Research Article

Deep Learning Based Automated Detection System For Diagnosing Glaucoma Disease

Yanna Prasanth[1], Dr. M. Kavitha[2]

[1,2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Guntur, AP, India
[1]Mail: 2401050081cse@gmail.com
[2]Mail: mkavita@kluniversity.in

Abstract

Glaucoma is an optic neuropathy that can cause permanent vision loss if not detected early. The goal of the present project is to design a controlled multi-architecture solution for automated glaucoma classification of retinal fundus images. A set of 7,815 images was collected across 6 public image repositories (ACRIMA, DRISHTI-GS, G1020, LAG, ORIGA, RIM-ONE) and duplicate images were eliminated at the file level by using a process of filename-label matching. This resulted in a dataset of 5,002 normal and 2,813 glaucomatous images that were divided into 6,252 training images, 781 validation images and 782 test images using two-stage stratified sampling. These models were tested: custom Basic CNN, hybrid CNN–Transformer (based on ResNet18), MobileNetV2, ResNet18, DenseNet121, EfficientNet-B0, Swin-T and squeeze-and-excitation attention-enhanced DenseNet121. The image were cropped to a size of 224 × 224, were normalized using the ImageNet statistics and were augmented in the training process by brightness-contrast, rotation and horizontal and vertical flipping. Adam optimizer, learning rate 2×10^{-4}, adaptive learning-rate reduction, checkpointing and early stopping were used for optimization. The best classification accuracy (93.22%), specificity (98.20%), and precision (96.36%) were obtained by ResNet18. DenseNet121 with attention had the best probability discrimination with AUC (0.9820; 95% CI: 0.9744–0.9882) and highest sensitivity (87.59%). Swin-T and EfficientNet-B0 also achieved good accuracy with 92.84% and 92.58% respectively. The results show that the residual learning can give highly reliable normal case exclusion and channel attention can enhance glaucoma sensitivity and ranking performance. The proposed benchmark is a repeatable framework that can aid in the identification of compact and efficient architectures for computer-aided glaucoma screening.

Keywords: Glaucoma detection, retinal fundus image, DL, CNN, vision transformer, transfer learning, squeeze and excitation attention, medical image classification, ROC analysis.

1. Introduction

Glaucoma is a chronic progressive disease of the optic nerve where patients can have some damage to their optic nerve without actually realizing that there is any significant vision loss. The resulting impairment is permanent and screening systems need to focus on early detection and clinically safe discrimination of glaucomatous and normal retinal appearance. The advantages of fundus photography for population-level screening are that it is non-invasive and readily deployable, but manual interpretation relies on specialist expertise,

and subtle changes to the optic disc, variability in images, and inter-observer differences affect the quality of the results. In recent years, deep learning has emerged as a key computational tool for glaucoma screening, glaucoma progression and glaucoma grading based on retinal images [1], [2], [17].

CNNs have shown to have great potential to learn optic-disc morphology, peripapillary texture, vascular configuration, and global retinal appearance directly from images. Transfer learning makes it more useful in practice by transferring representations trained on big collections of natural images to medical data with few annotations. In the context of retinal diagnosis, residual dense, mobile and compound-scaled networks have been extensively studied [4], [5], [16], [18], [19]. At the same time, transformer architectures have been developed to capture the long-range dependencies and wider spatial context, which can be achieved by using self-attention mechanisms. Some of the systems reported good results, such as transformer-based systems, foundation models [3] and CNN–transformer ensembles [6], [8], [10], [13], [15], but their relative advantage over well-tuned CNNs is dependent on the type of dataset used.

Although there has been progress, four methodological limitations are prevalent. First, most studies report on one repository or data split by architecture, which makes comparisons challenging. Second, there is a tendency to overemphasize accuracy without a corresponding discussion of the sensitivity, specificity, Matthews correlation coefficient (MCC), Cohen's kappa and the uncertainty of the area under the receiver operating characteristic curve

(AUC). Third, there are few studies that compare these lightweight networks, residual CNNs, hierarchical transformers and channel-attention models in a controlled protocol. Fourth, as is often the case, the impact of heterogeneous image sources is hidden behind repository-specific experiments, and not across a single multi-source corpus. Such restrictions make it difficult to reproduce and create architecture options for clinical screening.

This work seeks to close these gaps by systematically comparing eight architectures ranging from scratch-trained CNN, transfer-learning CNN, CNN–Transformer mixture, hierarchical Swin transformer, to an attention-enhanced DenseNet. The same preprocessing, splitting, optimization, checkpointing and evaluation protocol are used for all models. This design allows for direct investigation of the trade-offs between sensitivity, specificity, overall discrimination, architectural complexity and potential deployment cost.

This research makes the following main contributions:

1. A heterogeneous collection of 7,815 fundus images is created by compiling six public glaucoma repositories and performing file-level deduplication and reproducible stratified sampling (80:10:10).
2. Three, a custom CNN, five widely used transfer-learning backbones, a ResNet18–Transformer hybrid and a squeeze-and-excitation DenseNet121 are compared under a single experimental protocol.
3. The proposed hybrid model uses a transformer to map a 7 × 7 ResNet feature grid into 49 transformer tokens, and the proposed attention-enhanced DenseNet employs channel-recalibration modules after every dense block.
4. The assessment of the accuracy is not the only one required, clinically important sensitivity and specificity, F1-score, MCC, Cohen's kappa, AUC, and bootstrap 95% AUC CI are required.
5. Architecture effectiveness is interpreted together with the scale of the parameters, the patterns of confusion, the behavior of convergence and the likely needs for the screening deployment.

2. Literature Review

2.1 Systematic Evidence and Clinical Direction.

Deep learning has been regularly shown to be an effective approach to glaucoma detection using fundus images in the literature of systematic reviews and meta-analyses, which also show significant differences in the number of data sets used, reference standards, validation methods, and reporting [1], [2]. Another study on CNNs, vision transformers, attention mechanisms, and transfer learning also highlights the importance of having a lightweight and clinically interpretable algorithm when it is meant for real world screening [17]. The results of these studies provide a sense of both the promise of automated analysis and that while high performance can be reported, cross-source robustness and deployability are not guaranteed.

2.2 CNN and Transfer-Learning Approaches.

CNN-based studies have been investigated regarding task-specific networks and pretrained backbones. A multi-task graph-convolution network that integrated convolutional feature extraction and relational modeling performed well on a number of retinal repositories [4]. Another recent study on multi-architecture retinal disease classification has demonstrated the effectiveness of MobileNetV2 [5] and DenseNet121 for providing good representations in transfer-learning applications. A comparative transfer-learning study was conducted on a large Just RAIGS collection, which showed that the classic CNN architectures are still competitive when trained over a large-scale collection [16]. The use of custom CNNs with attention mechanisms and DenseNet121 for explainable systems also suggest that feature recalibration and dense connectivity can enhance the discriminatory ability of models, while preserving visual explanation [18], [19]. These studies inspire the adoption of a common benchmark for CNN families of different types: mobile, residual, dense, and attention-enhanced.

