Alkadhim Journal for Computer Science, Vol. 4, No. 3 (2026)

Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

Alkadhim Journal for Computer Science (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

Efficient Glaucoma Detection from Retinal Fundus Images Using a Lightweight DeiT-Small Vision Transformer and Test-Time Augmentation

Mothefer Majeed Jahefer

Imam Alkadhim University College – Iraq

Article information

Article history:
Received: August, 08, 2026
Accepted: August, 30, 2026
Available online: September, 25, 2026

Keywords:
Glaucoma detection,
Vision Transformer,
DeiT-Small,
Test-Time Augmentation,
retinal fundus images.

Corresponding Author:
Mothefer Majeed Jahefer
modafarmajed@iku.edu.iq

DOI:
https://doi.org/10.61710/pxx89244

This article is licensed under:
Creative Commons Attribution 4.0 International License.

Abstract

Glaucoma is one of the major causes of permanent blindness in the world, and since the progressive nature of the disease leads to considerable damage of the optic nerve before diagnosis, there is a need to develop precise and effective automated screening tools. This paper suggests the implementation of a lightweight vision transformer network for glaucoma screening based on the DeiT-Small architecture. The images were downsized to 224 × 224 resolution and transformed into patches. Four transformer blocks were frozen while the other blocks were fine-tuned via layer-wise learning-rate decay (LLRD) along with data augmentation and Test-Time Augmentation (TTA). Testing was performed on a balanced dataset with 9,540 images: 8,000 for training, 770 for validation and 770 for testing. In the independent test dataset, the model achieved 91.65% accuracy, 91.55% F1-score, 91.56% sensitivity, 95.06% specificity, and 95.84% AUC-ROC. Test-time augmentation was applied to enhance the reliability of predictions without re-training the architecture. According to the Grad-CAM visualization of predictions, the model made decisions based on clinically important features of the retinal structure, such as optic disc and optic cup. These results show the possibility of the effectiveness of a transformer-based technique in terms of automated glaucoma diagnosis, whereas further external validation using the datasets that represent the glaucoma distribution in the population is required.

1. Introduction

Glaucoma is considered to be the most common cause of permanent vision loss. Being a progressive optic neuropathy, it has a distinctive feature of retinal nerve fibre layer damage, thinning of the neuroretinal rim, and an increase in the size of the optic cup in comparison with the optic disc. In 2020, about 76 million people were diagnosed with glaucoma, with more than 3.2 million being completely blind, and 4.5 million suffering from moderate-to-severe visual impairment due to glaucoma [1]. Furthermore, epidemiologic predictions present an even more difficult scenario, with the global burden expected to increase to 112 million affected people by 2040, primarily fuelled by the growing number of ageing people and the increased prevalence of myopia, especially in

East and Southeast Asia [2]. The reason for glaucoma's danger is the long period during which it is asymptomatic. The optic nerve degenerates without any signs, so at least 50% of people in developed countries — and up to 90% in developing ones — do not even know about their diagnosis until the damage has become irreparable [3]. As glaucomatous damage is irreversible, the best ways to stop it are early detection and immediate treatment [4]. Traditionally, the diagnosis of glaucoma has been based on the assessment of the IOP, the visual field examination, and the evaluation of the optic nerve head by means of slit-lamp exam or fundus photography [5]. There are, however, some serious weaknesses in this approach. Elevated IOP is not an adequate marker for diagnosing glaucoma; there are cases where glaucoma develops despite normal IOP, and vice versa [6]. In addition, the onset of any abnormality in the visual field is generally noticed only after a tremendous reduction of 30–40% of retinal ganglion cells, indicating that the disease has already progressed to a point that is difficult to treat. In the case of direct assessment of the optic disc, there is a lot of subjectivity even among experienced ophthalmologists. Adding to these challenges, there is a lack of eye doctors, particularly in less privileged regions — highlighting the immediate necessity for an automatic tool [7]. To overcome this, deep learning is considered a paradigm-shifting technology in the field of automated glaucoma detection, providing an efficient means of achieving clinical-level precision in diagnosing glaucoma among populations without requiring an ophthalmologist to be physically present at the test site. Convolutional neural networks (CNNs) have been observed to be exceptionally proficient in extracting useful features from fundus images, such as cup-to-disc ratio, rim thinning, and RNFL degradation patterns [8].Transfer learning from large-scale collections of natural images such as ImageNet has become very popular, given the prevalent scarcity of available medical images; architectures such as Res Net, Inception, and Efficient Net are used to obtain a highly competitive baseline on publicly available datasets [9]. However, even though these techniques work well, standard CNN approaches still have a fundamental drawback caused by the local receptive fields of convolutional kernels, which analyse only small neighborhoods of the image at a time [10]. As a result, it becomes hard for CNNs to capture natural long-distance relationships that play an extremely important role for clinicians, such as the overall ratio between the cup and disc or symmetric patterns within the eye. In the case of Vision Transformers (ViTs), an input image is processed in a patch-wise manner by means of global self-attention, meaning that long-distance spatial relationships across the whole fundus image can be captured already from the first layer, resulting in much better generalization across different cameras and imaging conditions [11]. The only problem with ViTs is that they are notoriously data-hungry, requiring large datasets for training. Fortunately, DeiT addresses this issue through knowledge distillation during pertaining [12]. More specifically, DeiT-Small (with 22M parameters) offers a compact vision-transformer architecture that utilizes global self-attention with a design well suited to clinical-scale datasets. In addition, Test-Time Augmentation (TTA) an effective technique for improving prediction stability without any additional training, by generating different variations of the same image — gives us a chance to further enhance the model. Inspired by this, in this paper we propose a compact vision-transformer architecture for automated glaucoma detection based on DeiT-Small, which we optimize through layer freezing and test-time augmentation (TTA). All training and testing were carried out on a balanced clinical dataset consisting of 9,540 retinal fundus images (8,000 for training, 770 for validation, and 770 in the independent test set), with half of them being healthy and the other half glaucomatous. In the independent test set, our TTA-based model showed highly stable diagnostic capabilities in all metrics considered. The experimental results show that our method shows performance close to the best existing solutions. Contributions of this paper are stated below.

