# Contrastive Vision Transformer Combined with Hyperparameter Fine-Tuning and Interpretable AI for Glaucoma Assessment

R. Roopalakshmi *, Ayush Amarnath Bhagat, and Sambhav Nath Jain

School of Computer Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education (MAHE), Manipal, Karnataka, 576104, India

Email: roopalakshmi.r@manipal.edu (R.R.); ayush.mitmpl2022@learner.manipal.edu (A.A.B.); sambhav.mitmpl2022@learner.manipal.edu (S.N.J.)

*Corresponding author

***Abstract*—Glaucoma, often known as the 'silent thief of sight', is a leading cause of irreversible blindness, which affects around 80 million people worldwide and hence its early and reliable detection is critical for preventing permanent vision loss. Although Vision Transformer (ViT)-based deep learning models are employed by existing techniques for automated glaucoma screening, yet, they primarily rely on conventional supervised training and single-dataset evaluation, which result in limited feature discrimination, suboptimal generalization across heterogeneous images, and poor clinical interpretability. To address these limitations, this study proposes a novel contrastive learning-optimized ViT framework, which integrates supervised contrastive pre-training with systematic hyperparameter optimization to learn more discriminative retinal features, and fine-tuning for glaucoma classification. In addition, a unified preprocessing and patch-based representation strategy is introduced to mitigate domain shifts across multiple imaging devices and acquisition protocols. Unlike prior studies using single benchmarks, this framework is validated on comprehensive multi-dataset setting combining six public fundus datasets (including G1020, ORIGA, REFUGE, PAPILA) to assess real-world generalization. Experimental results demonstrate consistent and statistically significant improvements over Convolutional Neural Network (CNN), baseline ViT models, in terms of achieving up to 87.91% accuracy and performance gains of 3-16% across accuracy, precision, recall, and F1score metrics. Further, Layer-wise Relevance Propagation (LRP) is employed to generate clinically interpretable heatmaps, which confirms that the model focuses on anatomically meaningful regions such as the optic disc and optic nerve head. These findings prove that the proposed framework provides robust, explainable, and generalizable solution for automated glaucoma screening and highlights its potential for clinical deployment.**

***Keywords*—Glaucoma detection, vision transformers, contrastive learning, medical imaging, explainable AI, layer-wise relevance propagation, deep learning, convolutional neural network**

## I. Introduction

Glaucoma is one of the leading causes of blindness globally, which is marked by the progressive degeneration of optic nerve fibers and results in permanent vision loss [1]. Although glaucoma is progressive in nature, yet it remains a treatable disorder and early diagnosis followed by timely treatments can effectively prevent severe visual impairments. The term 'Glaucoma' originates from the Greek word 'glaukos', meaning 'grayish-blue'. Glaucoma is broadly described as a progressive optic neuropathy, which leads to irreversible blindness by damaging the optic nerves and the peripheral visual field [2]. In developed nations, it affects over 3% of the population and ranked as the second most common cause of permanent blindness worldwide [2]. Recent reports indicate, approximately 80 million individuals were affected by glaucoma in 2020, and this count is expected to increase to 111 million by 2040 [2]. In Asia alone, an estimated 27.8 million additional cases are expected by 2040, with India and China bearing the highest share of this burden [3]. In India specifically, glaucoma accounts for blindness in 1.2 million individuals, which represents about 5.5% of all blindness cases that result in irreversible vision loss [4]. The most common type of glaucoma disease is, *Primary Open-Angle Glaucoma (POAG)*, which results in irreversible vision loss at its end stage. Specifically, POAG is chronic in nature, which involves the loss of optic nerve fibers and progresses over time [5]. The major cause for POAG disorder is, high Intraocular Pressure (IOP) in the eye, which arises due to the imbalance between the production and drainage of fluid in the eye [5]. Typically, POAG begins with blockage in the drainage canal of eye, in which IOP increases, damages optical nerve and eventually leads to vision loss. When the glaucoma disease advances, the vision field deteriorates and generates blurred or darkened areas in the patient's sight.

