International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I12P114

Volume 73 Issue 12, 174-184, December 2025
© 2025 Seventh Sense Research Group®

Original Article

Enhancing Automated Glaucoma Detection: A Lightweight Hybrid Model with EfficientNetB3, CBAM, and Vision Transformers

Malla Sireesha[1], Meka James Stephen[2], P.V.G.D. Prasad Reddy[3]

[1]*Department of Information Technology and Computer Applications, Andhra University, Visakhapatnam, Andhra Pradesh, India.*
[2]*Dr. B.R. Ambedkar Chair, Andhra University, Visakhapatnam, Andhra Pradesh, India.*
[3]*Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.*

[1]*Corresponding Author : mallasireesha72@gmail.com*

Received: 12 August 2025 Revised: 20 November 2025 Accepted: 25 November 2025 Published: 19 December 2025

Abstract - *Glaucoma is one of the leading causes of permanent blindness globally, and has continued over the years as a dangerous eye condition that slowly damages the optic nerve. It is usually associated with high pressure in the eyes and can be affected by several factors, including age, family history, diabetes, and high blood pressure. Common tests such as tonometry, perimetry, and optical imaging are used to check for glaucoma, but they often find it difficult to detect the disease in its early stages, making early treatment harder. Recent advances in Deep Learning have resulted in several automated methods for detecting diseases from fundus images. There are many existing models that suffer from limitations such as a lack of clarity, generalization issues across datasets, and poor accuracy during times of confusion. The proposed EfficientNetB0 and Transformer architecture performed very well in automatically detecting glaucoma using normal fundus images. This architecture included an Explainable AI (XAI) method, which allows the decision-making process to be visually represented and enhances model accessibility by expanding the proposed model with EfficientNetB3 in place of EfficientNetB0, Vision Transformer (ViT), along with including the Convolutional Block Attention Model (CBAM), while working with a lightweight fundus image dataset. The Hybrid Deep Learning Model achieved the best AUC of 0.98 compared to all existing models.*

Keywords - *Lightweight Fundus Images, XAI, Grad-CAM, EfficientNetB3, CBAM, ViT.*

1. Introduction

Glaucoma is a vision-threatening disease that affects millions of people worldwide. The Optic Nerve (ON) transmits messages to the brain and is frequently damaged because of high pressure in the eyes. Since it does not show clear signs at first, people often do not realize they have it until their vision is already affected. Over 76 million individuals worldwide had been affected as of 2020, and by 2040, that figure will likely increase to over 111 million, making it the second most common factor in permanent blindness [1].

The main categories of glaucoma are Primary Open-Angle Glaucoma (POAG), which represents about 90% of all cases, Angle-Closure Glaucoma (ACG), Normal-Tension Glaucoma, and Secondary Glaucoma [2]. In the middle of each eye, there is a small bright round spot called the optic disc, which is shown in Figure 1 below. This is where tiny nerve fibers from the back of the eye join to form the ON, which carries messages from the eye to the brain, helping to visualize. In retinal images, the Optic Disc (OD) is usually the whitest spot, with blood vessels spreading out from it, as seen clearly in both the healthy eye (left) and the glaucoma-affected eye (right). The optic cup is a light-colored depression located at the center of the OD. In a normal or healthy eye (shown on the left), the central area of the ON the cup shows relatively small. But in an eye with glaucoma, some of these fibers are lost. As a result, the cup located at the center of the OD appears enlarged. This change is called cupping, and it is something doctors look for when checking for glaucoma.

While this can be done manually by examining images like this one, large-scale screening is time-consuming and leads to human error. That is why computers are used to detect and examine the optic disc and cup by looking at brightness, intensity, blood vessel patterns, and shapes, making the process faster and easier [3]. Doctors use many kinds of tests to detect glaucoma. Usually, they start with a test called visual field testing to check your line of sight, a procedure to measure the drainage position, and a tonometer to measure the level of pressure inside your eye.

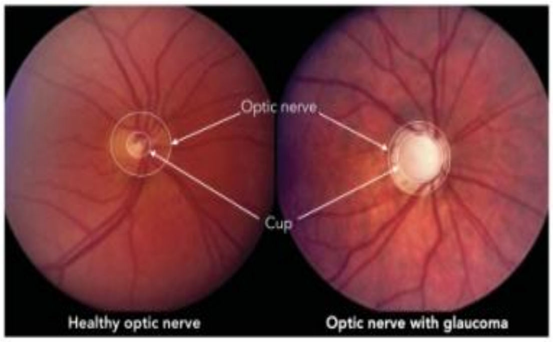

Fig. 1 Characteristics of fundus image

They may also take detailed images of your retina using specialized cameras (fundus photography) and scanning laser ophthalmoscopy to get a closer look at your optic nerve [4, 5]. These methods work well, but they can have trouble spotting the disease early, and they often depend a lot on the doctor's experience. Over the past decade, several models have been developed for the automated detection of glaucoma. Earlier methods mainly worked on machine learning classifiers, which were trained using manually designed features obtained from the OC and OD regions [6]. These models typically performed binary classification (glaucoma vs. normal) and achieved moderate accuracy, but encountered difficulties with manual feature dependency, generalization, and fluctuations in image quality. The field witnessed an evolution with the start of Convolutional Neural Networks (CNNs). One ResNet-50 model scored accuracy, sensitivity, and AUC [7].

