuming and costly. As a result, publicly available glaucoma datasets are relatively small and often lack diversity in terms of population demographics and imaging devices. This limits model generalization and increases the risk of dataset bias, particularly when models are deployed across different clinical settings [20], [22].

Class imbalance represents an additional challenge inherent to glaucoma screening tasks. In most datasets, images of healthy eyes substantially outnumber glaucomatous samples. Models trained on such imbalanced data tend to achieve high overall accuracy while exhibiting insufficient sensitivity to pathological cases, leading to clinically unacceptable false-negative rates [23].

More recent unpaired translation methods, such as contrastive unpaired translation (CUT), demonstrate improved preservation of local structures but require substantially higher computational resources and careful hyperparameter tuning to ensure training stability [24]. Diffusion-based generative models achieve even higher visual fidelity; however, their application in medical image synthesis remains constrained by significant computational demands and engineering complexity. Even with parameter-efficient adaptation techniques such as LoRA, diffusion models are often impractical for resource-limited clinical environments [25], [26].

Finally, the lack of interpretability and external validation remains a critical barrier to clinical adoption. Although visualization techniques such as Grad-CAM improve transparency by highlighting image regions contributing to model predictions, they do not fully address issues related to model calibration, robustness under domain shift, and long-term clinical reliability [23].

3. METHODS

3.1. Selection of methods, models, architectures, and approaches for problem solving

Comparative studies indicate that residual architectures frequently outperform earlier CNN models such as VGG16 and Inception-V3 in fundus image classification because residual connections improve optimization

stability and representation learning [23]. Strong performance reported on glaucoma benchmarks therefore supports the use of ResNet-50 as a suitable backbone for the present study. At the same time, the acquisition of high-quality ophthalmic images requires specialized equipment, standardized protocols, and expert annotation, which substantially limits the amount of labeled data available for model development [20].

These limitations are further complicated by variations in resolution, illumination, contrast, acquisition hardware, and imaging artifacts [21]. Such variability can negatively affect CNN-based classifiers, particularly when training data are limited. For this reason, the experimental design combines transfer learning with synthetic data augmentation in order to increase training diversity without introducing synthetic samples into the evaluation sets.

To mitigate data scarcity and class imbalance, synthetic data generation using GANs was incorporated into the training pipeline. Multiple studies report that synthetic retinal images can significantly improve classification robustness by increasing training diversity and balancing pathological classes [23].

Although more recent generative approaches such as CUT may provide improved preservation of local structures [21], and diffusion-based models can achieve higher visual realism [25], these methods typically require greater computational resources and more complex optimization procedures.

Model training and experimental evaluation were conducted on a personal workstation equipped with an NVIDIA GeForce RTX 4060 GPU (8 GB VRAM), an AMD Ryzen 5 5600X CPU, 32 GB of RAM (3200 MHz), and an AORUS B550 Elite V2 motherboard.

3.2. Architecture of the proposed model

The proposed classifier is based on ResNet-50 pre-trained on ImageNet. All fundus images were resized to 224×224 pixels and normalized before being provided to the network. Standard training-time augmentation, including rotation, horizontal flipping, and scaling, was applied to improve robustness to moderate appearance variations.

Because the number of labeled ophthalmic images was limited, CycleGAN-generated fundus images were incorporated into the classifier training set. Synthetic samples represented both normal and glaucomatous image domains and were used to increase dataset diversity and reduce class imbalance. Importantly, synthetic images were used exclusively during training, whereas the validation and test subsets contained only real fundus photographs. This design prevented synthetic-data leakage into the evaluation stage and allowed the effect of augmentation on real-image generalization to be assessed more reliably. The overall data processing and training pipeline is illustrated in Figure 4.

After CycleGAN training, 500 synthetic normal and 500 synthetic glaucoma fundus images were added exclusively to the classifier training set. The real validation and test sets were kept unchanged and contained no synthetic images, thereby preventing information leakage and ensuring that the reported performance represented generalization to real fundus photographs. Before each training epoch, real and synthetic samples were jointly shuffled and processed using the same resizing, normalization, and augmentation procedures. The ResNet-50 architecture, optimizer settings, and evaluation protocol were kept identical for the real-only and real-plus-synthetic experiments so that the contribution of synthetic augmentation could be evaluated independently.

3.3. Description of the retinal image database for optic nerve evaluation for deep learning dataset

The dataset is publicly available for research and educational purposes and represents an enhanced version of earlier RIM-ONE databases developed through collaborations between multiple academic institutions [21].

RIM-ONE DL consists of color fundus photographs acquired using ophthalmic cameras, with a particular focus on the ONH region. The dataset includes expert-annotated segmentation masks for both the optic disc and the optic cup, which are essential for structural analysis and computation of clinically relevant indicators such as the CDR.

The images were collected from three clinical centers: Hospital Universitario de Canarias (HUC), Hospital Universitario Miguel Servet (HUMS), and Hospital Clínic San Carlos (HCSC). In the hospital-based split, images from HUC are used exclusively for training, while images from HUMS and HCSC are reserved for testing. This strategy explicitly evaluates cross-center generalization and robustness to domain shifts caused by differences in acquisition devices, imaging protocols, and clinical environments [27], [28].