The main contributions of this work:

• Selective fine-tuning of a small transformer: The DeiT-Small backbone is selectively frozen up to the first four transformer blocks, while the last eight transformer blocks remain trainable. With the aid of a specially designed classifier head, the model is able to train from robustly pre-trained features and focus on retinal fundus properties exclusively.

• Test-Time Augmentation (TTA) study: A thorough study is conducted on the effect of TTA on fundus image classification. By applying systematic horizontal and vertical flips in multiple passes during the inference phase, we demonstrate how TTA improves sensitivity and AUC-ROC, without modifying any weight or the structure of the already trained model.

• Clinically Transparent Evaluation: The performance of the model for the test data consisting of 770 images is analyzed in-depth through confusion matrices (normal and TTA) along with the ROC curve. The proposed system manages to score an average accuracy of 93.51% by detecting 371 out of 385 cases of glaucoma (RG), which results in a glaucoma sensitivity of 96.36%.

This paper is structured as follows. **Section 2** presents the related work. **Section 3** describes the proposed method for glaucoma detection from retinal fundus images using a lightweight DeiT-Small Vision Transformer with test-time augmentation. **Section 4** discusses the experimental results using five evaluation metrics. Finally, **Section 5** concludes the paper.

2. Related Work

In the field of automated glaucoma diagnosis, many developments in terms of technology have been achieved, ranging from traditional machine learning algorithms to deep learning algorithms. The framework introduced by Krishnan and Faust [13] was among the pioneering frameworks based on raw fundus images, which used vessel segmentation with directional filtering, manually engineered clinical features, and ensembles of random forest chosen with the help of META-DES algorithm. The results obtained through this framework in the public and private dataset indicated that with the right engineering of the feature pipeline, one could achieve a strong baseline using non-deep learning approaches. MTL-IO (Multi-task Learning for glaucom A via Image-based segmentation) [14] is an approach presented by Pascal et al. in which the authors introduced a U-Net-based Encoder-Decoder model for multi-task learning of Optic disc segmentation, Cup segmentation, Fovea localization, and Glaucoma detection, with the use of VGG-16 based encoder and focal loss to mitigate class imbalance. Following this trend of research towards multimodal models, Li et al. [15] have proposed GMNN net, a model that combines clinical information with imaging using three parallel streams of clinical information mining, U-Net-based segmentation feature extraction, and Res Former-based raw image processing, followed by Cat Boost classification. Velpula and Sharma [16] performed ensemble learning using five pre-trained CNNs (ResNet50, Alex Net, VGG19, DenseNet-201, and Inception-ResNet-v2) combined by majority voting, extending classification beyond binary detection to multi-class staging (normal, early, and advanced) while demonstrating the benefits of assembling for stability and diagnostic capability. With deployment as a priority, Saha et al. [17] presented an end-to-end, two-step framework involving YOLO-Nano for optic-nerve-head localization and MobileNetV3-Small for classification on cropped images, comparing different backbone networks under transfer learning and training from scratch. The work of Fan et al. [18] investigated a fully transformer-based method, applying the Vision Transformer directly to fundus images without any convolutions. Therefore, the proposed model was more generalized than the conventional CNN classifier, indicating the effectiveness of global self-attention in modeling the diffuse changes in glaucomatous eyes. Consistent with this trend, one example would be the attention augmented dual channel DenseNet-121 model of Chakraborty et al. [19] that uses both CBAM attention for channel and spatial interactions as well as the CRM attention for edge perception through the use of Sobel statistics with emphasis on clinically significant features such as the optic disc and cup. Likewise, Brivitha and S. Sophia [20] used a U-Net with a ResNet34 encoder for optic disc and cup segmentation, followed by classification of the segmented and cropped region with a modified EfficientNetB0, which is more efficient than other models in terms of efficiency and representational capacity. Generally speaking, this shows a shift in the field of ophthalmic machine learning from feature engineering to multitask learning, multimodal integration, and attention mechanisms in CNNs. Most existing methods for the diagnosis of glaucoma are based on complex heavy models, However, this paper achieves a similarly high level of clinical performance using a lightweight DeiT-Small Vision Transformer. In order to adapt our model for retinal images, selective block unfreezing and Layer-wise Learning Rate Decay (LLRD) are used. Further robustness of the model comes from Test-Time Augmentation (TTA), and that without adding to the architecture of the model. Table 1 presents representative works in the field along with the hybrid CNN-Transformer approach used in this work.