The forecasting had been further improved by other models such as VGG19, Inception-ResNet-V2, and AG-CNN, but external validation efficiency was still slightly lower [8]. The primary purpose of these models was to perform binary classification, which distinguishes between glaucomatous and non-glaucomatous images. Minimal, low, and severe glaucoma have been categorized in a few attempts at multi-class classification; however, these efforts remain limited by both the dataset imbalance and the confusion in disease progression instances [9]. In this paper, a glaucoma classification model for normal fundus images is presented based on EfficientNetB0, a Transformer Architecture, and Explainable AI (XAI) techniques for improved interpretability. This arrangement required much work, despite its good performance. To optimize the approach for greater efficiency, we modified it for light fundus images by using EfficientNetB3 instead of EfficientNetB0 and incorporating the Vision Transformer (ViT) model. Additionally, we added a Convolutional Block Attention Module (CBAM) to enable the model to focus on significant areas of the retinal surface. CNN and Transformer models have improved glaucoma detection, but challenges remain. Many approaches rely on binary classification, making it challenging to differentiate between early, moderate, and severe stages of the disease. These models are not feasible for large-scale clinical application or real-time diagnosis due to their high computational resource requirements. Furthermore, their predictions are often opaque, as deep learning systems frequently function as black boxes, providing little understanding of how decisions are reached. To address these issues, this study proposes a lightweight hybrid deep learning framework combining EfficientNetB3, ViT, and CBAM for automated glaucoma detection. This model stands out because it merges convolutional layers for local feature extraction with transformer-based global attention and an attention mechanism that targets key retinal regions. This integration enhances diagnostic accuracy, increases transparency through XAI visualizations, and keeps computational requirements low, making it well-suited for practical use and resource-constrained medical settings.

2. Related Work

There are several publicly available eye image datasets that include clear retinal images with expert labels, which are useful for training and testing computer models to detect glaucoma. The Drishti dataset contains 101 high-resolution color fundus images (around 2896 × 1944 pixels) that clearly show the OD and OC areas, with annotations for their boundaries, soft segmentation maps, and cup-to-disc ratio provided by experts. In the study, CNN models that combine shape and texture details were used to analyze these regions [9]. The RIM-ONE-v1 dataset includes about 455 fundus images, featuring a mix of healthy eyes and eyes affected by glaucoma. In each case, trained experts manually traced the OD and OC to represent their shapes accurately. The RIM-ONE DL dataset is another collection of eye images like this. It has 313 normal eye images and 172 glaucoma images, and each one also includes expert markings of the disc and cup. A step-by-step deep learning model using curriculum learning was applied to RIM-ONE-v1 for measuring shapes and sizes in these images [10]. The EyePACS-AIROGS-light-V2 dataset is a smaller subset of the large Rotterdam EyePACS AIROGS dataset. This large dataset contains about 113,893 fundus images from over 60,000 subjects collected from many screening sites, while the light version has around 4,000 training images and about 385 images each for validation and testing, all resized to 512 × 512 pixels. Models such as the SwinTransformer and ConvNeXtTiny from the Kaggle Benchmark Study were used on this dataset [11, 12].

The AFIO dataset provides fundus images with detailed biological region information, including the retinal ganglion cell layer and the complex of the ganglion cell and inner plexiform layers. A CNN model combined with features from the retinal ganglion cell layer was used for analysis [13]. The Rotterdam EyePACS AIROGS dataset, mentioned earlier, merges images from the Rotterdam dataset and EyePACS AIROGS, offering varied image quality and optic disc appearances. The GARDNet multi-view CNN with advanced image preprocessing was applied to this dataset. Finally, the RIM-ONE DL dataset and RIM-ONE-r3 dataset both focus on disc and cup segmentation; RIM-ONE-r3 has 159 retinal images, with 84 from healthy individuals and 74 from glaucoma patients, including their computed cup-to-disc ratios. These datasets, along with others such as G1020 (1,020 images with glaucoma diagnosis, disc/cup segmentation, and neuroretinal rim measurements), LAG (5,824 images with glaucoma labels and attention maps), REFUGE (1,200 images with disc/cup segmentation and glaucoma labels), ONHSD (~100 images), Drions DB (~110 images), ORIGA (~650 images), and RIGA (~750 images), form a rich variety of publicly available data for glaucoma research. The RIM-ONE DL dataset was again used with GARDNet and contrast enhancement in deep learning experiments [14]. Perdomo et al. suggested a glaucoma diagnosis framework using fundus images, where a DCNN first segments the OD and OC, estimating morphometric features such as disc and cup areas, perimeters, axis lengths, and CDR-based ratios. These features are fed to an MLP classifier that categorizes images into healthy, suspicious, or glaucomatous cases. The approach has been tested on various datasets, including RIM-ONE and DRISHTI-GS1, with strong performance: about 89.4% accuracy on RIM-ONE_v1 and 0.82 AUC on DRISHTI-GS1; this outperforms earlier single-stage CNN approaches. However, the approach is limited by its dependence on high-quality images and accurate segmentations, a relatively small training dataset affecting generalization, and sensitivity to dataset variations. At the same time, the multi-stage pipeline adds complexity and can struggle in borderline “suspicious” cases [15]. Ajitha et al. developed a CNN-based model for automatic detection of glaucoma from fundus images, trained on labeled retinal photographs to classify eyes as glaucomatous or non-glaucomatous. The model achieved strong performance and supports the use of deep learning for automated glaucoma screening; it avoids manual optic disc/cup segmentation by learning features directly from raw images. However, this approach is limited by a small and narrow dataset affecting generalization, a lack of early-stage or longitudinal data for pre-clinical detection, and low interpretability of deep features, which might reduce clinical trust [16]. About these existing works explained through the below Table 1:

Table 1. Various existing works using different datasets and models with their respective results