To ensure reliable evaluation and to prevent data leakage, the RIM-ONE DL dataset provides two official data splitting strategies: a random split and a hospital-based split. The hospital-based strategy is particularly important for assessing cross-domain generalization, as images acquired in one clinic are excluded from the training set and used exclusively for testing. A summary of the dataset splitting strategies used in this work is presented in Table 4.

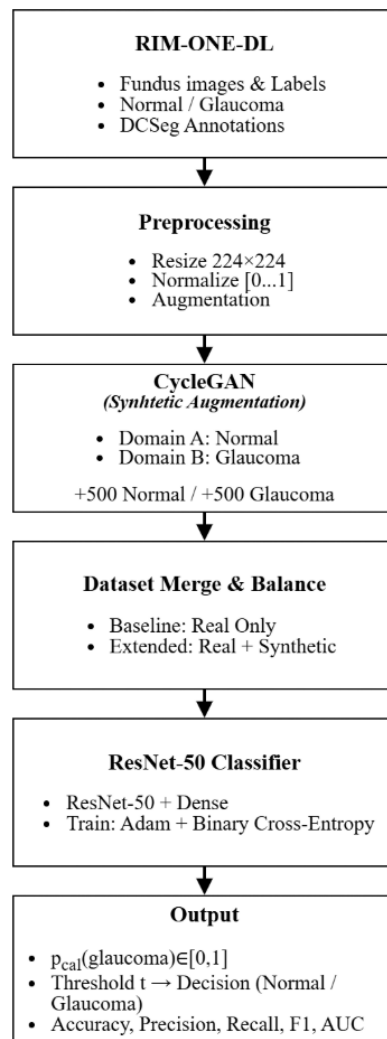


Figure 4. Model architecture

Table 4. Summary of RIM-ONE DL dataset splitting strategies

Splitting strategy	Dataset subset	Purpose
Hospital-based split	Training set (HUC)	Model training on images from a single clinic to evaluate generalization across institutions
Hospital-based split	Test set (HUMS and HCSC)	Independent evaluation on unseen clinical sources to assess domain robustness
Random split (70:30)	Training set	Baseline configuration with randomly sampled images for improved statistical generalization
Random split (70:30)	Test set	Evaluation on randomly held-out data with comparable class distribution

Fundus images in RIM-ONE DL capture fine anatomical details of the optic nerve head, enabling precise analysis of disc morphology, cup size, and vascular structure. Figure 5 illustrates representative examples of normal and glaucomatous fundus images from the dataset.

In normal eyes (Figure 5(a)), the optic disc exhibits well-defined boundaries, a relatively small cup, and a homogeneous coloration. Retinal vessels maintain regular thickness, direction, and branching patterns. In contrast, glaucomatous images (Figure 5(b)) demonstrate characteristic pathological changes, including enlarged optic cup, increased CDR, pallor of the optic disc, and thinning or displacement of retinal vessels. These morphological alterations reflect progressive loss of retinal ganglion cells and degeneration of optic nerve fibers, which ultimately leads to visual field deterioration and irreversible vision loss.

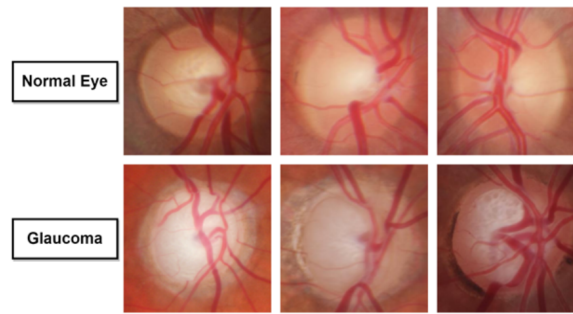


Figure 5. Examples of fundus images: (a) normal eye and (b) glaucomatous eye

The distinction between normal and glaucomatous appearances is not always visually obvious, particularly at early stages of the disease. Subtle structural changes may be difficult to detect even for experienced clinicians, posing significant challenges for automated systems. This highlights the importance of high-quality annotations and motivates the use of synthetic data generation techniques to enrich training datasets and improve sensitivity to early pathological features [29].

Modern segmentation methods rely primarily on CNNs, including U-Net, Mask R-CNN, and related architectures, which have demonstrated superior accuracy and robustness in optic disc and cup segmentation tasks [30].

The RIM-ONE DL dataset used in this study contains a total of 485 color fundus images, including 313 images of normal eyes and 172 images diagnosed with glaucoma. Two evaluation strategies were used to assess model robustness:

- Random split: 70% of the images were used for training and 30% for testing while preserving class distribution.
- Hospital-based split: images from HUC were used for training, while images from HUMS and HCSC were reserved for testing. This protocol evaluates cross-center generalization under different acquisition conditions.

3.4. Preprocessing pipeline

To prepare the dataset for model training, a sequence of standard preprocessing steps was applied. First, all fundus images were resized to a fixed spatial resolution of 224×224 pixels in order to ensure compatibility with the input requirements of deep convolutional neural networks (DCNNs). Second, pixel intensity values were normalized to the [0,1] range, which improves numerical stability during optimization and accelerates convergence.

When segmentation masks were utilized, additional preprocessing was performed to ensure spatial and semantic consistency between the images, masks, and class labels. All masks were resized and aligned with the corresponding fundus images prior to further processing.