Table (1): Summary of prior work on automated glaucoma detection

ID	Ref.	Authors	Method / Model	Modality	Methodology
1	[13], 2023	Krishnan & Faust	Ensemble Random Forest + META-DES	Fundus images	Directional-filter-based retinal vessel segmentation, ROI statistical feature extraction, texture and clinical feature analysis, followed by dynamic ensemble classification using random forests.
2	[14], 2022	Pascal et al.	MTL-IO (Multi-Task Learning with Independent Optimizers)	Fundus images	Multi-task deep-learning architecture based on VGG-16 U-Net, integrating retinal-structure segmentation, anatomical-landmark localization, and glaucoma classification with class-imbalance handling.
3	[15], 2022	Li et al.	GMNNnet (Multimodal Neural Network)	Fundus images + OCT + electronic medical records	Multimodal architecture with metadata processing, U-Net segmentation, Res Former feature extraction, feature fusion, and Cat Boost classification for glaucoma diagnosis.
4	[16], 2023	Velpula & Sharma	5-CNN ensemble (ResNet50, Alex Net, VGG19, DenseNet-201, Inception-ResNet-v2) + majority voting	Fundus images	Ensemble of five pre-trained CNN backbones trained by k-fold cross-validation and data augmentation techniques using majority voting algorithm; also extended from binary detection to multi-class glaucoma classification (normal, early, advanced).
5	[17], 2023	Saha et al.	Simplified YOLO-Nano + MobileNetV3-Small	Fundus images	Two-stage architecture with optic-nerve-head localization via YOLO-Nano followed by MobileNetV3-Small-based glaucoma classification using transfer-learning strategies.
6	[19], 2023	Fan et al.	Convolution-free Vision Transformer	Fundus photographs	Vision Transformer applied directly to fundus images without convolutional layers, evaluated for cross-dataset generalization relative to conventional CNN-based classifiers.
7	[20], 2024	Chakraborty et al.	Dual-Attention DenseNet-121 (CBAM + CRM)	Fundus images	Deep-feature extraction using DenseNet-121, refined through CBAM attention and CRM adaptive channel recalibration, followed by global pooling for glaucoma classification.
8	[18], 2025	K. J. Tina Brivitha and S. Sophia	U-Net (ResNet34) + EfficientNetB0	HRF	Pretrained U-Net–ResNet34 for OD/OC segmentation and ROI extraction, followed by EfficientNetB0 classification with feature pooling, dropout regularization, and image augmentation.

3. Materials and Methods

In this section, we discuss the DeiT-Small Vision Transformer architecture adopted for Glaucoma detection using retinal fundus images. The complete pipeline, as illustrated in Figure 1, consists of seven consecutive stages, namely image acquisition and dataset creation, noise removal and augmentation, DeiT-Small feature extraction, transformer layer fine-tuning, soft max binary classification, test-time data augmentation, and Grad-CAM explain ability.

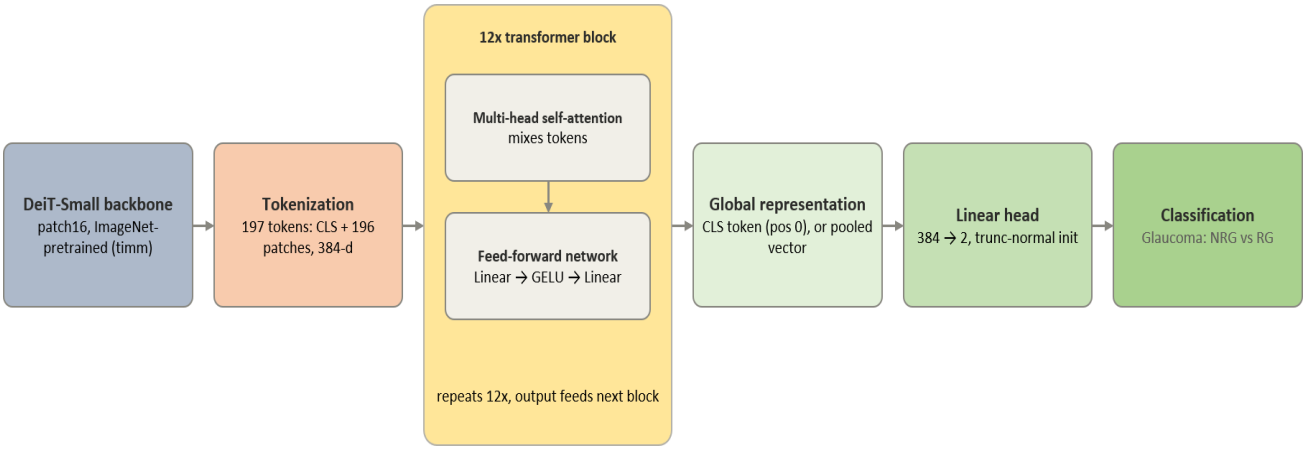


Figure (1): Proposed DeiT-Small glaucoma-detection pipeline

3.1 Dataset and Experimental Setup

The experiments were conducted using the EyePACS-AIROGS-light-V2 [21] dataset that is publicly accessible and contains retinal fundus images. There is total 9,540 images in the dataset, classified into two categories: non-referable glaucoma (NRG, 0), referable glaucoma (RG,1)