Paper Reference	Proposed work	Model(s) Used	Results	Limitations
[10] Pathan et al.	Combined shape and texture details for detecting glaucoma	CNN	Improved accuracy and dependability	Works well on Drishti, but not tested on other datasets, so performance on varied image quality is unknown.
[11] Raja et al.	CNN with retinal ganglion cell layer features for detection and severity	CNN + RGC features	F1-score 0.9577	Dependent on high-quality scans that capture the retinal ganglion layer; not suitable for low-resolution images
[12] Gavin D’Souza et al.	Swin Transformer for glaucoma detection	Swin Transformer	AUC 0.929	Very high processing power is needed, which limits use in low-resource clinics.
[13] Alsulami et al.	Hybrid DCGAN + DenseNet-based CNN for augmentation and classification	DCGAN + DenseNet CNN	Performance improved over baseline	GAN-based augmentation can create unrealistic samples, possibly misleading the classifier.
[14] Kaggle Benchmark Study	Tested multiple models on EyePACS-AIROGS-light-V2	ConvNeXtTiny	94.94% accuracy	Only tested on one dataset; may not handle images from different cameras or settings.
[15] Perdomo et al.	Curriculum learning-based model to measure shapes and sizes	Step-by-step deep learning	89.4% accuracy	Accuracy may drop on larger or more diverse datasets; the method was only tested on one dataset.
[16] Ajitha et al.	13-layer CNN with SoftMax and SVM classifiers	CNN + SoftMax + SVM	95.61% accuracy, 89.58% sensitivity	High accuracy reported, but no tests on external datasets to check generalization.
[17] Saha et al.	YOLO-MobileNet to detect the optic disc	YOLO-MobileNet	97.4% accuracy	Focuses only on optic disc detection; complete glaucoma classification still depends on extra processing.

[18] Afolabi et al.	Lightweight U-Net for segmentation + XGBoost for classification	U-Net + XGBoost	88.6% accuracy	Accuracy is lower than newer deep learning methods; it may miss finer texture features.
[19] Aljohani et al.	Hybrid ResNet50 + VGG-16 + Random Forest	ResNet50, VGG-16, RF	95.41% accuracy	Needs more processing due to multiple models; no tests on unseen datasets
[20] Islam et al.	EfficientNet on zoomed-in optic disc and cup	EfficientNet	96.52% accuracy	Requires very precise disc–cup localization; performance may suffer if segmentation is inaccurate.
[21] Juneja et al.	Two CNNs focusing separately on disc and cup	CNNs	Dice score: 95.8% (disc), 93% (cup)	Accuracy depends heavily on having both regions clearly visible and correctly cropped.
[22] Shoukat et al.	ResNet-50 on gray-channel images with data augmentation	ResNet-50	Accuracy improved with augmentation	Needs careful augmentation tuning; may still fail on poor-quality images
[23] Al Mahrooqi et al.	GARDNet multi-view CNN with preprocessing	Multi-view CNN	AUC 0.92 (Rotterdam), 0.9308 (RIM-ONE DL)	Heavy preprocessing and multiple CNNs make it slow for real-time use
[24] D'Souza et al.	AlterNet-K with MSA + ResNet layers	AlterNet-K	91.6% accuracy AUROC 0.968, F1 0.915	MSA sometimes focuses on irrelevant regions, making predictions harder to interpret

3. Methodology

3.1. About Dataset

The EyePACS-AIROGS-light-V2 dataset has color photos of the retina, which were taken from Kaggle, which was publicly available (https://www.kaggle.com/datasets/deathtrooper/glaucoma-dataset-eyepacs-airogs-light-v2) [25]. Each image is labeled as either Referable Glaucoma (RG), which means it might need a doctor's attention, or Non-Referable Glaucoma (NRG), which means it likely does not. There are about 4,000 images for training and around 385 images each for validation and testing. This gives a good approach for building and testing deep learning models. When researchers use this dataset, they usually apply several image-preprocessing steps so that the images are clean, focused, and uniform before training.

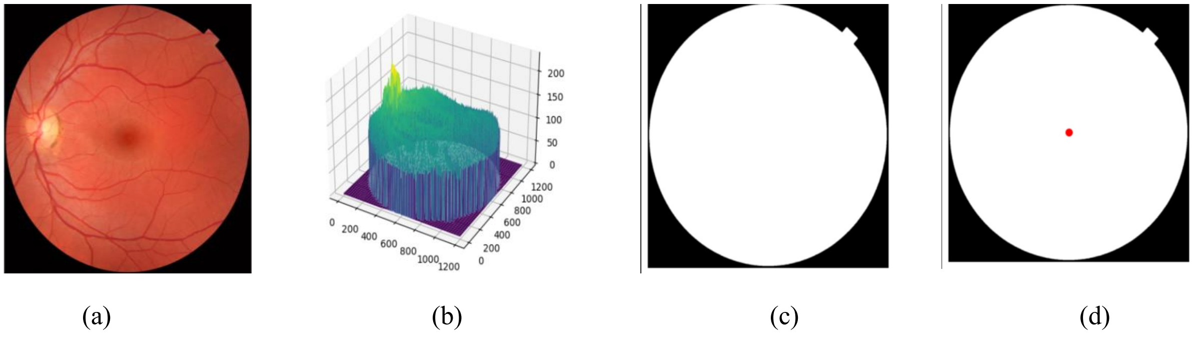

Fig. 2 First 4 preprocessing steps

The above Figure 2 follows steps 1 to 4, which are Original Image (a), Visualize Intensity Map in 3D (b), Approximate the Segmentation of the Background and Foreground (c), and Approximate Center of Segmentation Foreground (d).

Step 1: Original Image.

The raw fundus image, exactly as it was captured by the camera. The image might have extra background, uneven lighting, or be different in size, depending on how it was taken.

Step 2: Visualize the Intensity Map in 3D.

A 3D intensity map is made. This is like a height map showing the brightness of each pixel. It shows which parts of the image are bright, like the optic disc, and which parts are dark, like the background.

Step 3: Approximate the Segmentation of the Background and Foreground.

The image is roughly divided into two parts: the retina, which is the main part of the eye, and the background, which is everything outside the eye. This step removes black edges or camera borders, keeping only the important part-the retina.

Step 4: Approximate Center of Segmentation Foreground.

The center of the retina is estimated. This helps position the image properly, so the important area is in the middle.