3.5. Evaluation metrics and probability calibration

To quantitatively assess the performance of the proposed glaucoma detection model, several standard classification metrics were employed. Let TP , TN , FP , and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Overall classification accuracy was computed as (1):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

Precision, reflecting the reliability of positive predictions, was defined as (2):

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (2)$$

Recall (sensitivity), measuring the ability of the model to correctly identify glaucomatous cases, was calculated as (3):

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (3)$$

To provide a balanced evaluation under class imbalance conditions, the F1-score was used as the harmonic mean of precision and recall (4):

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

In addition to discrimination performance, the calibration quality of probabilistic predictions was explicitly evaluated. Calibration assesses how well predicted probabilities correspond to observed outcome frequencies, which is critical for clinical screening applications.

The Brier score was used to quantify the mean squared error between predicted probabilities and true labels:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2, \quad (5)$$

where $p_i \in [0, 1]$ denotes the predicted probability for sample i , $y_i \in \{0, 1\}$ is the corresponding ground-truth label, and N is the number of test samples.

Expected calibration error (ECE) was computed by partitioning predictions into M equally spaced probability bins:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (6)$$

where B_m denotes the set of samples in bin m , $\text{acc}(B_m)$ is the empirical accuracy, and $\text{conf}(B_m)$ is the mean predicted confidence within the bin.

Reliability diagrams were constructed to visualize calibration performance by comparing predicted probabilities with observed positive fractions across bins. Figure 6 illustrates the reliability curves before and after probability calibration.

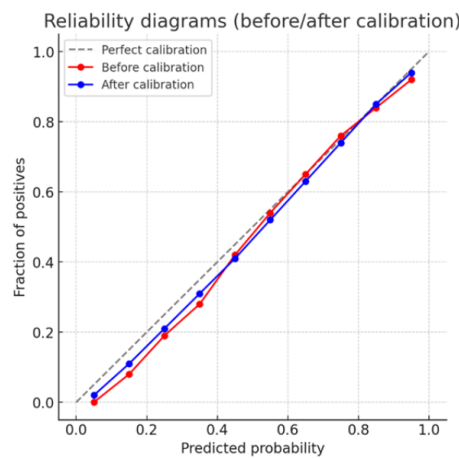


Figure 6. Reliability diagram of the proposed model before and after probability calibration

Prior to calibration, the model exhibited moderate overconfidence at high probability values and underestimation in the mid-probability range. Post-hoc calibration using isotonic regression substantially reduced calibration error, resulting in a reliability curve closely aligned with the diagonal of perfect calibration. To facilitate interpretation for readers from clinical and biomedical backgrounds, the main evaluation metrics are briefly summarized here.

- FID measures the similarity between the distributions of real and generated images in a deep feature space. Lower FID values indicate that synthetic images are closer to real images in terms of statistical characteristics.
- KID is a related metric that estimates the discrepancy between feature distributions using kernel methods and is more reliable for smaller datasets.
- LPIPS evaluates perceptual similarity between images using deep neural network features. Lower LPIPS values indicate that two images are perceptually closer.

- Grad-CAM is an interpretability technique that highlights regions of an image that contribute most strongly to the model's prediction, allowing visual verification that the classifier focuses on clinically relevant structures such as the optic disc and cup.

3.6. Reliability analysis before and after calibration

Prior to calibration, the model exhibited moderate miscalibration. Specifically, predicted probabilities above 0.7 tended to be overconfident, while probabilities in the 0.3–0.5 range were underestimated. The mean Brier score before calibration was 0.182, and the ECE reached 0.071. The corresponding reliability curve deviated noticeably from the diagonal line of perfect calibration, particularly in the high-confidence region.

To address this issue, isotonic regression was applied on the validation set as a post-hoc calibration method, and calibrated probabilities were subsequently evaluated on the test set. After calibration, the Brier score decreased to 0.136 and the ECE was reduced to 0.028. The reliability curve closely approached the ideal diagonal, and the maximum deviation was reduced from 0.19 to 0.08.

Improved calibration enabled more reliable selection of the operating threshold t for screening. For instance, under a target sensitivity of at least 0.90, the calibrated model maintained positive class precision in the range of 0.86–0.88, whereas prior to calibration precision dropped to 0.81–0.83 under comparable conditions.

4. SYNTHETIC DATA GENERATION

4.1. CycleGAN-based synthetic augmentation

To address limited training data and class imbalance in glaucoma screening from fundus photographs, a synthetic augmentation pipeline was implemented to generate additional retinal images. The primary objective of this stage was to increase the diversity of the training distribution, reduce overfitting on small datasets, and improve the generalization capability of the downstream classifier.

CycleGAN was selected as the generative framework because it enables image-to-image translation between two domains without paired supervision, which is particularly important in medical imaging where constructing aligned normal–pathological pairs is typically infeasible. In the present study, the two domains corresponded to fundus images from: i) eyes without glaucoma and ii) eyes with glaucoma-related changes.

4.2. Data preparation for generation

Prior to training the generative model, all images were resized to 256×256 pixels to standardize input resolution. Intensity normalization and mild contrast adjustment were applied to reduce acquisition-related variability (e.g., illumination differences and camera-dependent appearance). The dataset was then partitioned into the *normal* and *glaucoma* domains according to expert-provided diagnostic labels.