2.3 Transformer, Foundation, and Hybrid Architectures

The glaucoma screening research has moved from classifiers based on patches to segmentation-based and ensemble frameworks for transformers. A multi-backbone segmentation and vertical cup-to-disc ratio information was used for classification by an explainable transformer system [3]. A retinal foundation model was compared with a supervised CNN and showed that large-scale pretraining is far from always better than specialized convolutional learning [6] and a lightweight transformer architecture was created for early glaucoma staging [7]. A series of hybrid ensembles have been suggested to enhance the robustness against imbalance and noise [8]. Portable images obtained with the ophthalmoscope can also be used in comparative studies of different types of transformers, with the results showing

that the architecture of the transformer and the design of the ensemble play a significant role in the performance [10]. The benefits of global-context modeling and model visualization are further supported by foundation-model comparisons [13] and attention-infused screening systems [15].

2.4 Multi-Source Generalization, Imbalance, and Multimodal Learning.

One of the main challenges is generalization across image sources, as acquisition devices, field of view, illumination, population characteristics, and the annotation processes vary from image repository to image repository. Weighted-loss training over a large multi-source collection enabled better recall of glaucoma and pointed to the importance of dealing with class imbalance [9]. Multi-dataset training methods, such as sequential multi-dataset training with EfficientNet-B0, have also been studied for the purpose of achieving cross-dataset robustness [11]. In addition to fundus only classification, combination of fundus photos and optical coherence tomography (OCT) has shown the diagnostic utility of complementary modalities [12]. The results do not only complement the idea of heterogeneous data aggregation, but also show that eventually source-wise external validation should be complemented by internal stratification.

2.5 - Explanability and Alternative Automation Strategies

Clinically, interpretability is being viewed as a must-have. To evaluate if the models are focusing on the optic disc and cup instead of other irrelevant regions of the image, saliency analysis, Grad-CAM, and attention visualization are applied [6], [9], [13], [15], [18], [19]. While not specific to ophthalmology, transformation concepts developed for recurrent authentication can be used to showcase the usefulness of structured representation learning and reproducible transformation pipelines [14]. In addition, code-free automated machine learning systems have been tested for commonly seen eye diseases, showing a similar trend towards easily accessible model development [20]. However, performance alone without code is not a substitute for architecture-level controlled analysis to understand sensitivity, specificity and computational trade-offs.

2.6 Research Gap

Although there are significant advances in the literature, there is no single controlled comparison that features a scratch-trained CNN, several pretrained CNN families, a CNN–Transformer hybrid, a hierarchical transformer, and a channel-attention dense network on a common multi-source corpus. Several studies only report on accuracy or AUC, without any indication of clinically relevant false-negative behavior. To address this, the current study presents a reliable and reproducible benchmark of eight models that have been preprocessed, optimized, validated, and checked in the same manner, with a wide statistical analysis.

3. Methodology

3.1 Overall Framework

The suggested workflow is lopsided, comprised of multi-source data collection, duplicate data eradication, binary label harmonization, stratified partitioning, image preprocessing, augmentation at training time, independent optimization of models, held-out evaluation, and model interpretation comparison. The entire experimental workflow is summarized in figure 2. Most of the differences in the data handling and optimization of the models are intentional; differences in experimental settings are minimized to the extent that the design is consistent across model families and the behaviors are more architectural than experimental.

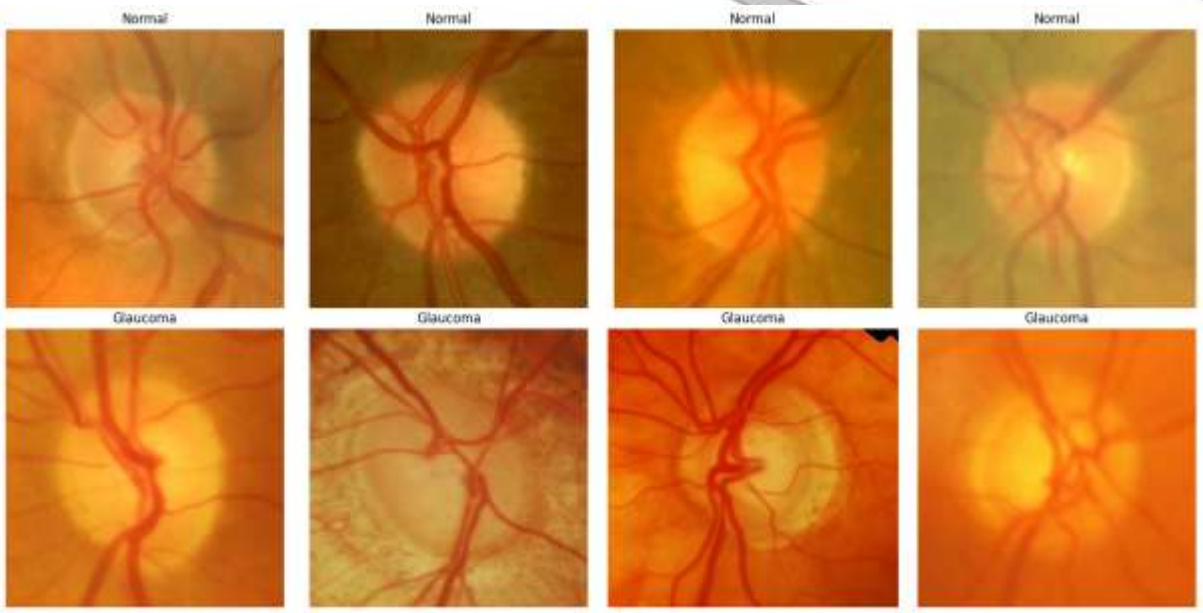

Figure 1. Representative normal and glaucomatous retinal fundus images from the unified dataset.