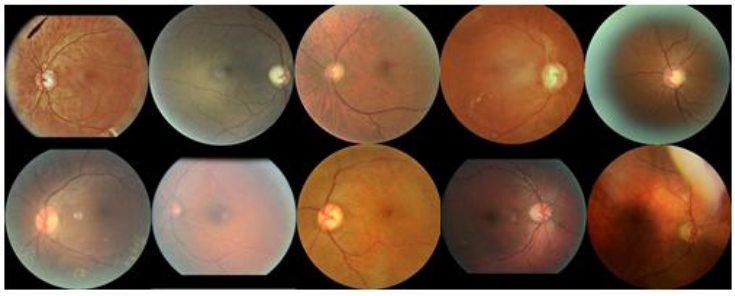


Figure (2): Retinal images sample: Normal/NRG (top row) and Glaucomatous/RG (bottom row).

By using stratified random sampling technique for ensuring class balance throughout, the data set was partitioned into three sets as follows: 8,000 training images (4,000 NRG and 4,000 RG), 770 validation images (385 NRG and 385 RG), and 770 test images (385 NRG and 385 RG). This is to ensure class balance in all data sets, as shown in Table 2.

Table (2): Dataset split across training, validation, and test partitions

Dataset Split	NRG (Normal)	RG (Glaucoma)	Total Images
Training	4,000	4,000	8,000
Validation	385	385	770
Test	385	385	770
Total	4,770	4,770	9,540

3.2 Data Preprocessing

All input images are resized to 224 × 224 pixels, matching the input resolution required by the DeiT-Small patch-embedding layer. Pixel intensities are standardized using the channel-wise statistics of the ImageNet dataset, with mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$. To increase generalization and

avoid overfitting to retinal fundus structures during training, we use a stochastic data-augmentation pipeline involving random horizontal and vertical flipping. Vertical flipping is particularly important for fundus photography, as it helps the network become invariant to camera orientation without altering pathological structures. In addition, to smooth decision boundaries, we apply Mixup and CutMix regularization at the batch level during training, with Mixup α = 0.8, CutMix α = 1.0, and a 50% probability of switching between the two techniques.

3.3 DeiT-Small Architecture

The classification backbone uses a DeiT-Small architecture with a 16×16 patch size (DeiT-Small with 16×16 patches and 224×224), pre-trained on ImageNet and obtained via the timm library [22]. DeiT-Small outputs a 384-dimensional feature representation of an image, which is mapped to the output class labels using a simple linear classification head. The classifier weights are initialized from a truncated-normal distribution (σ = 0.02) with zero bias. For an input image, the encoder returns a sequence of patch-token feature representations along with a classification (CLS) token. These tokens are fed sequentially into 12 transformer layers; in each layer, after multi-head self-attention, they pass through a feed-forward network (FFN) defined as:

$$FFN(x) = W_2\sigma(W_1x + b_1) + b_2 \qquad (1)$$

where x represents the input token representation, W_1 and W_2 are the projection weights of the two linear layers, b_1 and b_2 are their respective biases, and σ denotes the Gaussian Error Linear Unit (GELU) activation function. When the encoder output is a 3-D tensor of shape (batch × tokens × channels), the token at position zero — the CLS token — is used to represent the global image; if the encoder instead returns a pooled feature vector, that vector is used directly.

3.4 Selective Layer Freezing

Fine-tuning a model pre-trained on general natural images (e.g., ImageNet) for a highly specific task such as glaucoma diagnosis poses a serious risk of overfitting, given the comparatively small size of the target fundus dataset. To mitigate this risk and reduce computational cost, we use selective (partial) layer freezing before fine-tuning. This architectural design allows separation of the general visual representation extracted by the model from task-specific representations that need to be learned:

• Initial static components: the embedding layer, CLS token, and positional encoding layers are frozen; thus, no gradient is applied to these layers during the training process. The initial layers contain information about the general structure of visual data, which allows good generalization ability for different environments without fine-tuning.

• The freezing threshold (blocks 1-4): The transformer encoder is split at the threshold of 4. Four initial blocks are frozen; thus, maintaining its ability to model mid-level geometric features.

• Active training blocks (blocks 5-12): The last eight blocks are fully trainable, providing deep parts of the architecture with the ability to learn the clinically relevant structures, such as the optic disc and cup.

Finally, the last layer normalization layer is trainable. This subtle but crucial step enables the output feature distribution to be easily aligned with the new classification head.

3.5 Layer-Wise Learning Rate Decay (LLRD)

Learning rate decay per layer (LLRD) is used for fine-tuning the pre-trained vision transformer [23]. The approach uses lower learning rates for shallow layers and higher learning rates for deep layers. For an active block of the transformer at level i (from 0 to N−1, with N=12), the learning rate is computed as:

$$LR_i = LR_{base} \times \gamma^{N-i} \qquad (2)$$

where $LR_{base} = 5 \times 10^{-4}$ and γ=0.75. Early layers receive significantly smaller updates (e.g., $LR_0 \approx 1.58\times10^{-5}$), while later layers are fine-tuned at rates closer to LR_{base} (e.g., $LR_{11} \approx 3.75\times10^{-4}$). The classifier head and layer-

normalization weights are exempted from this decay and trained using the unscaled base rate of 5×10^{-4}. For improved efficiency, only the unfrozen parameters whose gradient is non-zero are updated. The optimization of all parameters takes place with AdamW [24], with a weight decay of 0.05 and an epsilon value of $\epsilon = 1 \times 10^{-8}$. A cosine-annealing scheduler reduces the learning rate from 5×10^{-4} 5×10^{-6} over 50 epochs to ensure stable convergence.