4.3. CycleGAN architecture and training configuration

The CycleGAN model consists of two generators and two discriminators trained with adversarial and cycle-consistency objectives. The generators followed a U-Net-style design to facilitate preservation of anatomical structure via multi-scale feature fusion, while discriminators were implemented as PatchGAN classifiers to enforce local realism at the patch level [30], [31]. Optimization was performed using Adam with learning rate $\text{lr} = 1 \times 10^{-5}$ and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Training was conducted for 2000 epochs with batch size 32.

The overall objective combined adversarial and cycle-consistency terms:

$$\mathcal{L} = \mathcal{L}_{\text{adv}}(G_{N \rightarrow G}, D_G) + \mathcal{L}_{\text{adv}}(G_{G \rightarrow N}, D_N) + \lambda \mathcal{L}_{\text{cyc}}(G_{N \rightarrow G}, G_{G \rightarrow N}), \quad (7)$$

where $G_{N \rightarrow G}$ and $G_{G \rightarrow N}$ denote generators translating between the normal (N) and glaucoma (G) domains, D_N and D_G are the corresponding discriminators, and λ controls the cycle-consistency strength. Model training was performed on a workstation equipped with an NVIDIA RTX 4060 GPU, providing stable convergence within the chosen configuration.

4.4. Synthetic dataset construction and qualitative inspection

After training, synthetic samples were generated for both classes, yielding 500 synthetic *normal* images and 500 synthetic *glaucoma* images. The generation process aimed to transfer domain-specific appearance while preserving clinically relevant anatomical structures, including the optic disc boundary and major retinal vessels. Qualitative inspection focused on glaucoma-associated morphological cues commonly used in clinical practice,

such as increased CDR, thinning of the neuroretinal rim, and changes in vessel trajectory around the optic nerve head. Representative synthetic examples are presented in Figures 7 and 8.

Overall, synthetic glaucoma images frequently exhibited enlarged excavation and reduced rim thickness, whereas synthetic normal images preserved sharp disc margins and physiologically consistent vessel patterns.

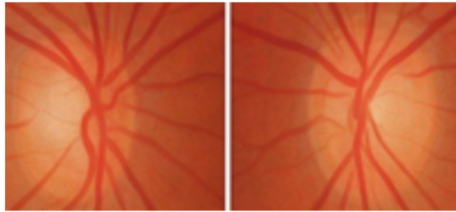


Figure 7. Examples of synthetically generated *normal* fundus images produced by CycleGAN

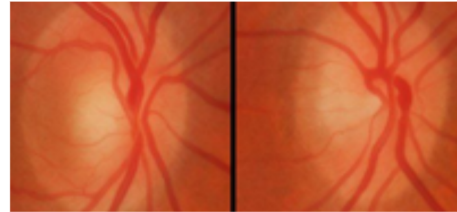


Figure 8. Examples of synthetically generated *glaucoma* fundus images produced by CycleGAN

4.5. Quantitative assessment of synthetic image quality

To complement the qualitative visual inspection, the similarity between real and synthetic fundus image distributions was quantitatively assessed using FID, KID, and LPIPS [32]–[34]. The evaluation was performed separately for the normal and glaucoma classes and additionally for the combined balanced dataset. To ensure a consistent and reproducible assessment of synthetic image quality, the following evaluation protocol was applied.

Evaluation protocol: i) comparison sets: for each class, subsets of real images were compared to the corresponding synthetic subsets; additionally, evaluation was repeated on the combined balanced dataset; ii) preprocessing: identical resizing and normalization were applied to both real and synthetic samples. Potential duplicates and near-duplicates were removed using an embedding-based cosine similarity threshold; iii) FID: computed from Gaussian approximations of deep feature embeddings extracted by a pre-trained Inception-V3 network [32]. This metric was used to quantify the distributional discrepancy between real and synthetic images, with lower FID values indicating greater similarity between the two distributions; iv) KID distance was estimated as an unbiased MMD² measure using a polynomial kernel in the same feature space [29]. KID complements FID by providing a distribution-level similarity estimate that is particularly suitable for evaluations involving relatively small image sets; v) LPIPS was calculated as a perceptual distance between real images and their nearest synthetic counterparts within the same class using deep features extracted from a pre-trained network [35]. This metric provides an additional assessment of whether generated images preserve perceptually relevant structural characteristics of real fundus photographs; and vi) uncertainty: 95% confidence intervals (CI) were obtained using bootstrap resampling with 1000 iterations and were aggregated across three independent generation runs. This procedure was used to account for variability in the generative process and to provide a more robust estimate of synthetic image quality. For FID, KID, and LPIPS, lower values indicate greater similarity between real and synthetic images.

The obtained FID and KID values fell within 25–30 and $(11\text{--}14) \times 10^{-3}$, respectively, as summarized in Table 5, indicating a relatively close match between synthetic and real distributions for fundus imaging. LPIPS values were in the range 0.23–0.25, suggesting preservation of clinically meaningful structural patterns such as optic disc contours and vessel configuration. Synthetic quality for the *normal* class was marginally higher than for the *glaucoma* class, which is plausibly explained by higher variability in glaucomatous morphology.

The obtained FID and KID values fall within the ranges typically reported for medical image synthesis tasks. Importantly, the perceptual similarity measured by LPIPS remained relatively low (0.23–0.25), indicating that key anatomical structures such as optic disc boundaries and vascular patterns were preserved. This suggests that the synthetic images provide meaningful variability for model training rather than introducing unrealistic artifacts.