Manuscript received October 13, 2025; revised November 20, 2025; accepted March 2, 2026; published June 17, 2026.

Glaucoma often remains asymptomatic, until the disease reaches an advanced stage. If it is left undiagnosed and untreated, then the disease progresses rapidly, causes early visual impairment and eventually leads to irreversible blindness. For instance, in India, nearly 90% of glaucoma cases are remaining undetected in the present decade [3]. Hence, early diagnosis and timely treatment are crucial to prevent the glaucoma-related vision loss. Common procedures of glaucoma diagnostics include Ophthalmoscopy, Tonometry, Gonioscopy, Perimetry, and Pachymetry [6]. Tonometry measures Intraocular Pressure (IOP), in which IOP greater than 21 mmHg is typically considered as glaucomatous, while ophthalmoscopy assesses optic nerve damage. Gonioscopy evaluates the drainage angle and helps to distinguish between open-angle and angle-closure glaucoma. Pachymetry is used to determine corneal thickness of the eye. Perimetry-commonly known as the visual field test, assesses the extent of visual field loss in glaucoma suspected patients [6]. In recent years, Optical Coherence Tomography (OCT) and color Fundus imaging are widely used for ocular diseases diagnosis including glaucoma disorder [7]. However, manual diagnosis of glaucoma from color fundus images is quite challenging, because of its cost-expensive nature and limited availability. Due to the shortage of ophthalmologists, color fundus imaging is becoming a preferred screening method, which offers greater accessibility and cost-effectiveness when compared to the other modalities. However, manual diagnosis of fundus images remains labor-intensive, which demands specialized medical expertise and extensive training. Techniques such as image segmentation, transfer learning, and image classification have demonstrated considerable success in this field, which support clinicians and expedite glaucoma detection process.

Thanks to the rapid developments in Artificial Intelligence (AI) field, recently Machine Learning (ML) and Convolutional Neural Network (CNN) models using fundus imaging are achieving higher performance in accurately detecting the glaucoma disease [8, 9]. Although CNNs-based glaucoma assessment techniques are performing better, yet, CNNs often fail to focus on the generalization aspects of unexplored fundus images [10]. Several practical cases indicate that, the optic nerve-head photographs based glaucoma detection carried out by CNNs may produce inconsistent results, even among experienced ophthalmologists. This highlights the urgent need for methods, which can enhance the generalization capability of prediction models for color fundus images. The segmentation-based approaches are also used in glaucoma prediction, which focus on Optic Disc/Cup localization without relying on explicit anatomical segmentation [9]. However they may result in lower accuracy in case of bright artifacts in fundus images. To address these challenges, the Vision Transformer (ViT) architecture [11] is recently introduced, which gained huge attention in the glaucoma detection literature. Typically, ViTs effectively capture relevant patterns from training data and encode structural information through positional embeddings. specifically, CNNs rely on local receptive fields confined to small grid regions, whereas ViTs utilize receptive field of full image and learns more comprehensive visual features and thereby results in improved prediction accuracies for glaucoma detection problem.

### *A. Motivation and Contributions*

In the existing literature on Glaucoma detection, only a few attempts have been made to predict glaucoma from fundus images, using vision transformer architectures [6, 8, 12, 13]. However, state-of-the-art techniques for glaucoma detection are less focused towards the issues such as: a) incorporating optimization strategies for performance enhancement, b) Improving feature representations during model training and c) Using Explainable AI frameworks for interpreting the predictions of the model. To solve these issues, this study proposes a new Contrastive Learning-based optimized ViT architecture combined with fine-tuning of hyperparameters, which scores better results for the glaucoma detection problem. Specifically, our study advances the existing literature in the following ways:

Methodological Novelty: Unlike state-of-the-art glaucoma detection studies that directly fine-tune Vision Transformers using supervised learning, this article propose a contrastive learning–driven feature representation framework integrated with ViT, which enables the model to learn discriminative retinal patterns before classification. This improves inter-class separability and robustness to cross-dataset variations, which has not been systematically explored in earlier ViT-based glaucoma studies.