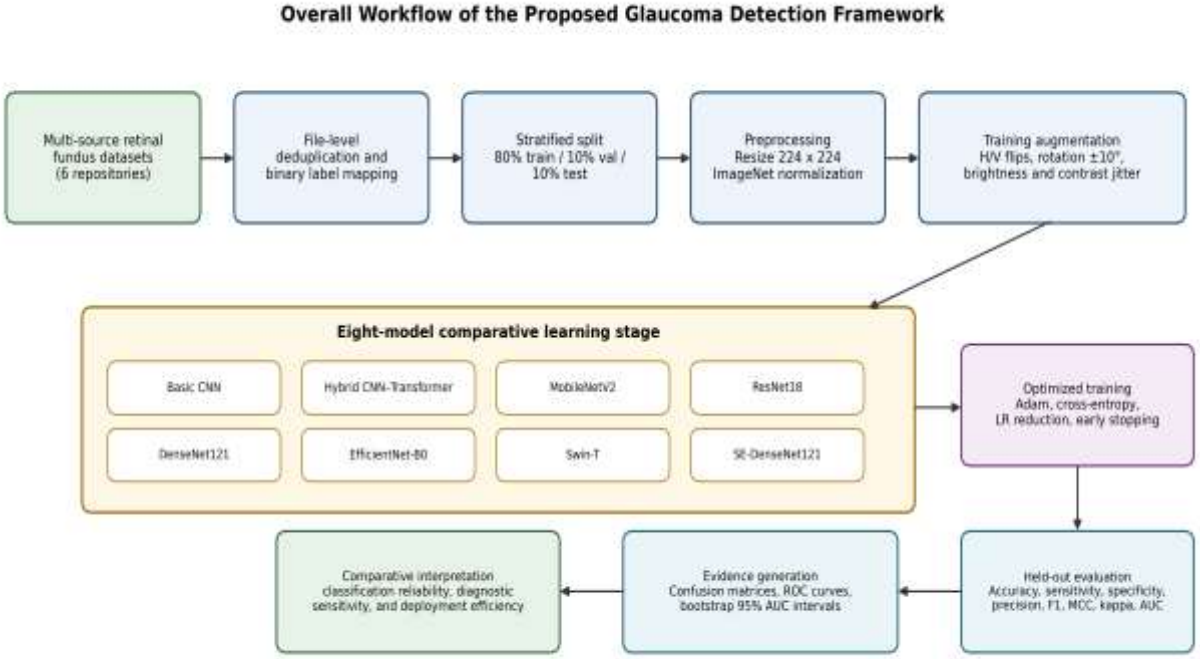

Figure 2. Overall workflow of the proposed multi-architecture glaucoma detection framework.

3.2 Dataset Description

The dataset was assembled from ACRIMA, DRISHTI-GS, G1020, LAG, ORIGA, and RIM-ONE. Images stored in normal folders were mapped to class 0 and images stored in glaucoma folders were mapped to class 1. Supported image formats included JPEG, PNG, and BMP. To reduce repeated entries originating from partitioned and non-partitioned directory structures, records sharing the same filename and class label were retained only once. This process yielded 7,815 images: 5,002 normal and 2,813 glaucomatous samples.

Dataset	Number of images	Contribution (%)
ACRIMA	705	9.02
DRISHTI-GS	101	1.29
G1020	1,020	13.05
LAG	4,854	62.11
ORIGA	650	8.32
RIM-ONE	485	6.21
Total	7,815	100.00

The class distribution corresponds to 64.01% normal images and 35.99% glaucoma images. This moderate imbalance was preserved through stratification so that validation and test performance remained representative of the complete corpus.

3.3 Stratified Dataset Partitioning

Stratified Shuffle Split procedure was performed in two stages using the fixed random seed 42. In the first phase, 80% of the samples were put into the training set, while the remaining 20% were temporarily set aside as a validation set. The holdout set was then divided into two parts: half for validation and the other half for testing. There are no enhancement images for use for validation or as test images validation or as test images.

Table 2. Stratified class distribution in the training, validation, and test partitions.

Partition	Normal	Glaucoma	Total	Share (%)
Training	4,002	2,250	6,252	80.00
Validation	500	281	781	9.99
Test	500	282	782	10.01
Total	5,002	2,813	7,815	100.00

3.4 Image Preprocessing and Augmentation

All fundus images were converted to RGB and reduced to 224 × 224 size. Pixel tensors were normalized using the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225). Random horizontal flipping, random vertical flipping, rotation in the range of +-10° and color jitter (brightness and contrast factors of 0.2) were applied to the training pipeline. The validation and test pipelines included only resizing, tensor conversion and normalization. The separation guaranteed that the evaluation was performed on the non-modified holdout images and that the training set was amplified with the geometric and photometric variability.

Table 3. Preprocessing and augmentation configuration.

Stage	Operation	Configuration
All partitions	Color mode	RGB
All partitions	Resize	224 × 224 pixels
All partitions	Tensor normalization	Mean = [0.485, 0.456, 0.406]; SD = [0.229, 0.224, 0.225]
Training only	Horizontal flip	Random, default probability 0.5

Training only	Vertical flip	Random, default probability 0.5
Training only	Rotation	Random angle within ±10°
Training only	Color jitter	Brightness = 0.2; contrast = 0.2

3.5 Deep Learning Architectures

The chosen architectures include complementary designs such as conventional convolution, residual learning, depthwise separable (DW) convolution, dense connectivity, local-window self-attention, squeeze and excitation (Squeeze Excitation) channel recalibration, and compound scaling. All models with fixed backbones were initialized with weights pre-trained on ImageNet while the Basic CNN was trained using the standard training procedure. For all models, the last layer was substituted with the two output head: normal and glaucoma classes.

3.5.1 Basic CNN

The Basic CNN has 4 convolutional blocks. The first two blocks have 3x3 convolution, batch normalization, rectified linear activation, and 3x3 max-pooling and stride of 3. The third block applies a 2 × 2 pooling, while the fourth block applies convolution, batch normalisation and activation without additional pooling. The last 128 × 9 × 9 is flattened and then projected to 512 hidden units, followed by Relu, dropout layer (with rate 0.5), and a 2-neuron output layer. The network serves as a non-pretrained baseline with which the benefits of transfer learning can be quantified.

Table 4. Layer-level configuration of the Basic CNN.

Layer group	**Configuration**	**Output representation**
Block 1	Conv 3→32, 3×3; BN; ReLU; MaxPool 3×3/3	32 × 74 × 74
Block 2	Conv 32→64, 3×3; BN; ReLU; MaxPool 3×3/3	64 × 24 × 24
Block 3	Conv 64→128, 3×3; BN; ReLU; MaxPool 2×2/2	128 × 11 × 11
Block 4	Conv 128→128, 3×3; BN; ReLU	128 × 9 × 9
Classifier	Flatten; Linear 10,368→512; ReLU; Dropout 0.5; Linear 512→2	Two-class logits

3.5.2 Hybrid CNN–Transformer

Hybrid architecture consist of a pre-trained ResNet18 with no global pooling and fully connected layers. The convolutional backbone generates a feature map of 7 x 7 size with 512 channels for an input of 224 x 224. This map is flattened into 49 feature tokens, with each token having 512 dimensions. A learnable positional embedding is added on, and a learnable classification token is added at the beginning to form a sequence of length 50. It passes through two layers of Transformer encoder, each having eight attention heads and a model dimension of 512. The classification token is passed to a classification network consisting of a two-output linear head and layer normalization. This design integrates inductive bias of convolution with long-range token interactions from self-attention.

3.5.3 Transfer-Learning CNNs

MobileNetV2 is the smallest pretrained model in the comparison as it uses inverted residual bottlenecks and depthwise separable convolutions. ResNet18 uses identity shortcuts to make propagation of gradients more stable and to retain discriminative residual features. A DenseNet121 is a type which reuses features by appending the output of each layer to the layers that follow it. EfficientNet-B0 optimizes the network depth, width and resolution by compound scaling and mobile inverted bottlenecks. For each architecture, the original ImageNet classifier was substituted with a two-neuron linear layer and the network is fine-tuned.

3.5.4 Swin Transformer

Swin-T is a hierarchical Vision Transformer where the self-attention is calculated in local windows and the windows are divided into regular and shifted windows. The spatial resolution decreases and representation depth increases with the patch merging process. Replaced original classification head with two-output layer. Swin-T does not rely on a convolutional backbone for learning its hierarchy, unlike the CNN – Transformer hybrid.