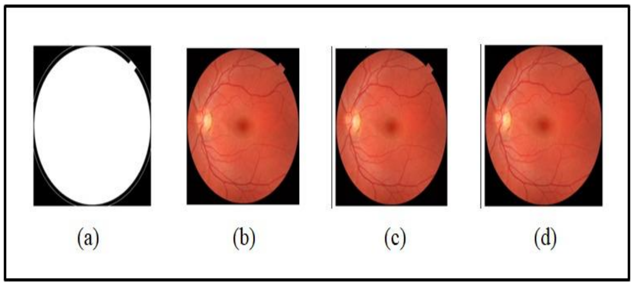

Fig. 3 Next 4 preprocessing steps

Figure 3 represents the remaining steps, which are Approximate Radius of Segmentation Foreground from Radius (a), Crop the Minimal Bounding Square Around the Foreground (b), Resize to Desired Dimensions (ex., 512x512) (c), and Apply Mask (d).

Step 5: Approximate Radius of Segmentation Foreground from Radius.

The size of the retina is measured by estimating its radius. This ensures that cropping and scaling maintain the correct eye proportions.

Step 6: Crop the Minimal Bounding Square Around the Foreground.

Using the center and radius, the smallest possible square that contains the whole retina is cropped. This helps remove space around the eye and zooms in to focus on the eye area.

Step 7: Resize to Desired Dimensions (ex., 512x512).

The image is cut and resized to a fixed size, typically 512 × 512 pixels, ensuring all images appear uniform. This helps computer models work with them more easily.

Step 8: Apply Mask.

A mask is applied to the image, allowing only the retina to be visible. The corners and extra background are made black, leaving a clear, round image of the eye.

4. Proposed Architecture

Figure 4 below represents the proposed architecture, which begins with an input retinal fundus image, which may represent either a normal eye or one affected by glaucoma. This hybrid deep learning model combines EfficientNet-B0, a pretrained image analysis network, with a Transformer Encoder to improve fundus image interpretation. EfficientNet-B0 extracts key visual details from the images, while the Transformer captures the relationships between different parts of the image. By working together, the model learns both fine details and overall patterns, making it better at understanding complex retinal structures and supporting accurate disease detection. The final classifier features adaptive average pooling and a fully connected layer with two output nodes for binary classification.

After obtaining the classification result, which integrated XAI using Grad CAM to visualize the model's decision-making process. XAI is a technique that helps make machine learning models understandable to humans by explaining why a model makes a certain prediction. This is particularly important in fields such as healthcare, where doctors must verify AI results before using them for diagnosis or treatment. XAI provides several methods to explain model behavior, including visualization-based techniques such as Grad-CAM, Grad-CAM++, Saliency Maps, and Occlusion Sensitivity, as well as feature attribution methods like LIME and SHAP, and surrogate models that approximate complex models with simpler interpretable ones.

Used Grad-CAM instead of other methods like Saliency Maps, LIME, or SHAP because Grad-CAM works faster with CNN models, gives clear and focused heatmaps, and is great for analyzing medical images. Saliency Maps can be noisy, and LIME and SHAP are better suited for textual or tabular data, rather than detailed images like eye scans. In the process of detecting glaucoma using a Hybrid CNN, we applied XAI because a simple output like "Glaucoma" or "Normal" is not sufficient for medical decision-making. For better understanding, see which part of the retina the model focused on for the prediction. Grad-CAM is one of the XAI techniques that is chosen because it generates class-specific heatmaps that graphically highlight the image's most important areas.

Grad-CAM calculates how the predicted class changes by measuring the gradients on the last convolutional feature maps, applies global average pooling to generate importance weights, and produces a class activation heatmap that is overlaid on the original image. The red/yellow area, also called the "hot spot," is primarily concentrated at the top region, and it focuses directly on the optic disc area (the tail pointing to the OD). Yellow and Red denote the areas that the model considers most important for glaucoma class prediction, and green indicates less important regions. The model's focus on medically important areas has been verified by the fact that the highlighted areas in glaucoma-positive images were mostly located around the optic disc. By integrating XAI, the proposed model improves trust and confidence for medical applications by producing both clear visual explanations and precise guidance. By adding Grad-CAM to the model, it is very easy to understand the regions that impacted the model's predicted outcome, improving clinical accuracy. This ensures that decisions are based on medically relevant areas, such as the optic disc for glaucoma detection, increasing trust and confidence in AI-assisted diagnosis. Furthermore, these visual explanations support educational purposes, help in model validation and error analysis, and bridge the gap between automated predictions and clinical decisions, making the system more suitable for practical healthcare applications.