Optimization strategy: We introduce a structured hyperparameter optimization and training strategy combined with supervised contrastive objectives, instead of standard end-to-end training, which results in consistent performance gains over baseline ViT architectures.

Comprehensive multi-dataset generalization analysis: Most previous works evaluate on single datasets. We perform cross-dataset training and evaluation across six benchmark fundus datasets, which demonstrates improved generalization and clinical reliability. Specifically, model evaluations are conducted on a comprehensive database, combining images from 6 different benchmark datasets including G1020, ORIGA, REFUGE and PAPILA datasets, when compared to its counterparts which employ only limited datasets. The results of experiments, clearly demonstrate the efficiency of the proposed model in terms of Accuracy, precision, Recall, F1-Score, sensitivity and specificity.

Explainability integration: Beyond accuracy improvements, we incorporate Explainable AI technique known as, Layer-wise Relevance Propagation (LRP) [14] to provide clinically interpretable heatmaps, for validating that the model focuses on anatomically meaningful regions (optic disc and cup). This bridges the gap between deep learning predictions and clinical trust, thereby demonstrates model's robustness and potential for future clinical integrations.

In this way, these contributions emphasize both a new learning framework and practical insights into robust,

explainable glaucoma screening, which extends beyond incremental architectural modifications.

## II. Literature Review

Early studies on glaucoma diagnosis primarily relied on conventional machine learning methods such as K-means clustering for detecting the disease. For example, Praveena and Babu [15] developed a K-means-based framework integrated with a hill-climbing strategy to extract optic disc contours for detecting glaucoma, which suffers due to scalability limitations. In recent years, CNNs have gained prominence in glaucoma detection domain due to their superior ability to capture local features which are crucial for the image classification task. For instance, An *et al.* [7] applied a CNN-based approach on OCT data and fundus images using retinal nerve fiber layer, ganglion cell thickness maps, which scored better detection results, yet faces generalization challenges. Similarly, Fan *et al.* [16] employed CNNs with convolutional layers to represent fundus images as 1-D visual features, which offers enhanced spatial representations but misinterprets few features due to inadequate pixel connection encoding. Nayak *et al.* [17] proposed an ECNet-based framework by combining Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers, which suffers due to scalability issues. Christopher *et al.* [18] investigated three pre-trained CNN models- InceptionV3, ResNet50, and VGG-16, which are trained with randomly initialized weights for glaucoma detection. Phan *et al.* [19] analysed trained CNNs to identify glaucomatous conditions from fundus images, by examining the effect of image size and quality on the detection process. Further, Li *et al.* [20] introduced an automated system for detecting glaucomatous optic neuropathy using fundus images and Inception-V3 based CNN models.

In the current glaucoma diagnosis research, Vision Transformer (ViT) architectures are less explored in the domain. For example, Yu *et al.* [21] proposed a multiple-instance learning based ViT framework for retinal disease detection, which reports marginal improvements compared to conventional CNN approaches. Chincholi and Koestler [1] applied ViTs for glaucoma diagnosis using limited open-access datasets, while Wassel *et al.* [6] evaluated their method on six to seven datasets but relied solely on the baseline ViT architectures. Mallick *et al.* [12] investigated various transformer-based models for glaucoma assessment, yet their experiments are restricted to limited datasets. More recently, in 2023, Piyush *et al.* [8] employed lightweight ViTs for glaucoma diagnosis, though their study was limited to the base architecture and a relatively small dataset of 1700 images.