3.5.5 Attention-Enhanced DenseNet121

The attention enhanced DenseNet121 is an extension of DenseNet network that has dense blocks 1-4 followed by attention network (Attention DenseNet). Global Average Pooling results in a C-dimensional descriptor for a feature tensor with C channels. A channel bottleneck is used to rescale the original tensor with channel weights that are sigmoid-normalized, which is the case when a two-layer channel bottleneck with reduction ratio r = 16 is applied. SE modules were inserted for channel dimensions 256, 512, 1,024, and 1,024. Such a process enables the model to enhance the glaucomatous structure channels and to reduce weak or redundant signals.

Table 5. Architectural characteristics and trainable parameter counts.

Model	Principal design	Pretraining	Parameters (M)
Basic CNN	Four convolutional blocks + 512-unit classifier	No	5.551
Hybrid CNN–Transformer	ResNet18 feature map + 2-layer, 8-head Transformer	ImageNet	17.772
MobileNetV2	Inverted residual and depthwise separable blocks	ImageNet	2.226
ResNet18	Residual basic blocks	ImageNet	11.178
DenseNet121	Dense feature connectivity	ImageNet	6.956
EfficientNet-B0	Compound-scaled MBConv network	ImageNet	4.010
Swin-T	Hierarchical shifted-window Transformer	ImageNet	27.521
SE-DenseNet121	DenseNet121 + four SE channel-attention blocks	ImageNet	7.262

3.6 Training Strategy

Cross entropy loss was used as the loss function and the Adam optimizer was used to optimize the models. The initial learning rate was 2×10^{-4}. The validation loss was monitored by a ReduceLROnPlateau scheduler, with its learning rate being reduced every 4 epochs of training if no improvement was observed. The training could be carried on for a maximum of 50 epochs. A parameter state with the lowest validation loss was kept and the training was terminated after eight epochs with no improvement in the validation loss. For all models except EfficientNet-B0, a batch size of 32 was chosen, and for EfficientNet-B0 a batch size of 16 was used. The data loading was done with two worker processes. The seeds for the random number generators in Python, NumPy and PyTorch were set to the constant value 42, while the behavior of CUDA was set to deterministic.

Table 6. Common training configuration.

Component	Setting
Objective function	Cross-entropy loss
Optimizer	Adam
Initial learning rate	0.0002

Scheduler	ReduceLROnPlateau, factor = 0.5, patience = 4
Maximum epochs	50
Early-stopping patience	8 epochs
Checkpoint criterion	Minimum validation loss
Batch size	32; EfficientNet-B0 = 16
Random seed	42
Execution	CUDA-enabled GPU environment

3.7 Evaluation Methodology

Glaucoma was the only positive class. Given a confusion matrix with true positives (TP), true negatives (TN), false positives (FP) and false negative (FN), the following were calculated:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Sensitivity = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Precision = TP / (TP + FP)
- F1 score = 2 × (Precision × Sensitivity) / (Precision + Sensitivity).

A balanced correlation measure MCC was used to measure the correlation under class imbalance and Cohen's kappa was used to measure agreement over chance. Glaucoma probabilities were used to create receiver operating characteristic curves, which were used to summarize threshold-independent discrimination ability using AUC. A 95% AUC confidence interval was estimated from 1,000 bootstrap replications (with the exception of those based on a single class). The confusion matrices were kept to illustrate the clinical distribution of the false negative and false positive decisions.

3.8 Implementation Environment

The system was built in Python, and the model was defined and trained with PyTorch, Torchvision and scikit-learn was used for stratified sampling and performance measurements, Pillow was used for loading images, NumPy and Pandas were used for handling the numerical data and Matplotlib/Seaborn was used for visualizing the results. The experiments were run on a Kaggle platform with a CUDA-enabled worker. This set-up enabled fine-tuning of the model using the GPU and ensured deterministic control of the experiments.

4. Results

4.1 Convergence and Checkpoint Selection

Models were selected based on the lowest validation loss, and not the final epoch. The transfer-learning models typically merged with each other significantly earlier than the Basic CNN. The two best results were achieved by efficientnet-b0 with the lowest validation loss (0.1842) and swin-t with the highest validation accuracy at the point of lowest loss (92.70%). The Basic CNN needed 42 epochs (with early stopping) to reach the desired accuracy.

Table 7. Training convergence and selected validation checkpoints.

Model	Early-stop epoch	Best-loss epoch	Minimum val. loss	Val. accuracy at checkpoint (%)
Basic CNN	42	34	0.2489	88.99
Hybrid CNN–Transformer	15	7	0.2066	91.55
MobileNetV2	18	10	0.1960	91.55
ResNet18	20	12	0.1979	91.29
EfficientNet-B0	18	10	0.1842	92.06
Swin-T	30	22	0.1909	92.70

SE-DenseNet121	23	15	0.2029	90.78

4.2 Overall Classification Performance

The main diagnostic parameters are given in Table 8. ResNet18 achieved the highest accuracy (93.22%), specificity (98.20%), and precision (96.36%). These values represent exceptional reliability in excluding glaucoma in normal images, and a small false-positive load. The best recovery of glaucomatous cases was seen in SE-DenseNet121 with a sensitivity of 87.59%. Swin-T achieved the second-best accuracy (92.84%) and a balanced F1 score (89.71%). EfficientNet-B0 was found to be very close with 92.58% accuracy. DenseNet121 had a balanced classification profile with 91%, whereas the Basic CNN was the most specific with a low sensitivity.

Table 8. Test performance comparison of the eight deep learning models.

Model	Accuracy (%)	Sensitivity (%	Specificity (%	Precision (%)	F1-score (%)
Basic CNN	90.79	78.72	97.60	94.87	86.05
Hybrid CNN–Transformer	91.56	82.62	96.60	93.20	87.59
MobileNetV2	91.94	86.52	95.00	90.71	88.57
ResNet18	93.22	84.40	98.20	96.36	89.98
DenseNet121	91.00	88.00	93.00	88.00	88.00
EfficientNet-B0	92.58	86.88	95.80	92.11	89.42
Swin-T	92.84	86.52	96.40	93.13	89.71
SE-DenseNet121	92.46	87.59	95.20	91.14	89.33

The transfer-learning models always outperformed the scratch-trained Basic CNN in terms of sensitivity and F1 score. The hybrid network achieved superior sensitivity of 3.90 percentage points when compared with the Basic CNN, indicating that the fusion of pretrained convolutional features and token-level self-attention resulted in an improvement. The pure ResNet18 model still outperformed the others in terms of overall accuracy and also specificity, indicating that their moderate amount of data was appropriate for residual convolutional representations.

4.3 Advanced Agreement and Discrimination Metrics

Results of the probability-based evaluations are summarized in Table 9 in terms of MCC, Cohen's kappa, AUC, and bootstrap confidence intervals for the probability. ResNet18 had the highest MCC (0.8532) and kappa (0.8489), showing a good balance across both classes. SE-DenseNet121 attained the highest AUC (0.9820), followed by ResNet18 (0.9795). AUCs overlap, suggesting that although the models have different threshold-dependent sensitivity and specificity, the overall ranking performance of the leading models are not significantly different.