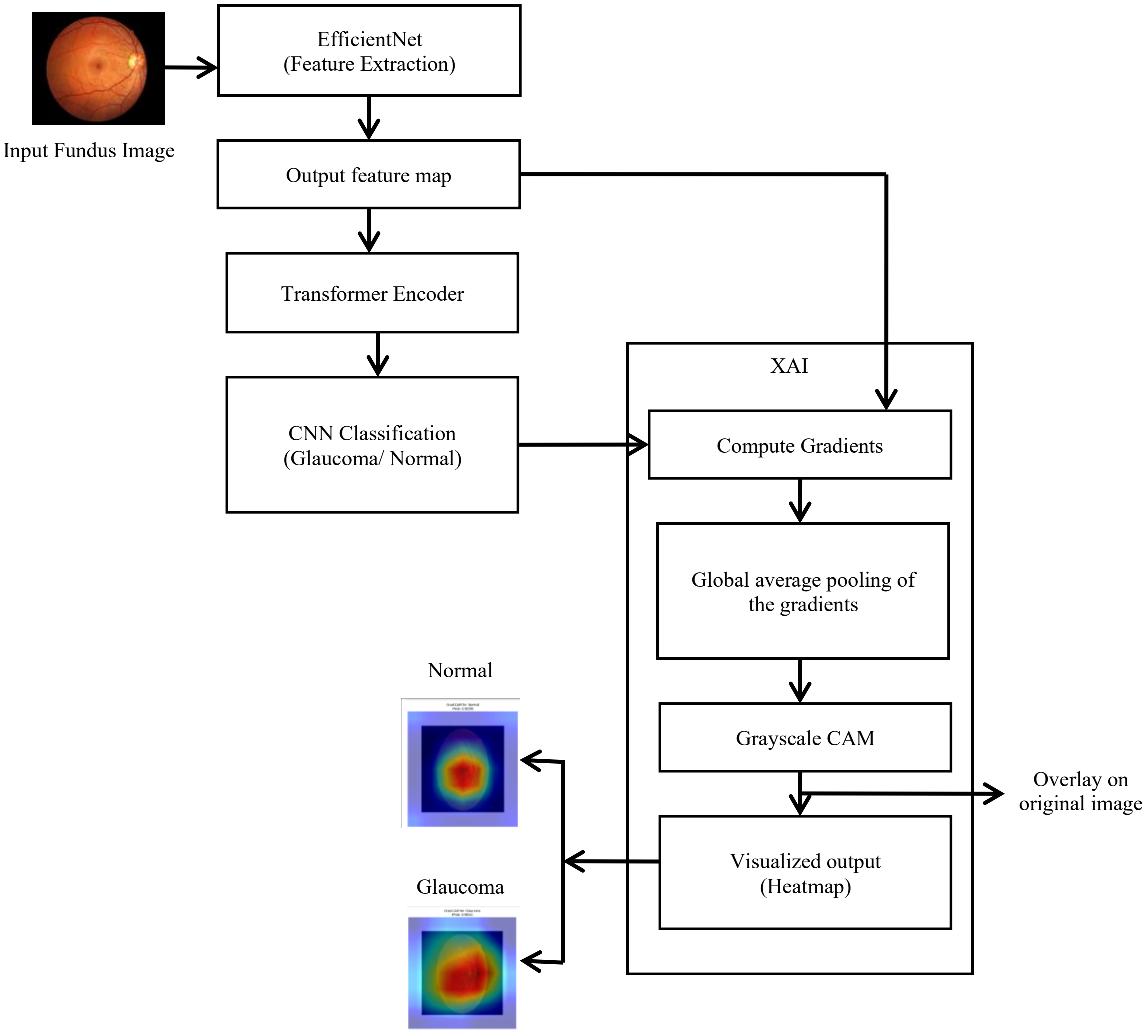

Fig. 4 Proposed model architecture

The proposed architecture above works on a standard fundus image dataset. Here, not only normal fundus images, but also lightweight fundus images with a Hybrid Deep Learning Model. Lightweight fundus images are reduced or optimized versions of standard fundus images that maintain important retinal details while minimizing storage needs and allowing quicker processing. They work by minimizing image size or complexity while retaining the essential details required for accurate analysis. They are used instead of regular images to save storage space and improve device performance, especially in situations with limited memory or processing power. This study examines how the dataset affects model performance. In the Hybrid Deep Learning Model, EfficientNetB0 and EfficientNetB3 are used because EfficientNetB3 is larger, scaled up in depth and width. This model has more channels and deeper blocks with a greater number of parameters.

5. Hybrid Architecture - EVC (EfficientNetB3 + ViT + CBAM)

The proposed hybrid model integrates three main components: Vision Transformer (ViT) for learning global relationships, CBAM for emphasizing significant features, and EfficientNetB3 for extracting spatial characteristics. This combination uses the strengths of both CNN and Transformer architectures to achieve higher detection accuracy and better model explainability. Figure 5 below represents a well-structured hybrid deep learning model for glaucoma detection that uses a new dataset called EyePACS-AIROGS-light-V2, which is available on Kaggle. This Hybrid Model is a combination of EfficientNetB3, Convolutional Block Attention Model (CBAM), and Vision Transformer (ViT). The proposed architecture shows that it takes an eye fundus image from the dataset as input, performs preprocessing techniques like resizing all images to 224x224 and normalizing each image, and then returns the image and its label.