In Ref. [22], the authors used CNN architecture with separable convolutional layers and increased filter size for improving the classification accuracy. However, this model manually crops the optic disc, which causes data loss related issues. Very recently, Chen *et al.* [23] introduced a ResNet-50 based algorithm to detect glaucoma in highly myopic eyes, in which Grad-CAM was utilized for visual interpretation. Though this methods scores well, yet it fails to deal with explainability aspects of predictions. Recently, Shoukat *et al.* [24] utilized ResNet-50 based deep learning model for identifying subtle indicators in early-stage glaucoma across multiple datasets. Though their technique resulted in high accuracy and versatility, yet it suffers from issues such as less training data, single architecture, and lack of external, clinical validations. To summarize, although very few studies are carried out on glaucoma detection in the literature using ViT architectures, yet they suffer from optimization, model training and labeled-data dependency issues. From another perspective, when ViTs are trained using smaller datasets, they suffer from inductive bias issues and result in slightly lesser scores, when compared to CNNs based approaches [9, 10]. To solve these problems, this study proposes an efficient glaucoma diagnosis framework, which utilizes contrastive learning based optimized ViT architecture integrated with hyperparameter fine-tuning, as detailed in the subsequent sections.

## III. Methods and Materials

### A. Proposed Methodology

The block diagram of the proposed glaucoma detection framework is shown in Fig. 1, which specifies the customized version of ViT architecture for Glaucomatous eye disease prediction. Precisely, the input image is divided into image patches, which are flattened and linearly projected on to the transformer block by the patch encoder. For each of the flattened patch, positional embedding is also generated which is class-specific learnable embedding in nature. Once positional embedding for all flattened patches of the given image is generated, it is considered as the input sequence and fed in to the transformer encoder block. Specifically, input image patch and its positional embedding are indicated as Orange, pink colored oval shapes numbered from 0–9 in Fig. 1, which are inputted into the encoder block. The transformer encoder consists of three main processing elements- Norm, Multi-head attention and Multilayer Perceptron (MLP) blocks, which are indicated in the left portion of Fig. 1.

Norm block, is also called as, Layer Norm facilitates the training process and enhances adaptability of model towards multiple variations in the training images. The Multi-head Attention block is mainly responsible for the creation of attention maps from the input embedded visual tokens. The attention maps are essential for focusing on the most critical sections in the image, which needs to be classified. The MultiLayer Perceptron (MLP) block is a classification network with Gaussian Error Linear Unit (GeLU) at the end. Precisely, Position-wise Feed-Forward Network is employed in this framework for this glaucoma classification task. The Last stage of MLP block is also termed as MLP head, which represents the output of transformer in the form of classification labels with help of softmax layers. Specifically, the output classification labels including Glaucomatous Eye and Normal Eye are indicated as “Glaucoma” and “Normal”, as indicated in orange colored oval shapes in Fig. 1.

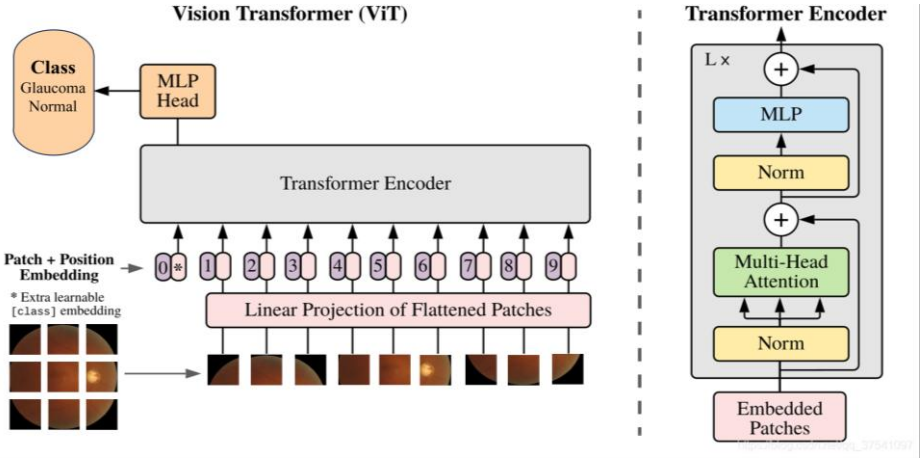



Fig. 1. Customized version of ViT architecture for Glaucoma detection.