Table 9. Agreement statistics and ROC discrimination.

Model	MCC	Cohen's kappa	AUC	AUC 95% CI
Basic CNN	0.8003	0.7926	0.9644	0.9536–0.9754
Hybrid CNN–Transformer	0.8157	0.8123	0.9726	0.9626–0.9809
MobileNetV2	0.8241	0.8235	0.9745	0.9650–0.9824
ResNet18	0.8532	0.8489	0.9795	0.9712–0.9866
EfficientNet-B0	0.8380	0.8371	0.9751	0.9655–0.9835
Swin-T	0.8436	0.8423	0.9754	0.9669–0.9840

SE-DenseNet121	0.8354	0.8350	0.9820	0.9744–0.9882

4.4 Confusion-Matrix Analysis

The confusion matrices obtained during the evaluation are presented in figure 3. The highest number of glaucoma misdiagnosis was from the Basic CNN with 60 false negatives. The hybrid approach lowered the number of false negatives to 49. The number of false negatives for MobileNetV2 is 38, for Swin-T is 38, and for EfficientNet-B0 it was 37, while for SE-DenseNet121 it was 35. With a mere 44 false-negatives, ResNet18's results in terms of specificity and precision were outstanding, but its performance of 44 false-negatives fell short of the attention-enhanced DenseNet. This indicates that these architectures could not optimally minimize all the clinically-relevant types of errors simultaneously.

Table 10. Confusion-matrix counts for the probability-based test evaluations.

Model	TN	FP	FN	TP
Basic CNN	488	12	60	222
Hybrid CNN–Transformer	483	17	49	233
MobileNetV2	475	25	38	244
ResNet18	491	9	44	238
EfficientNet-B0	479	21	37	245
Swin-T	482	18	38	244
SE-DenseNet121	476	24	35	247

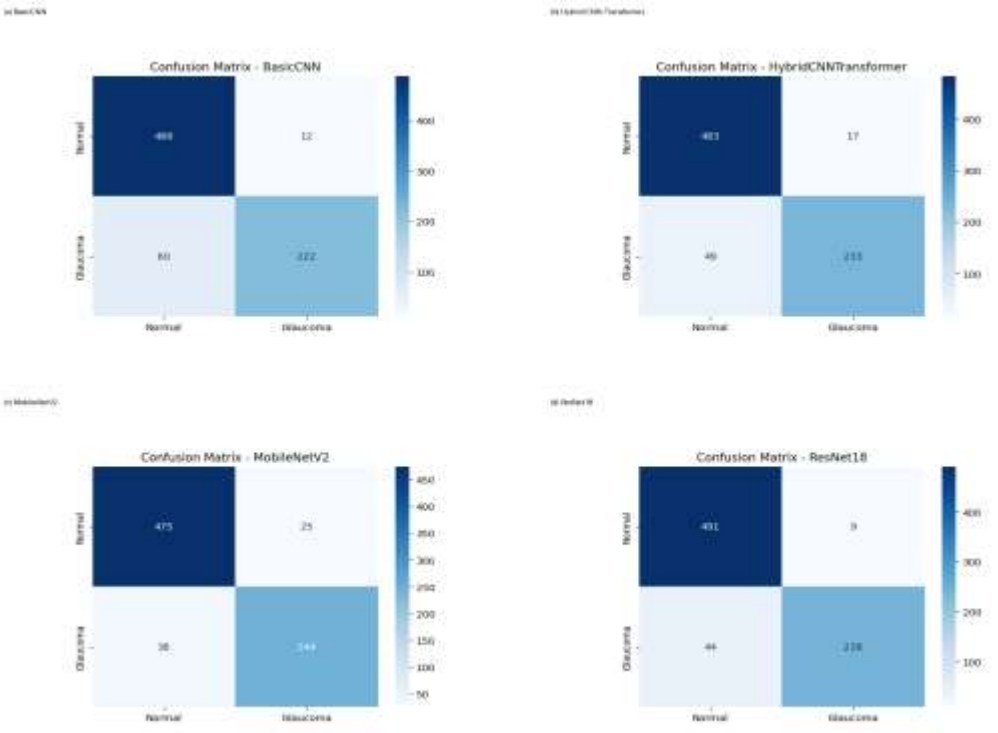

Figure 3. Confusion matrices for (a) Basic CNN, (b) Hybrid CNN–Transformer, (c) MobileNetV2, and (d) ResNet18. The horizontal axis denotes predicted class and the vertical axis denotes true class.

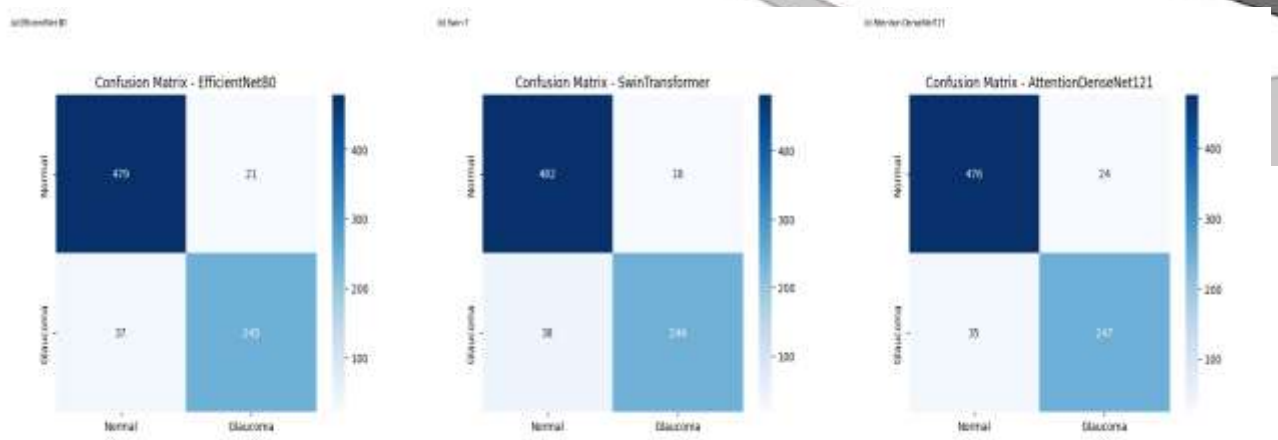

Figure 3 (continued). Confusion matrices for (e) EfficientNet-B0, (f) Swin-T, and (g) SE-DenseNe

4.5 ROC and Precision–Recall Evidence

The ROC curves in Figure 4 stay near the upper left corner, which is indicative of high sensitivity-specificity trade-offs at various decision thresholds. The largest AUC value is obtained by SE-DenseNet121 for the best ordering of glaucoma probabilities, and the best fixed-threshold specificity is achieved by ResNet18. The confusion matrix, ROC curve, precision–recall curve, and metric profile are all found in the Basic CNN dashboard shown in figure 5. Its mean rank quality is good, and the plots are oriented towards recall, showing sub-optimal glaucoma detection compared to the pretrained alternatives.