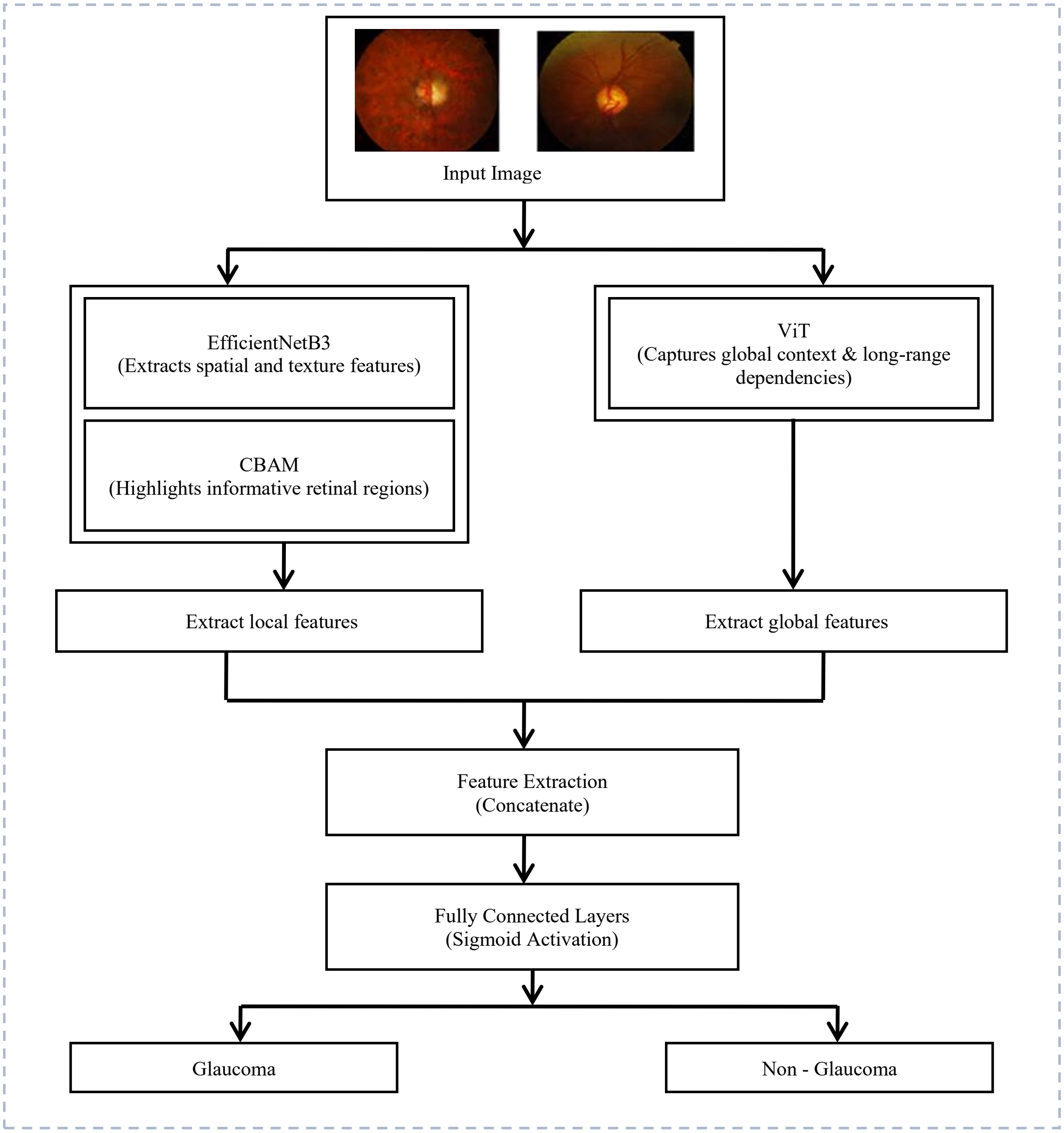

Fig. 5 EVC hybrid architecture

Those preprocessed images travel through the first model, EfficientNetB3, built using multiple MBConv(Mobile Inverted Bottleneck Convolution) blocks with Squeeze-and-Excitation (SE) modules for channel-wise feature recalibration. It follows the compound scaling principle, which increases network depth (by adding more MBConv layers), width (by increasing the number of channels), and input resolution (from 224×224 to 300×300 for B3) in a balanced manner. The model starts with basic convolution to process the image. Then it adds several specialized blocks (MBConv1 and MBConv6) that leverage efficient approaches to reducing computation cost; breaking up convolutions, emphasizing features (SE blocks), and information reuse (skip connections). Ultimately, a small 1×1 convolution is performed, all the features are averaged, and a final prediction is obtained through a fully connected layer. The B3 model is envisioned with a significantly higher number of filters and image resolution, enabling the identification of finer details through an efficient process, compared to the B0 model.

The CBAM is an additional module that draws attention in two stages: channel attention and spatial attention. In the CAM, the feature map obtained from a CNN block undergoes Global average and max pooling to generate two descriptors, which are fed into a shared MLP. The model first utilizes a sigmoid function to highlight important features across channels, a technique known as channel attention. Then, the Spatial Attention Module (SAM) finds key areas in the image by combining average and max pooling, applying a 7×7 convolution, and using another sigmoid to create spatial attention. These attention maps are multiplied with the input to help the model focus on both what and where to look.

The ViT splits the input image into equal-sized patches, converts each patch into a one-dimensional vector, and projects it into a patch embedding using a linear layer. Positional embeddings preserve spatial information, and the sequence goes through a Transformer encoder. Each encoder block contains Multi-Head Self-Attention (MSA) layers that learn global relationships between patches and Feed-Forward Neural Networks (FFN) for feature transformation, with residual connections and layer normalization after each step. The sequence's first token, the (CLS) token, pulls together global information for classification.

The output from the Transformer is passed through a classification layer, using SoftMax for final classification purposes. ViT treats the entire image in parallel, making it easier to maintain long-range relationships and global patterns. Each module in this hybrid model plays a separate role in the architecture. EfficientNetB3 captures detailed spatial and texture features from retinal fundus images. Then, the CBAM module refines the feature maps extracted using channel and spatial attention, allowing the network to emphasize the most important regions within the image that are most relevant to the task. Unlike CNNs, ViTs capture relationships and global context across image patches, expanding the locally based learning process of EfficientNetB3.

The EfficientNetB3-CBAM and ViT models produce their respective outputs, which we concatenate into a single feature vector. The feature vector proceeds to a dense layer that identifies whether the input image has a functional representation of glaucoma or normality. The dense layer uses a Binary Cross-Entropy with Logits Loss function, and the output of the dense layer is converted into a probability by means of the sigmoid function. If the probability is greater than 0.5, then the model will classify the image as glaucoma (1), and if not, it will be classified as normal (0). A schematic diagram is also included to illustrate how the spatial, attention-enhanced, and global features are combined prior to the final classification. This image provides a simple and effective representation of how the three models EfficientNetB3, CBAM, and ViT combine and share information.

6. Results

The Hybrid EVC model that was proposed achieved an AUC 0.9852, precision 0.947, recall 0.955, F1 score 0.9501, and accuracy 94.9% which is better than previous results and sets a new standard. The Receiver Operating Characteristic (ROC) curve is a graph that indicates the manner in which the model does its classification of separating the positive class from the negative class. A perfect AUC is close to 1, meaning that the model correctly predicts both positive and negative cases, showing great predictive power. An AUC of 0.5 means that the model performed as if it were randomly guessing.