Table 5. Quantitative evaluation of synthetic image quality using FID and KID metrics

Comparison set	FID ↓	95% FID	KID ($\times 10^{-3}$) ↓	95% KID
Normal: real vs. synthetic	24.8	22.5–27.2	11.3	9.5–13.0
Glaucoma: real vs. synthetic	28.7	26.1–31.4	13.6	11.7–15.5
Combined (balanced dataset)	26.9	25.0–28.8	12.4	11.0–13.8

The quantitative results indicate that the generated images remain reasonably close to the distribution of real fundus photographs. The normal class achieved the lowest FID and KID values, with FID = 24.8 and KID = 11.3×10^{-3} , whereas the glaucoma class produced slightly higher values of 28.7 and 13.6×10^{-3} , respectively. This difference is consistent with the greater morphological variability associated with glaucomatous changes. The combined balanced dataset showed intermediate values, while LPIPS scores in the range of 0.23–0.25 further suggested preservation of important perceptual and anatomical image characteristics.

4.6. Impact of synthetic augmentation on ResNet-50 performance

A central objective of this study was to determine whether CycleGAN-based augmentation improves glaucoma classification by ResNet-50 under limited-data conditions. The classifier was therefore evaluated in two configurations: training on real RIM-ONE DL images only and training on the same real data supplemented with 500 synthetic normal and 500 synthetic glaucoma images. The classifier architecture, optimization settings, preprocessing pipeline, and real-image test set were kept unchanged between experiments.

With real images only, the model achieved an accuracy of 76.8%, precision of 75.5%, recall of 74.2%, F1-score of 74.8%, and AUC-ROC of 0.825. After synthetic augmentation, accuracy increased to 91.4%, precision to 90.2%, recall to 89.7%, F1-score to 89.9%, and AUC-ROC to 0.943. Thus, synthetic augmentation produced improvements across all evaluated classification metrics, with a particularly important increase in sensitivity for glaucoma screening.

The statistical significance of the AUC improvement was evaluated using DeLong's test, while 95% CI for classification metrics were estimated through bootstrap resampling. AUC-ROC increased from 0.825 (95% CI: 0.785–0.861) to 0.943 (95% CI: 0.917–0.965), corresponding to $\Delta\text{AUC} = 0.118$ and $p = 0.0031$. This result indicates that the observed improvement was statistically significant rather than being attributable solely to random variation in the test sample.

Overall, CycleGAN-based synthetic data generation: i) reduces overfitting risk under limited data, ii) increases robustness by enlarging appearance variability, and iii) mitigates class imbalance between normal and glaucomatous images. These factors jointly contribute to the observed improvement in diagnostic performance of the ResNet-50 classifier.

5. USER STUDY: EXPERT ASSESSMENT OF SYNTHETIC FUNDUS IMAGES

5.1. Participants

A total of $n = 12$ participants took part in the expert evaluation study. The cohort included practicing ophthalmologists and residents with varying clinical experience. The sample was gender-balanced (6 female, 6 male). Participants' age ranged from 24 to 35 years (mean ≈ 26.7 years). The reported weekly workload of fundus image assessment varied from approximately 10 to 50+ images.

5.2. Study design and procedure

The study was conducted as a structured questionnaire for expert visual assessment of fundus images. Each participant evaluated:

- Per-image realism/quality using a 10-point ordinal scale (1 — not similar to real fundus images and non-informative; 10 — fully consistent with real clinical images).
- Forced-choice realism comparisons for five paired examples (Variant 1 vs Variant 2), where experts selected the more realistic image and optionally provided reasons (e.g., vessel continuity, optic disc appearance, texture).
- Clinical action recommendation for five examples under a primary glaucoma screening scenario (e.g., sufficient for risk estimation, repeat acquisition due to poor quality, additional tests such as OCT/perimetry, or uncertain/suspicious due to artifacts).

The questionnaire focused on identifying whether synthetic images contain visually detectable artifacts that may hinder clinical interpretation and whether the generated images are plausible enough to support training data augmentation. The expert evaluation was conducted using a structured online questionnaire (Figure 9).

5.3. Outcome measures

The following outcomes were analyzed:

- Distribution of realism scores across evaluated images (median and interquartile range per image).
- Agreement in forced-choice realism selections (majority vote per pair).

- Distribution of recommended clinical actions, emphasizing the proportion of responses requiring re-acquisition or additional examinations.

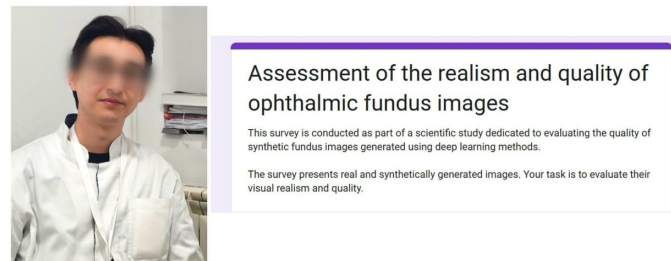


Figure 9. Survey interface used for expert assessment of realism and quality of real and synthetic fundus images

5.4. Key observations (qualitative synthesis)

Expert comments repeatedly highlighted several visual cues associated with synthetic artifacts:

- Vessel discontinuities and unnatural branching, often described as “incorrect vessel trajectory” or disrupted vascular architecture.
- Blurred or “washed-out” optic disc boundaries and insufficient visibility of excavation (cup), which reduces diagnostic informativeness.
- Texture and color inconsistencies compared to typical fundus photographs, including overly smooth regions and atypical contrast.