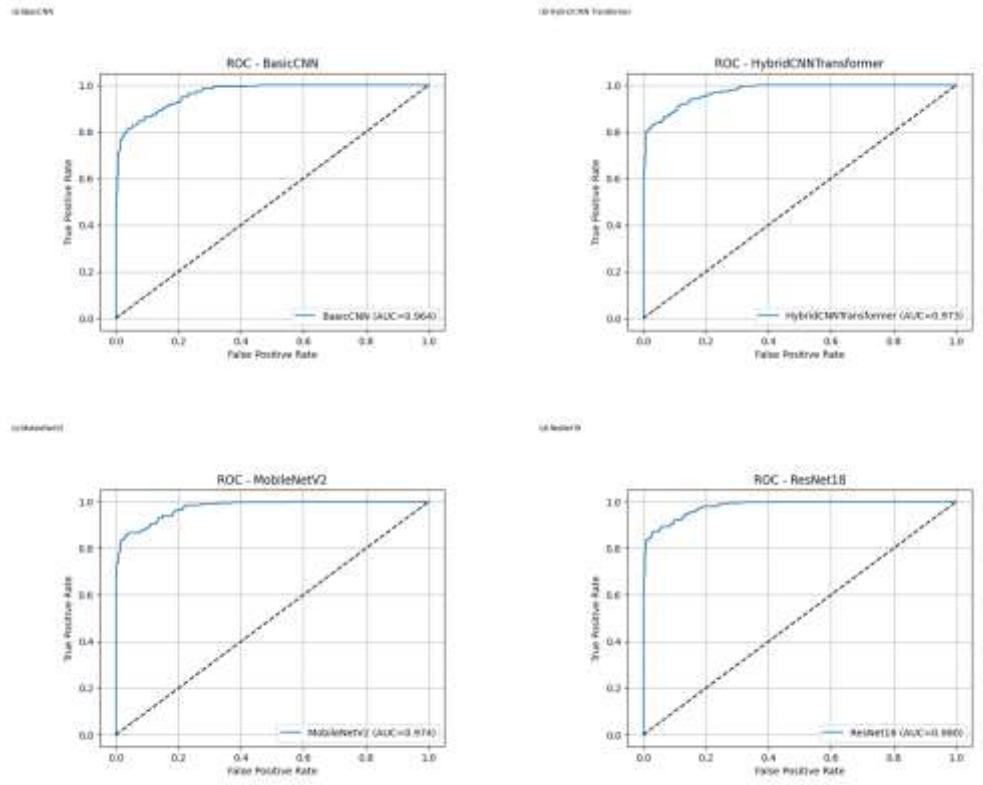

Figure 4. ROC curves for (a) Basic CNN, (b) Hybrid CNN–Transformer, (c) MobileNetV2, and (d) ResNet18.

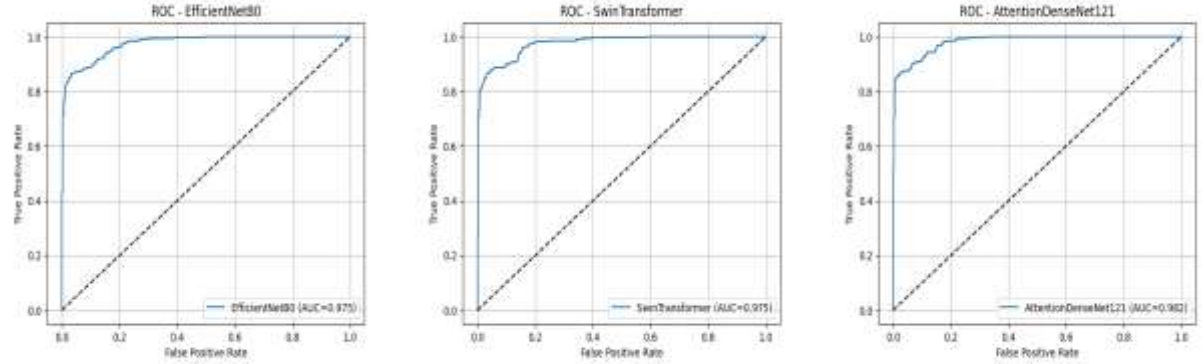

Figure 4. ROC curves for (e) EfficientNet-B0, (f) Swin-T, and (g) SE-DenseNet121.

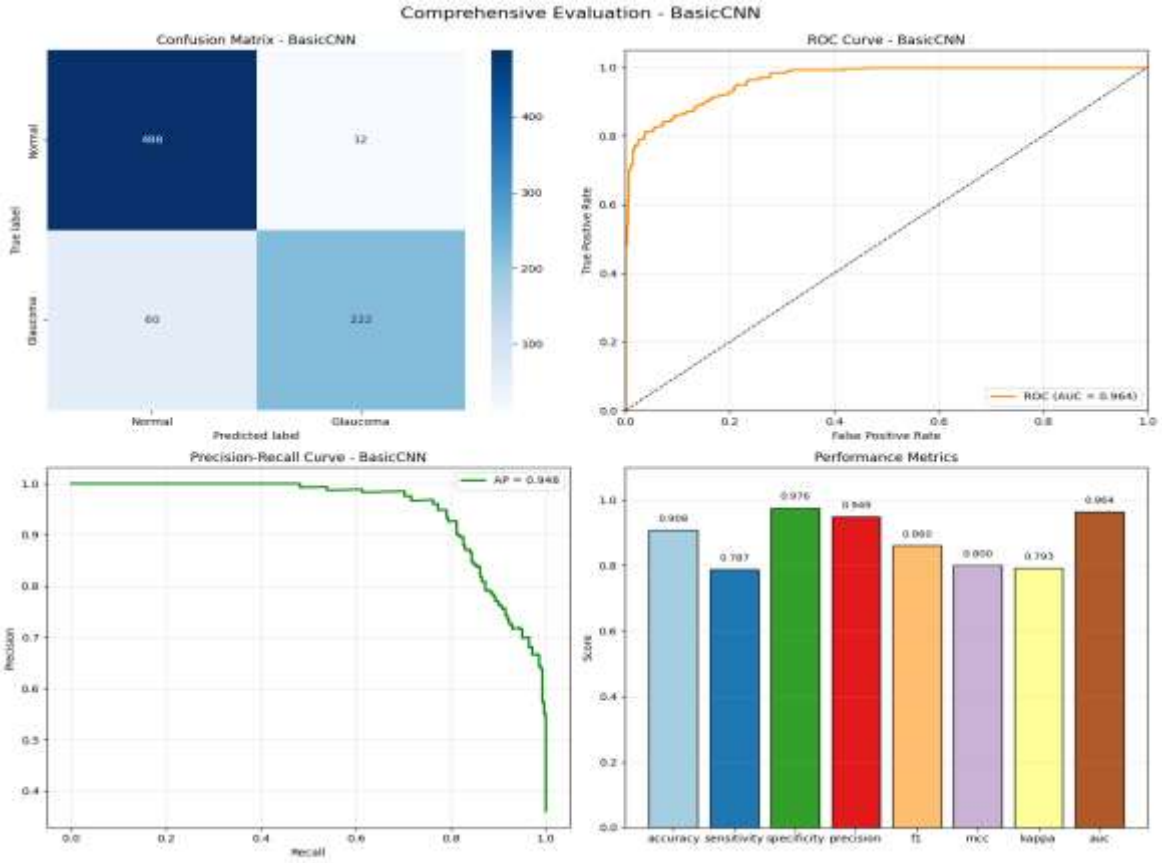

Figure 5. Comprehensive evaluation dashboard for the Basic CNN, including confusion matrix, ROC curve, precision–recall curve, and performance metrics.