TPR (True Positive Rate) is determined by calculating the ratio of correctly predicted positive samples to the total actual positive cases. The formula is given as:

$$\text{TRP=True Positives / (True Positives + False Negatives)} \quad (1)$$

The FPR (False Positive Rate) tells us how many negative cases the model wrongly marks as positive. It is calculated by dividing the number of false positives by all real negative cases. The formula is:

$$\text{FPR= False Positives / (False Positives + True Negatives)} \quad (2)$$

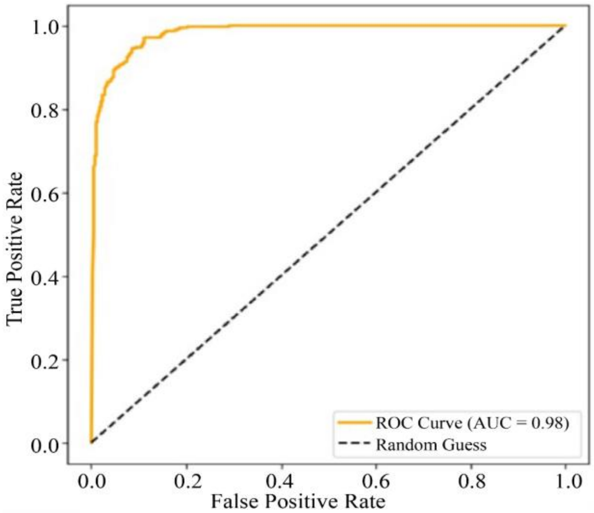

Fig. 6 ROC curve- glaucoma detection

The above Figure 6 shows the ROC curve for glaucoma detection. Here, this curve shows how the model differentiates glaucoma from non-glaucoma cases. The y-axis shows the TPR, which means the model correctly predicts the glaucoma when it is present. The x-axis indicates the FPR, which is when the model wrongly predicts glaucoma if it is not there.

When the ROC curve is nearer to the top-left corner, it signifies stronger model performance, while the black dashed line shows the outcome of random guessing. With an AUC score of 0.98, the model achieved the best results in correctly classifying glaucoma cases while reducing the false positives.

Table 2. Comparison of the proposed EVC model's performance with existing models

Model Name	Accuracy	AUC	F1 Score
AlterNet-K [24]	91.6%	0.968	0.915
ResNet50 [26]	~80%	-	-
DG2Net [27]	~91.2%	0.963	0.91
MobileNetV3 [28]	92.6%	-	-
Proposed Hybrid Deep Learning Model- EVC	**94.9%**	**0.9852**	**0.9501**

Table 2 shows different deep learning models that detect glaucoma using the same dataset, along with their respective accuracy, AUC, and F1 score.

The proposed Hybrid Deep Learning Model (EVC) achieved the highest performance, attaining an accuracy of 94.9%.

It performed better than all other models, including AlterNet-K, ResNet50, DG2Net, and MobileNetV3.

The hybrid model achieved the best accuracy among all other machine learning and deep learning models, along with XAI techniques, which is Grad-CAM, produced the output as a heatmap along with the probability of predicted results as follows:

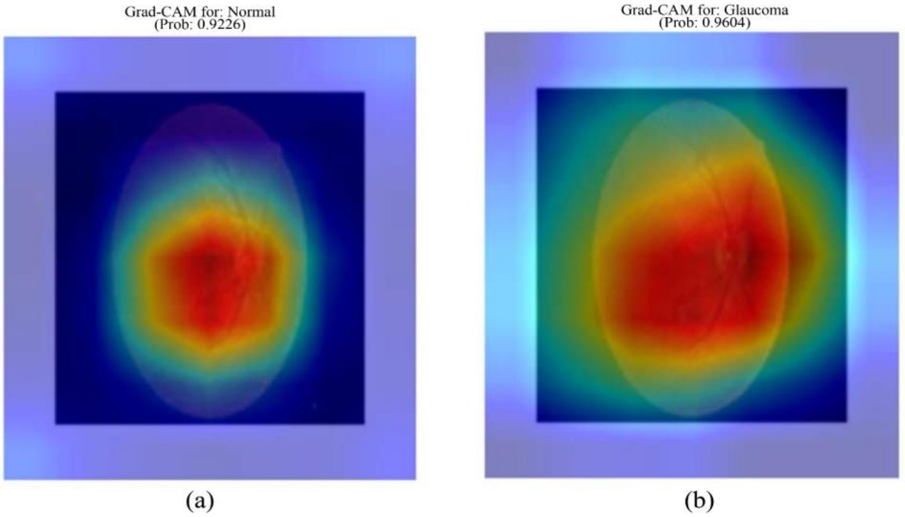

Fig. 7 Grad-CAM prediction for normal and glaucoma

Figure 7(a) shows the Grad-CAM visualization for a Normal retinal fundus image, with a predicted probability of 0.9226. The heatmap highlights the regions that the Hybrid CNN model focused on while classifying the image as Normal. The red and yellow areas indicate the regions with more important input to the prediction, while the blue and purple areas represent regions with minimal effect.

In this case, the central area of the retina is highlighted, showing that the model considered the relevant retinal region before confidently predicting the image as Normal. Figure 7 (b) shows the Grad-CAM visualization for a Glaucoma retinal fundus image, with a predicted probability of 0.9604. The heatmap highlights the areas that had the highest contribution to the model's decision. The red and yellow regions indicate where the Hybrid CNN focused the most to classify the image as glaucoma, while the blue and green regions had little to no influence. In this case, the heatmap mainly covers the OD and the surrounding retinal area, which are clinically relevant for the detection of glaucoma.

7. Conclusion

The results of this study show that combining convolutional and transformer-based networks can significantly improve glaucoma detection accuracy. This advancement could help doctors make faster and more reliable diagnoses by using AI support in screening and early detection. The use of Grad-CAM makes the model more transparent, allowing doctors to see which regions of the image influenced the prediction clearly. This enhances clinical trust and makes the system practical for hospital use or large-scale telemedicine-based glaucoma screening. Future research could aim to develop models that classify multiple stages of glaucoma, train them on larger and more varied datasets, and adapt the system for efficient mobile or real-time diagnostic use. In conclusion, this study presents an accurate glaucoma detection model that combines EfficientNet and Transformer architectures, along with Explainable AI (XAI) using Grad-CAM to enhance the clarity and interpretability of the results. The developed EVC model (EfficientNetB3 + ViT + CBAM) effectively captures both small details and overall features from eye images, yielding more accurate and easily interpretable results than existing methods. Testing on the EyePACS-AIROGS-light-V2 dataset showed that the model performs reliably across images with different quality levels. Its explainable visual outputs enable doctors to understand how predictions are made, strengthening confidence in clinical use. In the future, this model can be improved to identify various stages of glaucoma, made faster for real-time use, and developed into a mobile or web app for large-scale eye screening.