3.6 Training and Evaluation Protocol

The model is trained for 50 epochs on a dataset comprising 8,000 training images, 770 validation images, and 770 images in a separate test set, using mini-batches of 32 images. To avoid memorization of the training data a major concern in medical image analysis . we apply Mixup and CutMix augmentations. Because these augmentations generate mixtures of images, they produce soft, fractional targets rather than strict one-hot labels; accordingly, we use soft-target cross-entropy for the loss calculation. We also apply gradient clipping to stabilize training, update parameters using the AdamW optimizer configured according to the LLRD scheme, and step the cosine-annealing scheduler after each epoch. To preserve the objectivity of the generalization assessment, the model is evaluated on the validation set after every epoch, tracking cross-entropy loss and top-1 accuracy. Rather than simply retaining the weights from the final epoch, we adopt validation-based checkpointing, saving a new weight file whenever a new peak validation accuracy is reached. In addition, the model is passively evaluated on the independent test set every 5 epochs, providing a systematic view of generalization performance during training. Because checkpoints are selected using validation metrics only, these test-set evaluations remain strictly passive and do not influence the training process.

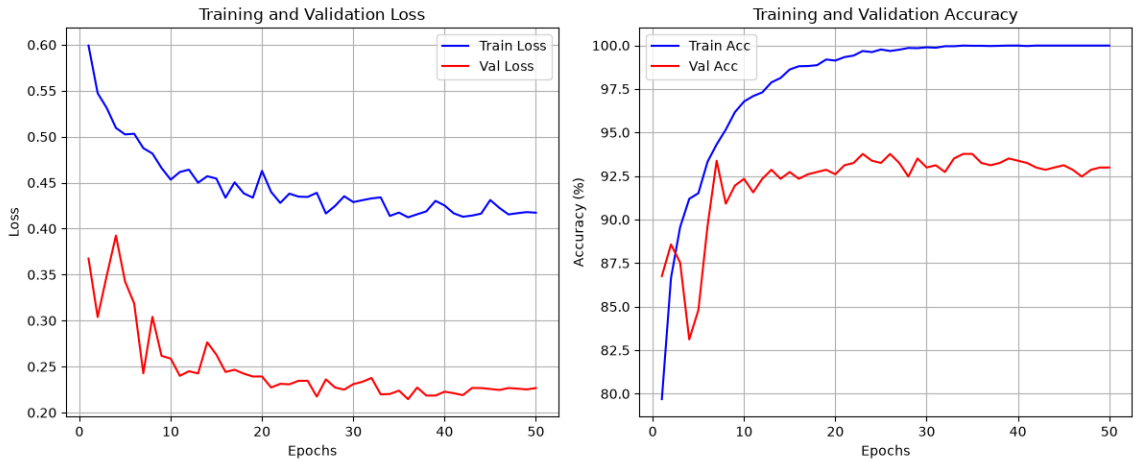


Figure (3): Training and validation curves over 50 epochs of fine-tuning

These dynamics are shown in Figure (3) below. It can be seen that stable convergence was achieved during all the training. The training accuracy starts from 79.69% and sharply increases, getting close to a plateau. The validation accuracy also sharpens and reaches the maximum value of 93.51% at Epoch 23, after which the best checkpoint for the model is chosen according to the validation dataset. At the same time, the test accuracy grows from 86.36% at Epoch 5 to 93.25% at Epoch 35 and finally settles down at the value of 92.86% at the final epoch. The validation loss is rather low and stable at the level of 0.22, while the training and validation performance are rather consistent, meaning that there was almost no overfitting.

3.7 Implementation Details

The entire training pipeline is implemented in PyTorch. The DeiT-Small backbone is configured and loaded through the timm library, which is also used for the Mixup/Cut Mix data-augmentation techniques (timm.data) and the soft-target cross-entropy loss function (timm. loss). All computations are performed on an NVIDIA GeForce RTX 5060 Laptop GPU with 8 GB of VRAM using CUDA acceleration. To make the experiments fully reproducible, all hyper parameters are listed in Table 3.

Table (3): Summary of training configuration and hyperparameters.