4.6 Architecture and Deployment Trade-Offs

In the mobile model, the biggest, MobileNetV2, has around 2.23 million trainable parameters and 91.94% accuracy and 86.52% sensitivity. EfficientNet-B0 is able to reach a higher accuracy of 92.58% using around ~4.01M parameters, making it a good accuracy-to-size comparison. ResNet18 takes about 11.18 million parameters, and achieved the highest overall accuracy and agreement. The most noticeable and biggest model is swin-T, which has around 27.52 million parameters and accuracy of 92.84% – this shows that hierarchical self-attention is competitive, but might be large for low-resource deployments. The SE-DenseNet121 achieves the highest sensitivity and AUC with an increase of only ~0.31 million parameters in DenseNet121, demonstrating the effectiveness of lightweight channel attention in enhancing

clinically relevant behaviour without substantial growth in model size.

5. Discussion

5.1 Interpretation of the Main Findings

Comparative results show that there is a difference between threshold-dependent classification reliability and threshold-independent discrimination. ResNet18 was the most accurate model with the highest value for specificity, precision, MCC and kappa with the selected decision threshold. It has residual shortcuts which enable stable fine-tuning and maintain the hierarchical features of the optic disc without the need for a parameter scale of larger transformers. On the other hand, SE-DenseNet121 has the best sensitivity and AUC. Parallel and dense feature reuse provides multi-level representations in the retina, and channel attention selectively highlights informative features. This is a good ratio to use when the clinical goal is to minimize missed glaucoma cases.

The hybrid CNN–Transformer outperforms the Basic CNN in terms of sensitivity and F1, showing that self-attention can enhance the global reasoning ability from features learned from mature CNN networks. However, the hybrid did not surpass ResNet18, EfficientNet-B0 or Swin-T in accuracy. One likely reason is that the 49 spatial tokens obtained from the last feature map in ResNet represent a relatively low-level representation; better multi-scale tokenization might be needed to truly make the most of transformer attention. The results show that Swin-T achieved excellent performance due to the use of shifted windows, which help capture local structure while progressively adding more context.

Transfer learning was consistently beneficial. The Basic CNN achieved a 90.79% accuracy and a sensitivity of 78.72%, while all the pretrained models were significantly better in recalling glaucoma. This is significant because false-negative predictions are more important in a screening system than false alarms. This supports the idea of pre-trained residual, efficient, or attention-based dense architectures over a baseline architecture.

5.2 Relation to Previous Studies

The performance observed is within the range of the other general ranges found in the systematic reviews [1], [2], [17] and points to the fact that well-tuned CNNs are very competitive with transformers. The well-performing MobileNetV2, ResNet18 and DenseNet-family models are consistent with the results of earlier CNN studies [4], [5], [16], [18] and [19] on transfer-learning. The competitive, but not necessarily better, performance of the hybrid and Swin architectures also aligns with comparison between transformer, foundation and ensemble approaches [3], [6], [8], [10], [13], [15]. Instead of as many previous experiments, the present study is used to evaluate these model families in a single multi-source dataset and single optimization protocol, to obtain a more direct comparison at the architecture level.

The results also corroborate the importance of class imbalance and multi-source learning [9], [11]. The dataset is skewed toward normal images, but stratification and wide reporting on metrics make accuracy do not obscure poor sensitivity. The study is limited to fundus images and does not include the complementary structural information available from OCT fusion [12]. However, the multi-source fundus corpus has implications for low cost screening systems where OCT might not be available.

5.3 Clinical and Practical Significance

Ideally, a glaucoma screening model should avoid false negatives and have high specificity, to prevent over-referral. The findings in this study tell of two complementary deployment profiles. ResNet18 is appropriate if overall agreement is not as important but high specificity is. When sensitivity and probability ranking are highlighted, SE-DenseNet121 is preferable. For devices with limited resources, there are more compact variants: EfficientNet-B0 and MobileNetV2. These systems should be considered decision-support/triage systems and positive predictions must be corroborated by clinical examination.

5.4 Strengths

- The study does not rely on any one repository, but rather a multi-source, heterogeneous corpus.
- A similar preprocessing, split, training and checkpoint protocol are applied to all major architectures.
- The model set includes models that are based on scratch CNNs, transfer-learning CNNs, a hybrid transformer, a hierarchical transformer and explicit channel attention.
- Performance reporting consists of clinically meaningful error measures, agreement statistics, ROC analysis and bootstrap confidence intervals.
- Parameter counts and confusion patterns are used together to aid in model selection for deployment.

5.5 Limitations

There are a lot of LAG images in the dataset, which can make the learned representation mimic the characteristics of the LAG acquisition process. Duplicate removal is based on filename and label instead of perceptual image hashing and this means that it can keep visually similar images with different filenames. The current evaluation does not employ a completely isolated source-wise external test set, but rather the internal stratified partitions. Demographic, camera, disease stage and image quality metadata was not included. The models are not explicitly segmenting the optic disc and/or optic cup, quantifying uncertainty, assessing probability calibration, nor are they actually classifying the whole image. Lastly, potential clinical evaluation and human-reader comparison are needed prior to deployment.

5.6 Future Work

To improve future research, leave one dataset out cross-validation, prospective multicenter testing, source-aware sampling and perceptual duplicate detection should be carried out. The hybrid CNN–Transformer might benefit from multi-scale token fusion and class-sensitive or focal objectives could further enhance glaucoma recall. Issues of model calibration, uncertainty estimation and choice of thresh-holding for clinical triage should be investigated. Predictions that can be explained using segmentation assisted learning, explainability techniques, and ophthalmologist validation can illuminate if the prediction is anatomically valid. In the case of edge deployment, quantization, pruning, knowledge distillation, and latency measurement should be done on MobileNetV2, EfficientNet-B0 and compact residual models. Multimodal fusion of OCT and clinical risk factors is another avenue for a more holistic glaucoma examination.

6. Conclusion

In this study, the authors present a single deep learning model for binary glaucoma detection using retinal fundus images from six publicly available databases that contain 7815 images. A controlled preprocessing, augmentation, stratified-split, optimization, checkpointing and evaluation protocol was used to investigate eight architectures. It also included a comparison with a scratch-trained CNN, a CNN–Transformer hybrid based on ResNet18, MobileNetV2, ResNet18, DenseNet121, EfficientNet-B0, Swin-T and one with SE attention enhancement on DenseNet121.