References

[1] Xiao Chun Ling et al., “Deep Learning in Glaucoma Detection and Progression Prediction: A Systematic Review and Meta-Analysis,” Biomedicines, vol. 13, no. 2, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Allen Beck, and Ta Chen Peter Chang, “Glaucoma: Definitions and Classification,” American Academy of Ophthalmology, 2015. [Online]. Available: https://www.aao.org/education/disease-review/glaucoma-definitions-classification

[3] Muhammad Naseer Bajwa et al., “Two-Stage Framework for Optic Disc Localization and Glaucoma Classification in Retinal Fundus Images using Deep Learning,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Diagnosing Glaucoma, The Glaucoma Foundation, 2018. [Online]. Available: https://glaucomafoundation.org/about-glaucoma-2/diagnosing-glaucoma/

[5] Glaucoma - Diagnosis and Treatment, Mayo Clinic, 2025. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/glaucoma/diagnosis-treatment/drc-20372846

[6] Tareek Pattewar et al., “Glaucoma Detection using Machine Learning with Fundus Images,” *Communications on Applied Nonllinear Analysis*, vol. 32, no. 9s, pp. 1171-1184, 2025. [CrossRef] [Publisher Link]

[7] Venkatesh Guntreddi, and Sivakumar V, “Deep Learning based Glaucoma Detection Using Majority Voting Ensemble of ResNet50, VGG16, and Swin Transformer,” *Results in Engineering*, vol. 28, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sajib Saha, Janardhan Vignarajan, and Shaun Frost, “A Fast and Fully Automated System for Glaucoma Detection using Color Fundus Photographs,” *Scientific Reports*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Guangzhou An et al., “Comparison of Machine-Learning Classification Models for Glaucoma Management,” *Journal of Health care Engineering*, vol. 2018, pp. 1-8, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Sumaiya Pathan et al., “Automated Segmentation and Classifcation of Retinal Features for Glaucoma Diagnosis,” *Biomedical Signal Processing and Control*, vol. 63, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Stefan Maetschke et al., “A Feature Agnostic Approach for Glaucoma Detection in Oct Volumes,” *Plos One*, vol. 14, no. 7, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Gavin D’Souza, P.C. Siddalingaswamy, and Mayur Anand Pandya, “AlterNet-K: A Small and Compact Model for the Detection of Glaucoma,” *Biomedical Engineering Letters*, vol. 14, pp. 23-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Fairouz Alsulami, et al., “HiGAN-CNN: A Hybrid Generative Adversarial Network and Convolutional Neural Network for Glaucoma Detection,” *International Journal of Computer Science and Network Security*, vol. 22, no. 9, pp. 23-30, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kaggle, *EyePACS-AIROGS-light-V2 Leaderboard- ConvNeXtTiny Benchmark*, 2023. [Online]. Available: https://github.com/TheBeastCoding/glaucoma-dataset-metadata/blob/main/benchmark-eyepacs-airogs-light-v2.md

[15] Oscar Perdomo et al., “Glaucoma Diagnosis from Eye Fundus Images based on Deep Morphometric Feature Extraction,” *Computational Pathology and Ophthalmic Medical Image Analysis*, Lecture Notes in Computer Science, pp. 319-327, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] S. Ajitha et al., “Identification of Glaucoma from Fundus Images using Deep Learning,” *Indian Journal of Ophthalmology*, vol. 69, no. 10, pp. 2702-2709, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sajib Saha, Janardhan Vignarajan, and Shaun Frost, “A Fast and Fully Automated System for Glaucoma Detection using Color Fundus Photographs,” *Scientific Reports*, vol. 13, no. 1, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Oluwatobi Joshua Afolabi et al., “The Use of U-Net Lite and Extreme Gradient Boost (XGB) for Glaucoma Detection,” *IEEE Access*, vol. 9, pp. 47411-47424, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Abeer Aljohani, and Rua Y. Aburasain, “A Hybrid Framework for Glaucoma Detection through Federated Machine Learning and Deep Learning Models,” *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mir Tanvir Islam et al., “Deep Learning-Based Glaucoma Detection with Cropped Optic Cup and Disc and Blood Vessel Segmentation,” *IEEE Access*, vol. 10, pp. 2828-2841, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Mamta Juneja et al., “Automated Detection of Glaucoma using Deep Learning Convolutional Network (G-net),” *Multimedia Tools and Applications*, vol. 79, no. 21-22, pp. 15531-15553, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ayesha Shoukat et al., “Automatic Diagnosis of Glaucoma from Retinal Images using Deep Learning Approach,” *Diagnostics*, vol. 13, no. 10, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ahmed Al-Mahrooqi et al., “GARDNet: Robust Multi-View Network for Glaucoma Classification in Color Fundus Images,” *arXiv Preprint*, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Gavin D’Souza, P. C. Siddalingaswamy, and Mayur Anand Pandya, “AlterNet-K: A Small and Compact Model for the Detection of Glaucoma,” *Biomedical Engineering Letters*, vol. 14, no. 1, pp. 23-33, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Riley Kiefer, Glaucoma Dataset: EyePACS-AIROGS-light-V2, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/deathtrooper/glaucoma-dataset-eyepacs-airogs-light-v2

[26] Chandra Nugraha, and Sri Hadianti, “Glaucoma Detection in Fundus Eye Images using Convolutional Neural Network Method with Visual Geometric Group 16 and Residual Network 50 Architecture,” *Medical Informatics Technology*, vol. 1, no. 2, pp. 36-41, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Yu Feng, Cong Wu & Yuan Zhou, “DG2Net: A MLP-Based Dynamixing Gate and Depthwise Group Norm Network for Glaucoma Detection,” *Pattern Recognition,* pp. 295-308, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] DiagIA, MobileNetV3 Model for Glaucoma Detection Kaggle Pipeline, 2025. [Online]. Available: https://github.com/DiagIA/retina-datasets.