For multiple examples, experts recommended additional examinations (e.g., OCT or perimetry) or repeat acquisition when image quality was perceived as insufficient, indicating that realism and diagnostic usability are not guaranteed across all synthetic samples.

5.5. Implications for synthetic augmentation

The user study suggests that synthetic augmentation can be promising for dataset expansion; however, expert feedback indicates a need for quality control criteria (artifact screening) before incorporating generated images into model training. In subsequent experiments, only synthetic images passing minimum visual plausibility thresholds are considered to reduce the risk of learning spurious patterns.

6. RESULTS

6.1. Baseline experiment reproduction

The baseline experiment proposed by the authors of the RIM-ONE DL dataset was reproduced using the original data splits and evaluation protocol. The task consisted of binary classification of fundus images into normal and glaucomatous classes. The reproduced results demonstrate a noticeable performance gap compared to the metrics reported in the original publication.

The baseline experiment was performed using only real RIM-ONE DL images and the original evaluation protocol. Table 6 summarizes the resulting classification performance on the real-image test set. The ResNet-50 model achieved an accuracy of 0.768, precision of 0.755, recall of 0.742, F1-score of 0.748, and AUC-ROC of 0.825, demonstrating that the limited real dataset alone was insufficient to reproduce the higher accuracy reported in the original publication.

Table 6. Baseline ResNet-50 performance on RIM-ONE DL (real data only)

Metric	Value
Accuracy	0.768
Precision	0.755
Recall (sensitivity)	0.742
F1-score	0.748
AUC-ROC	0.825

The obtained accuracy did not exceed 80%, which is substantially lower than the results originally reported by the dataset authors.

6.2. Comparison with reported results

A direct comparison between the reproduced accuracy and the accuracy reported in the original RIM-ONE DL publication is presented in Table 7. To compare our results with previously reported findings, Table 7 summarizes the difference between reproduced and reported accuracy values.

Table 7. Accuracy comparison with the original publication

Model	Reported accuracy	Reproduced accuracy
ResNet-50	0.917	0.768

The reproduced ResNet-50 accuracy of 0.768 was lower than the value of 0.917 reported in the reference study. Several factors may contribute to this difference, including the limited number of training images, class imbalance, differences in preprocessing and optimization settings, and sensitivity to the particular data partition. Rather than treating the reported result as directly reproducible under all experimental configurations, the observed performance gap emphasizes the importance of explicitly reporting the training protocol and evaluating strategies for improving robustness under restricted-data conditions.

6.3. Confusion matrix analysis

To further analyze classification errors, a confusion matrix was computed on the test set and is reported in Table 8. The confusion matrix provides additional insight into the baseline classifier errors. Among the glaucomatous cases, 21 images were incorrectly classified as normal, producing a clinically important number of false-negative predictions. In a screening scenario, such errors are particularly undesirable because affected patients may not be referred for additional examination. The observed error pattern therefore supports the need for methods that improve sensitivity to the minority pathological class.

Table 8. Confusion matrix for baseline glaucoma classification

	Predicted normal	Predicted glaucoma
Actual normal	110	28
Actual glaucoma	21	51

6.4. Effect of synthetic data augmentation

Synthetic fundus images generated by CycleGAN were subsequently added to the training set while the real-image evaluation protocol remained unchanged. The expanded training set included an additional 500 normal and 500 glaucomatous synthetic images. As summarized in Table 9, augmentation improved accuracy from 76.8% to 91.4%, precision from 75.5% to 90.2%, recall from 74.2% to 89.7%, F1-score from 74.8% to 89.9%, and AUC-ROC from 0.825 to 0.943.

The improvement was also statistically significant. DeLong's test demonstrated an AUC-ROC increase from 0.825 (95% CI: 0.785–0.861) to 0.943 (95% CI: 0.917–0.965), with $p = 0.0031$. Bootstrap resampling with 5,000 iterations was additionally used to estimate 95% confidence intervals for accuracy, recall, and F1-score. Together, these results provide evidence that the performance gain associated with synthetic augmentation was not limited to a single evaluation metric. The classification performance before and after synthetic data augmentation is summarized in Table 9.

Table 9. Performance comparison before and after synthetic data augmentation

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
Real images only	76.8	75.5	74.2	74.8	0.825
Real+synthetic images	91.4	90.2	89.7	89.9	0.943

The inclusion of synthetic data resulted in a substantial improvement across all evaluation metrics. Accuracy increased by nearly 15 percentage points, while recall improved by more than 15 percentage points, indicating a significant reduction in false negative predictions.

6.5. Summary of findings

Synthetic augmentation substantially improved glaucoma classification on RIM-ONE DL, although external validation on independent datasets such as DRISHTI-GS and ORIGA remains necessary to assess cross-device and cross-center generalizability.

7. DISCUSSION

CycleGAN-based augmentation substantially improved ResNet-50 performance under limited-data conditions, particularly recall and AUC-ROC. The reduction in false negatives is especially important for glaucoma screening, where missed cases may delay treatment.