Hyper parameter	Value
Backbone architecture	DeiT-Small (patch16, 224×224)
Embedding dimension	384
Number of classes	2 (glaucomatous / non-glaucomatous)
Input resolution	224 × 224
Batch size	32
Training epochs	50
Base learning rate	5×10^{-4}
LLRD layer-wise decay rate	0.75
Weight decay	0.05
Optimizer	AdamW ($\varepsilon = 1 \times 10^{-8}$)
LR schedule	Cosine annealing ($\eta min = 1 \times 10^{-6}$)
Frozen transformer blocks	First 4 blocks (patch embedding, CLS token, positional embedding also frozen)
Mixup α	0.8
CutMix α	1.0
Mixup / CutMix application probability	1.0 (switch probability 0.5)
Label smoothing	0.1
Gradient clipping (max norm)	1.0

3.8 Soft max Classification

The output of the custom classification head is a two-dimensional logit vector $z \in \mathbb{R}^2$, passed through a soft max function to produce the final class-probability distribution:

$$P(y = k \mid x) = \frac{e^{zk}}{\sum_{j=1}^{2} e^{zj}} \qquad (3)$$

The model's predicted class label ŷ corresponds to the index of the highest probability:

$$\hat{y} = argmax_k \; p\,(y = k|x) \qquad (4)$$

During standard, single-pass evaluation, the computed probability for the positive (glaucoma) class, P(y = 1 | x), serves directly as the risk score, used to construct Receiver Operating Characteristic (ROC) curves and to calculate the Area Under the Curve (AUC-ROC).

3.9 Test-Time Augmentation

For optimal performance and stability in production use of our model, Test-Time Augmentation (TTA) technique was applied. The main benefit of TTA is the fact that the model and the dataset do not need to be changed at all.

N = 5 copies of each test sample are created using random horizontal and vertical flips and then run through the network individually, leading to five probability vectors which when averaged element-wise give

$$P_{TTA} = (y = k|x) = \frac{1}{N}\sum_{n=1}^{N} P(y = k|Aug\,, n(x)) \qquad (5)$$

Here, $Aug_n(x)$ represents the n-th version of the input image x that has been randomly flipped. The class label predicted by TTA is based on the highest average probability:

$$\hat{y}_{TTA} = \ argmax_k \ P_{TTA}\,(y = k|x) \qquad (6)$$

To evaluate the clinical relevance, the average probability P_{TTA} (y = 1|x) of the positive class for the TTA approach represents the risk score of glaucoma used for the creation of the TTA-based ROC curve. The usage of TTA can be considered as an additional insurance; it increases the accuracy of classifiers since the prediction becomes independent of the image rotation angle. It plays a vital role especially for the case of using images taken from different cameras and positions of patients.

3.10 Evaluation Metrics

Model performance is evaluated on a held-out test set using six metrics, with glaucoma (RG) treated as the positive class (TP, TN, FP, and FN denote true/false positive and negative predictions, respectively).

$$\text{Accuracy} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \qquad (7)$$

$$\text{Recall} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \qquad (8)$$

$$\text{Specificity} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \qquad (9)$$

$$\text{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \qquad (10)$$

$$\mathrm{F1-score} = \frac{2 \times (\text{Precision} \times \text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})} \qquad (11)$$

To assess the model's ability to discriminate between classes, we use the AUC-ROC curve, plotting sensitivity against (1 − specificity). In clinical practice, sensitivity and specificity matter most, since the cost of an error differs: missing a glaucoma patient can lead to irreversible vision loss, whereas a false alarm results only in an additional examination. We report results for both standard and TTA-based inference.

4. Results and Discussion

The DeiT-Small model is evaluated on an independent test set of 770 fundus images (385 normal and 385 glaucomatous), constituting 15% of the total dataset. Performance is compared using traditional single-pass inference and Test-Time Augmentation (TTA, N = 5).

4.1 Quantitative Evaluation on the Test Set

Performance details of the architecture with respect to the 770-image dataset are provided in Table (4) . Baseline performance was measured using the standard single-pass inference for all metrics used. The use of Test-Time Augmentation (TTA), which included five random passes, consistently led to improvements in all metrics, especially sensitivity and AUC-ROC. From a clinical perspective, the increase in sensitivity is of utmost importance as it shows that TTA is capable of picking up on more cases of glaucoma than single-pass inference. False negatives are especially problematic in glaucoma detection because they might lead to missing a diagnosis and proper assessment/treatment of the patient, possibly resulting in irreversible vision impairment.

Table (4): Model test-set performance: standard inference vs. Test-Time Augmentation.

Evaluation Metric	Standard	With TTA	Improvement
Accuracy	91.65%	93.51%	+2.60 pp
F1-score	91.55%	93.50%	+2.76 pp
Sensitivity (Recall)	91.56%	93.51%	+1.95 pp
Specificity	95.06%	96.36%	+1.30 pp
AUC	95.84%	96.66%	+0.82 pp

4.2 Confusion Matrix Analysis

In order to have a better understanding of where the model succeeds and fails, the confusion matrix for both standard and TTA modes was analyzed.

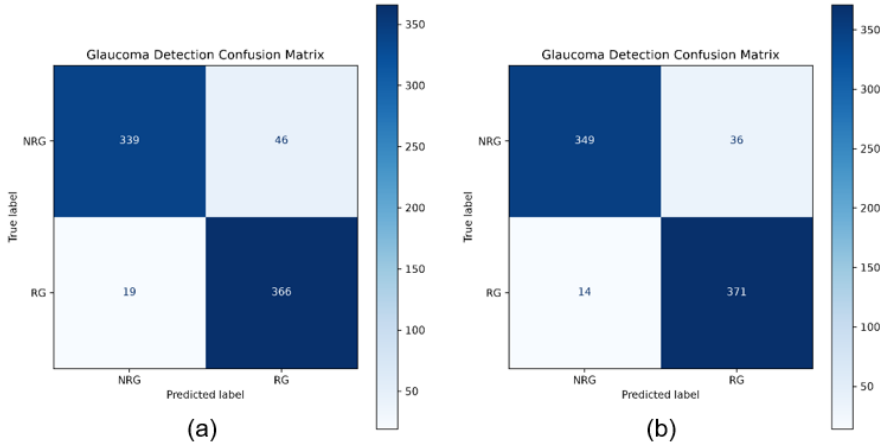


Figure (5): Confusion matrices for glaucoma classification (a) without TTA and, (b) with TTA