ResNet18 achieved the highest accuracy of 93.22%, specificity of 98.20%, precision of 96.36%, MCC of 0.8532, and Cohen's kappa of 0.8489. Channel attention had the best sensitivity of 87.59%, and AUC of 0.9820, showing that it is clinically useful for glaucoma recovery in glaucomatous cases. Swin-T and EfficientNet-B0 also performed well in terms of accuracy, while MobileNetV2 offered the most lightweight and effective performance. The results demonstrate the continued usefulness of residual convolution for accurate glaucoma classification and the enhanced sensitivity of attention-based dense connectivity and its threshold independence for glaucoma discrimination. The suggested criterion offers a strict framework for external validation, explanation, calibration of the model and for use in limited-resource computer-aided glaucoma screening.

References:

1. X. C. Ling et al., "Deep Learning in Glaucoma Detection and Progression Prediction: A Systematic Review and Meta-Analysis," Biomedicines 2025, Vol. 13, vol. 13, no. 2, Feb. 2025, doi: 10.3390/biomedicines13020420.
2. E. Arrieta-Rodriguez et al., "Deep Learning for Glaucoma Classification and Grading: A Comprehensive Review on Fundus Imaging Approaches," IEEE Access, vol. 13, pp. 163699–163730, Jan. 2025, doi: 10.1109/ACCESS.2025.3605819.
3. H. Alasmari, G. Amoudi, and H. Alghamdi, "Explainable Transformer-Based Framework for Glaucoma Detection from Fundus Images Using Multi-Backbone Segmentation and vCDR-Based Classification," Diagnostics, vol. 15, no. 18, p. 2301, Sep. 2025, doi: 10.3390/diagnostics15182301.
4. S. Lenka, Z. L. Mayaluri, and G. Panda, "Glaucoma detection from retinal fundus images using graph convolution based multi-task model," e-Prime - Advances in Electrical Engineering, Electronics and Energy, vol. 11, p. 100931, Mar. 2025, doi: 10.1016/j.prime.2025.100931.
5. S. A. Hussein, A. A. Farouk, and M. M. Saeid, "Intelligent retinal disease detection using deep learning," Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 43282-, Dec. 2025, doi: 10.1038/s41598-025-28376-w.
6. K. Bolo et al., "Comparison of RETFound and a Supervised Convolutional Neural Network for Detection of Referable Glaucoma from Fundus Photographs," Ophthalmology Science, vol. 6, no. 2, p. 101008, Feb. 2025, doi: 10.1016/j.xops.2025.101008.
7. M. Yurdakul, K. Uyar, and Ş. Taşdemir, "MaxGlaViT: A Novel Lightweight Vision Transformer-Based Approach for Early Diagnosis of Glaucoma Stages From Fundus Images," Int. J. Imaging Syst. Technol., vol. 35, no. 4, p. e70159, Jul. 2025, doi: 10.1002/ima.70159.

8. F. Tariq, H. Hameed, T. Malik, M. Mollel, M. A. Imran, and Q. H. Abbasi, “Hybrid ResNet–Vision Transformer ensemble for early glaucoma detection from fundus images,” IEEE Sens. J., pp. 1–1, Jan. 2026, doi: 10.1109/jsen.2025.3650333.
9. D. J. Nugraha, N. Yudistira, and A. W. Widodo, “Weighted loss for imbalanced glaucoma detection: Insights from visual explanations,” Comput. Biol. Med., vol. 196, no. Pt C, Sep. 2025, doi: 10.1016/j.compbiomed.2025.110875.
10. R. O. C. Costa, P. A. F. França, A. C. P. Pessoa, G. Braz Júnior, J. D. S. de Almeida, and A. Cunha, “Comparative Analysis of Transformer Architectures and Ensemble Methods for Automated Glaucoma Screening in Fundus Images from Portable Ophthalmoscopes,” Vision, vol. 9, no. 4, p. 93, Dec. 2025, doi: 10.3390/vision9040093.
11. R. P. Chowdhury and N. S. Karkera, “Early Glaucoma Detection using Deep Learning with Multiple Datasets of Fundus Images,” ArXiv, p. arXiv:2506.21770, Jun. 2025, Accessed: May 11, 2026. [Online]. Available: http://arxiv.org/abs/2506.21770
12. S. Islam, R. C. Deo, P. D. Barua, J. Soar, and U. R. Acharya, “Novel Deep Learning Model for Glaucoma Detection Using Fusion of Fundus and Optical Coherence Tomography Images,” Sensors (Basel), vol. 25, no. 14, p. 4337, Jul. 2025, doi: 10.3390/s25144337.
13. K. Bolo et al., “Comparison of Foundation and Supervised Learning-Based Models for Detection of Referable Glaucoma from Fundus Photographs,” Aug. 2025, doi: 10.1101/2025.08.21.25334170.
14. P. Ramu, P. Parida, P. Bellamkonda, and M. K. Panda, “Keystroke Authentication Using a Novel RTS Framework,” 2024 IEEE 4th International Conference on Applied Electromagnetics, Signal Processing, and Communication, AESPC 2024, 2024, doi: 10.1109/AESPC63931.2024.10872345.
15. R. Swaminathan, “An Attention Infused Deep Learning System with Grad-CAM Visualization for Early Screening of Glaucoma,” Jan. 2026, Accessed: May 11, 2026. [Online]. Available: https://arxiv.org/pdf/2505.17808v2
16. R.-M. FLORESCU, D.-O. ALEXANDRU, and M.-S. ŞERBĂNESCU, “Deep Learning with Transfer Learning for Automated Glaucoma Detection in Fundus Images,” Appl. Med. Inform., vol. 47, no. Suppl. 1, May 2025, Accessed: May 11, 2026. [Online]. Available: https://doaj.org/article/a85528847a474a44927e3ff0fcbf5b6d
17. S. Saidah, A. Rizal, and I. Wijayanto, “Artificial intelligence in glaucoma detection system based on fundus images,” PeerJ Comput. Sci., vol. 12, p. e3705, Mar. 2026, doi: 10.7717/peerj-cs.3705.
18. V. K. Velpula et al., “Efficient and Explainable Glaucoma Detection Using an Attention-Enhanced Custom CNN on Retinal Fundus Images,” Engineering Proceedings 2026, Vol. 124, vol. 124, no. 1, p. 110, Apr. 2026, doi: 10.3390/engproc2026124110.
19. H. C. Pathmakumara and G. Perera, “Explainable Deep Learning for Glaucoma Detection: A DenseNet121-Based Classification with Grad-CAM Visualization,” medRxiv, p. 2025.10.08.25337634, Oct. 2025, doi: 10.1101/2025.10.08.25337634.
20. T. Lin and T. Leng, “Code-Free Machine Learning for the Detection of Common Ophthalmic Diseases,” Transl. Vis. Sci. Technol., vol. 14, no. 9, p. 16, Sep. 2025, doi: 10.1167/tvst.14.9.16.
21. Donepudi,R., & Kavitha, M. (2025). A Hybrid CNN-BiGRU-GAN Framework for EnhancedAutomated Analysis of Cervical Cancer in Medical Imaging. InternationalJournal of Advanced Computer Science & Applications, 16(12), 1082.·.
22. Rohini,D., & Kavitha, M. (2024). ABC-Optimized CNN-GRU Algorithm for ImprovedCervical Cancer Detection and Classification Using Multimodal Data. InternationalJournal of Advanced Computer Science & Applications, 15(9),701