Qualitative analysis of the generated images suggests that CycleGAN was able to preserve key anatomical structures of the optic nerve head, including the disc boundary, cup morphology, and major vascular patterns. These characteristics are essential for glaucoma assessment and likely contributed to the improved classification performance. Quantitative evaluation using FID, KID, and LPIPS further supports the visual findings, indicating a close distributional and perceptual similarity between real and synthetic images, particularly for the normal class. Slightly higher divergence observed for glaucomatous images can be attributed to the inherently greater morphological variability of pathological cases.

Despite the demonstrated benefits, several limitations should be acknowledged. First, although synthetic data improved overall performance, the generation process may still introduce subtle artifacts, such as vessel smoothing or overly sharp disc boundaries, which do not always correspond to true clinical appearances. While such artifacts were partially mitigated through visual inspection and filtering, their presence may influence model learning. Second, the study focused on a single dataset and a single classifier architecture; therefore, the generalizability of the results to other datasets, imaging devices, or model families remains to be fully validated.

This experimental setup allows preliminary assessment of model robustness under domain shift, which is an important step toward external validation. Future work will include direct evaluation on independent datasets to further confirm generalization performance.

Overall, the discussion confirms that synthetic data augmentation represents a viable and effective strategy for enhancing deep learning-based glaucoma detection in data-constrained scenarios. The results highlight the importance of combining generative modeling with rigorous quantitative and qualitative evaluation to ensure both performance gains and clinical relevance.

8. CONCLUSION

In this study, a glaucoma screening framework combining ResNet-50 classification with CycleGAN-based synthetic data augmentation was developed and evaluated using the RIM-ONE DL dataset. Adding 500 synthetic normal and 500 synthetic glaucoma images to the training set increased classification accuracy from 76.8% to 91.4%, recall from 74.2% to 89.7%, F1-score from 74.8% to 89.9%, and AUC-ROC from 0.825 to 0.943. Statistical testing further confirmed a significant improvement in AUC-ROC.

The results demonstrate that synthetic augmentation can be an effective strategy for mitigating class imbalance and limited training-data availability in automated glaucoma screening. Quantitative image-quality assessment and expert evaluation nevertheless showed that generated images may contain structural artifacts, emphasizing the need for appropriate quality-control procedures before synthetic samples are incorporated into medical training datasets.

The proposed framework should therefore be considered a supplementary decision-support approach rather than a replacement for clinical diagnosis. Future research will focus on independent cross-dataset validation, evaluation with additional classifier and generative architectures, improved explainability and calibration, and assessment of the framework under broader real-world clinical conditions.

FUNDING INFORMATION

The authors did not receive any financial support for the research, authorship, or publication of this article.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Eduard Schiller	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Sabina Rakhmetulayeva	✓	✓		✓						✓		✓	✓	
Anastasiya Tyunina	✓	✓				✓				✓				

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The RIM-ONE DL dataset used in this study is publicly available for research and educational purposes through the official RIM-ONE DL repository maintained by the Medical Image Analysis Group of the University of La Laguna. The repository provides the retinal fundus images, official training and testing partitions, and reference optic disc and optic cup segmentations. The dataset can be accessed at the RIM-ONE DL GitHub repository. The synthetic images generated during the present study and the derived experimental results are available from the corresponding author upon reasonable request [8], [21].

REFERENCES

- [1] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014, doi: 10.1016/j.optha.2014.05.013.
- [2] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," *JAMA*, vol. 311, no. 18, pp. 1901–1911, 2014, doi: 10.1001/jama.2014.3192.
- [3] World Health Organization, *World Report on Vision*. Geneva: World Health Organization, 2019.
- [4] D. S. W. Ting *et al.*, "Artificial intelligence and deep learning in ophthalmology," *British Journal of Ophthalmology*, vol. 103, no. 2, pp. 167–175, 2019, doi: 10.1136/bjophthalmol-2018-313173.
- [5] R. Nunez *et al.*, "Artificial Intelligence to Aid Glaucoma Diagnosis and Monitoring: State of the Art and New Directions," *Photonics*, vol. 9, no. 11, 2022, doi: 10.3390/photonics9110810.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [7] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, T. N. Rao, and U. R. Acharya, "Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images," *Information Sciences*, vol. 441, pp. 41–49, 2018, doi: 10.1016/j.ins.2018.02.051.
- [8] F. Fumero, S. Alayón, J. L. Sánchez, J. Sigut, and M. González-Hernández, "RIM-ONE: An open retinal image database for optic nerve evaluation," in *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, 2011, pp. 1–6, doi: 10.1109/CBMS.2011.5999143.
- [9] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [10] P. Costa *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2018, doi: 10.1109/TMI.2017.2759102.
- [11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232, doi: 10.1109/ICCV.2017.244.
- [12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE International Conference on Computer Vision (ICCV)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
- [14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [15] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, and J. Liu, "Joint optic disc and cup segmentation based on multi-label