The results are illustrated in Figure 5. In standard (one-pass) mode, the model is efficient, and it is able to detect 339 out of 385 true negatives for NRG cases, and 366 out of 385 true positives for RG cases, while generating 46 false positives and 19 false negatives. with the application of Enabling Test-Time Augmentation, there is a noticeable improvement in all four quadrants:

- TP improves from 366 to 371 (+5 detections of positives).
- TN improves from 339 to 349 (+10 detections of healthy people).
- FP decreases from 46 to 36 (-10 cases creating anxiety to the patients).
- FN decreases from 19 to 14 (-5 false negatives).

Clinical significance lies in the following: with 14 false negative cases, the sensitivity of the model for glaucoma is 96.36%. In any screening process, a failure to detect the patient is considered the worst scenario since glaucoma may lead to permanent blindness if left untreated.

4.3 Clinical Significance

With the TTA system, the recall of glaucoma is estimated at 96.36%, with successful detection of 371 out of 385 glaucoma patients, or, in other words, about 96 out of every 100 glaucoma patients. High recall is an essential characteristic for a screening system because any undetected or late detection of glaucoma can prevent proper treatment and lead to irreversible vision loss. Nevertheless, the system achieves a specificity of 90.65%, which means correct identification of 349 out of 385 healthy patients. The false-positive rate of the model is 9.35%, or, in other words, about 9 out of every 100 healthy patients are incorrectly diagnosed as having glaucoma and need additional examination by the clinician. This is a decent specificity for a screening system, especially in clinical

settings. moreover, the maximum AUC-ROC score of 96.66% indicates excellent discrimination capacity of the system regardless of the classification threshold. This enables selecting the most appropriate threshold according to the clinical needs.

4.4 Benefit of Test-Time Augmentation

The benefit of TTA is that it delivers a performance boost at zero cost without any extra training data, architectural changes, or retraining. On the 770-image test set, enabling TTA reduced the number of missed glaucoma patients (false negatives) from 19 to 14, a 26.3% decrease.

Similarly, TTA reduced the number of false-positive glaucoma predictions from 46 to 36, a 21.7% decrease. Because this technique only requires a few quick geometric transforms of the input image, Benefits of TTA have been observed, which validate the use of TTA in a clinical setting; however, its usefulness must be validated externally using real world data sets that show glaucoma distribution.

4.5 Limitations and Future Work

The primary limitation of this work is that the model was tested on a single dataset and has not been validated on external benchmark datasets such as REFUGE, RIM-ONE DL, or ACRIMA. Demonstrating performance across other camera types and demographic groups will require cross-dataset validation in future work. The system will also be extended for glaucoma severity staging and multimodal data input, including optical coherence tomography images and intraocular pressure measurements.

4.6 Grad-CAM Explain ability

To verify interpretability and clinical reliability, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to all test cases in both output classes. As shown in Figure (6) , the resulting activation maps confirm that the network focuses on anatomically relevant regions — the optic disc, optic cup, and peripapillary retinal nerve fibre layer — the structures actually affected by glaucomatous damage. Anatomically relevant activations are observed in both the Referral Glaucoma (RG) and Normal (NRG) classes; in particular, the RG class shows highly concentrated salient regions along the cup-disc border and areas of rim thinning.

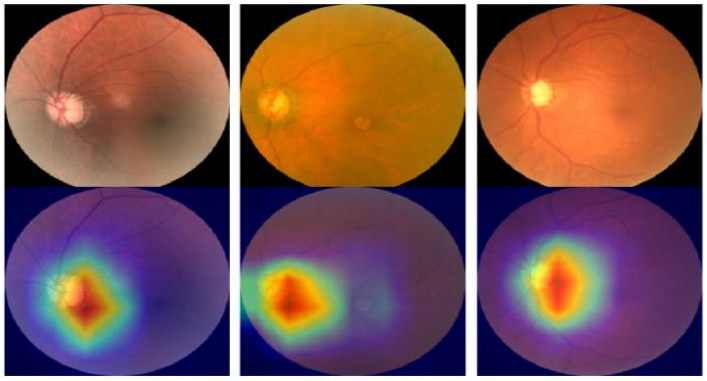


Figure (6): Grad-CAM activation maps for representative fundus images illustrating model attention over the optic disc.

5. Conclusion

This paper proposed an efficient glaucoma-classification pipeline by fine-tuning a DeiT-Small Vision Transformer. The first four transformer layers were frozen while the rest were fine-tuned using LLRD and a classification head as an efficient fine-tuning strategy. On the independent test data consisting of 770 images, the architecture obtained an accuracy of 91.65%, F1-score of 91.55%, sensitivity of 91.56%, specificity of 95.06% and AUC-ROC of 95.84%. Test-Time Augmentation showed promising results in being an inference time technique for increasing robustness without extra model training and architecture changes. this paper illustrated that it is possible to attain high glaucoma screening performance with low computational cost, showing the feasibility of utilizing small-efficient transformers for practical applications in healthcare. But since the dataset used in the evaluation process was artificially balanced, the predictive performance obtained from the evaluation

may not be realistic due to the imbalance of the actual screening data. Further research will be carried out to externally validate our approach using real-world imbalanced datasets and multimodal inputs.