- deep network and polar transformation,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1597–1605, 2018, doi: 10.1109/TMI.2018.2791488.
- [16] M. Hesham, G. Kareem, and M. Hadhoud, “Enhanced real-time glaucoma diagnosis: dual deep learning approach,” *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 3, pp. 1846–1857, 2025, doi: 10.11591/eei.v14i3.8495.
- [17] Z. Baishemirov, A. Berdyshev, and A. Ryskan, “A Solution of a Boundary Value Problem with Mixed Conditions for a Four-Dimensional Degenerate Elliptic Equation,” *Mathematics*, vol. 10, no. 7, 2022, doi: 10.3390/math10071094.
- [18] M. Nurtas, Z. Baishemirov, S. Alpar, and F. Tokmukhamedova, “Numerical simulation of wave propagation in mixed porous media using finite element method,” *Journal of Theoretical and Applied Information Technology*, vol. 99, pp. 4163–4172, 2021.
- [19] A. Berdyshev, R. Aloev, D. Bliyeva, S. Dadabayev, and Z. Baishemirov, “Stability Analysis of an Upwind Difference Splitting Scheme for Two-Dimensional Saint–Venant Equations,” *Symmetry*, vol. 14, no. 10, 2022, doi: 10.3390/sym14101986.
- [20] T. Krzywicki, P. Brona, A. M. Zbrzezny, and A. E. Grzybowski, “A Global Review of Publicly Available Datasets Containing Fundus Images: Characteristics, Barriers to Access, Usability, and Generalizability,” *Journal of Clinical Medicine*, vol. 12, no. 10, art. 3587, 2023, doi: 10.3390/jcm12103587.
- [21] J. Camara, R. Rezende, I. M. Pires, and A. Cunha, “Retinal Glaucoma Public Datasets: What Do We Have and What Is Missing?,” *Journal of Clinical Medicine*, vol. 11, no. 13, art. 3850, 2022, doi: 10.3390/jcm11133850.
- [22] F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayon, R. Arnay, and D. Angel-Pereira, “RIM-ONE DL: A unified retinal image database for assessing glaucoma using deep learning,” *Image Analysis & Stereology*, vol. 39, no. 3, pp. 161–167, 2020, doi: 10.5566/ias.2346.
- [23] M. Ashtari-Majlan, M. M. Dehshibi, and D. Masip, “Glaucoma diagnosis in the era of deep learning: A survey,” *Expert Systems with Applications*, vol. 256, 2024, doi: 10.1016/j.eswa.2024.124888.
- [24] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in *The European Conference on Computer Vision (ECCV)*, 2020, pp. 319–345, doi: 10.1007/978-3-030-58536-5_19.
- [25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695, doi: 10.1109/CVPR52688.2022.01042.
- [26] E. J. Hu *et al.*, “LoRA: Low-rank adaptation of large language models,” *arXiv preprint*, 2021, doi: 10.48550/arXiv.2106.09685.
- [27] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.
- [28] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs,” *PLOS Medicine*, vol. 15, no. 11, p. e1002683, 2018, doi: 10.1371/journal.pmed.1002683.
- [29] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomputing*, vol. 321, pp. 321–331, 2018, doi: 10.1016/j.neucom.2018.09.013.
- [30] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
- [31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976, doi: 10.1109/CVPR.2017.632.
- [32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [33] M. Bińkowski, D. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.
- [35] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

BIOGRAPHIES OF AUTHORS



Eduard Schiller    graduated from the Faculty of Engineering and Information Technology of the Kazakh-German University (Deutsch-Kasachische Universität) with a degree in Telematics in 2024. During his studies, he actively participated in academic projects and was recognized for his contributions to the student council. In 2024, he enrolled in a master's degree program in Machine Learning and Data Science at Satbayev University in Almaty at the Faculty of Information Technology. His professional experience includes working as a developer-analyst at proper data management (PDM) since October 2023 and completing an internship at bank soft systems (BSS), where he focused on software development, data storage optimization, and data visualization. He possesses strong skills in programming (Python, Java, JavaScript, C, and C++), database management (MySQL, PostgreSQL, Greenplum, Oracle), and the deployment of modern IT infrastructures (Linux, Kubernetes, Docker, Apache Airflow, and CI/CD). He is fluent in Russian (native), English, and German, and has good proficiency in Kazakh. He can be contacted at email: c2238@mail.ru.



Sabina Rakhmetulayeva    graduated with honors from the Faculty of Physical and Mathematical Sciences of Abai Kazakh National Pedagogical University in 2006. She completed her master's degree in Computer Science at the same university in 2008 and entered the doctoral program in Information Systems at Narxoz University. From 2009 to 2015, she worked as a senior lecturer at the Department of Applied Informatics at Narxoz University. In 2015, she successfully defended her Ph.D. dissertation in the dissertation council at the Eurasian National University in the specialty "6D070300-Information Systems," earning her Ph.D. degree. In addition to her teaching activities, she actively engages in research. Her scholarly work has resulted in over 30 publications, including 2 monographs and 10 articles in international peer-reviewed scientific journals classified within the first, second, and third quartiles according to Journal Citation Reports by Clarivate Analytics. She also has published more than 20 articles in international peer-reviewed scientific journals indexed in Scopus with a CiteScore percentile of at least 35. She can be contacted at email: ssrakhmetulayeva@gmail.com.



Anastassiya Tyunina    graduated from the Faculty of Engineering and Information Technology of the Kazakh-German University (Deutsch-Kasachische Universitaet) with a degree in Telematics in 2024. In 2024, she enrolled in a scientific and pedagogical master's degree at the Satpayev Research University in Almaty at the Faculty of Computer Engineering, specializing in machine learning and data science. She can be contacted at email: st.demiu@bk.ru.