References

[1] R. R. A. Bourne et al., "Global estimates on the number of people blind or visually impaired by glaucoma: A meta-analysis from 2000 to 2020," Eye, vol. 38, no. 11, pp. 2036–2046, 2024, doi: 10.1038/s41433-024-02995-5.

[2] 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, Nov. 2014, doi: 10.1016/j.ophtha.2014.05.013.

[3] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," British Journal of Ophthalmology, vol. 90, no. 3, p. 262, Mar. 2006, doi: 10.1136/bjo.2005.081224.

[4] American Academy of Ophthalmology, Primary Open-Angle Glaucoma Preferred Practice Pattern Guidelines, San Francisco, CA, USA, 2020.

[5] European Glaucoma Society, Terminology and Guidelines for Glaucoma, 5th ed., British Journal of Ophthalmology, vol. 105, Suppl. 1, pp. 1–169, 2021.

[6] D. R. Anderson, "Normal-tension glaucoma (low-tension glaucoma)," Survey of Ophthalmology, vol. 48, no. 2, pp. S52–S57, 2003.

[7] R. S. Harwerth, E. L. Carter-Dawson, J. K. Shen, J. M. Smith, and R. L. Crawford, "Ganglion cell losses underlying visual field defects from experimental glaucoma," Progress in Retinal and Eye Research, vol. 18, no. 4, pp. 443–464, 1999.

[8] Z. Li et al., "Development and validation of a deep learning system for detecting glaucomatous optic neuropathy using fundus photographs," JAMA Ophthalmology, vol. 136, no. 12, pp. 1363–1370, 2018.

[9] M. Raghu et al., "Transfusion: Understanding transfer learning for medical imaging," Nature Medicine, vol. 25, pp. 764–769, 2019.

[10] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models reveals their strong shape bias," arXiv preprint arXiv:1904.00760, 2019.

[11] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.

[12] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," ICML, 2021.

[13] S. Pathan, P. Kumar, R. M. Pai, and S. V. Bhandary, "An automated classification framework for glaucoma detection in fundus images using ensemble of dynamic selection methods," Progress in Artificial Intelligence, vol. 12, no. 3, pp. 287–301, 2023, doi: 10.1007/s13748-023-00304-x.

[14] L. Pascal, O. J. Perdomo, X. Bost, B. Huet, S. Otálora, and M. A. Zuluaga, "Multi-task deep learning for glaucoma detection from color fundus images," Scientific Reports, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-16262-8.

[15] Y. Li, Y. Han, Z. Li, Y. Zhong, and Z. Guo, "A transfer learning-based multimodal neural network combining metadata and multiple medical images for glaucoma type diagnosis," Scientific Reports, vol. 13, no. 1, p. 12076, 2023, doi: 10.1038/s41598-022-27045-6.

[16] V. K. Velpula and L. D. Sharma, "Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion," Frontiers in Physiology, vol. 14, 2023, https://doi: 10.3389/fphys.2023.1175881 .

[17] S. Saha, J. Vignarajan, and S. Frost, "A fast and fully automated system for glaucoma detection using color fundus photographs," Scientific Reports, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-44473-0.

[18] R. Fan et al., "Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization," Ophthalmology Science, vol. 3, no. 1, p. 100233, 2023, doi: 10.1016/j.xops.2022.100233.

[19] S. Chakraborty, A. Roy, P. Pramanik, D. Valenkova, and R. Sarkar, "A dual attention-aided DenseNet-121 for classification of glaucoma from fundus images," arXiv preprint arXiv:2406.15113, Jun. 2024.

[20] K. J. Tina Brivitha and S. Sophia, "Integrated Deep Learning Framework for Automated Glaucoma Detection, Optic Disc/ Cup Segmentation, and CDR Calculation," in Proceedings of International Conference on Visual Analytics and Data Visualization, ICVADV 2025, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 887–894. doi: 10.1109/ICVADV63329.2025.10961500.

[21] R. Kiefer, “Glaucoma dataset: EyePACS-AIROGS-light-V2,” Kaggle Dataset, 2024. Available: https://www.kaggle.com/datasets/deathtrooper/glaucoma-dataset-eyepacs-airogs-light-v2
[22] H. Touvron, M. Cord, and H. Jégou, “DeiT III: Revenge of the ViT,” in Computer Vision – ECCV 2022, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 516–533.
[23] Y. Li and Q. Zhou, “Estimation of Doppler velocity from incoherent scatter spectra using context-aware transformers,” Atmos Meas Tech, vol. 19, no. 11, pp. 3865–3874, Jun. 2026, doi: 10.5194/amt-19-3865-2026.
[24] I. Loshchilov and F. Hutter, “DECOUPLED WEIGHT DECAY REGULARIZATION.” [Online]. Available: https://github.com/loshchil/AdamW-and-SGDW