Fig. 2 flowchart illustrates the step-wise procedure involved in the proposed study. Initially, the input images from glaucoma database are converted into image patches as shown in Fig. 2. Then, the image patches are subjected to pre-processing activities such as normalization, rescaling and data augmentation. The resultant input patches are split into training and testing datasets for evaluation purposes. After this step, positional embeddings are added for every patch and the resultant flattened patches are fed as input into the transformer encoder block, as indicated in Fig. 2.

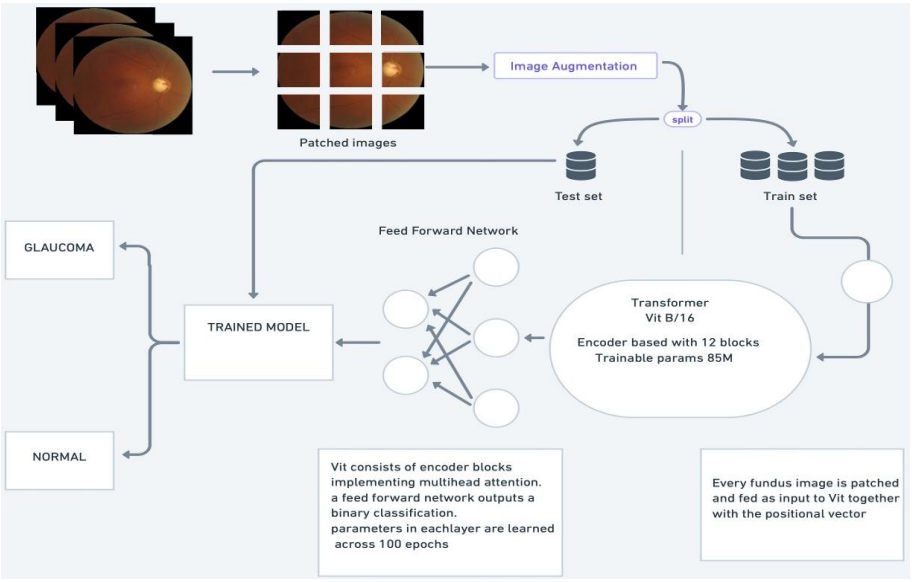



Fig. 2. Flowchart showing working of the proposed model.

Specifically, the ViT B16 model is utilized in the proposed study, in which encoder employs 12 blocks for optimizing the binary classification results. More specifically, each encoder block includes the Layer Normalization, Multi-Head Self-Attention mechanism and Position-wise Feed-Forward Network modules for implementing the classification task. The final classification layer is a Dense layer with a 'sigmoid' activation function, which optimizes the final binary classification result of the trained model in terms of output classes- Glaucoma and Normal images respectively as illustrated in Fig. 2.

ViT Configuration & Architectural Details: To promote experimental transparency, the complete architectural and training configuration of the proposed framework is described as follows: This framework employs ViT-B16 backbone, where each fundus image is resized to 224×224 pixels and divided into 16×16 non-overlapping patches, which results in 196 tokens per image. Each patch is linearly projected to a 768-dimensional embedding space, and learnable positional embeddings are added. The encoder consists of 12 transformer blocks, in which each block comprises 12 attention heads, the hidden size of 768, the MLP dimension of 3072, followed by Layer Normalization and residual connections. The final classification head consists of a fully connected layer with sigmoid activation for binary prediction. During Preprocessing and Data Augmentation stage, all images are resized to 224×224, and the pixels are normalized to [0,1] range. During training, stochastic augmentations

including random horizontal flip, random rotation (±15°), random crop (scale 0.8–1.0), color jitter are applied with probability 0.5. These augmentations are used both for supervised training and for generating contrastive pairs.

### *B. Contrastive Learning-Based Optimization Algorithm for ViT-B16 Architecture*

In the proposed glaucoma detection framework, Optimization of ViT-B16 architecture is carried out by considering different strategies including data augmentation, feature selection, contrastive loss function calculation as detailed in Algorithm 1. In general, contrastive learning is a self-supervised learning strategy, that helps to learn useful representations by contrasting positive and negative pairs [25]. The core idea of contrastive learning is to bring the representations of similar (positive) pairs closer together in the embedding space while pushing the representations of dissimilar (negative) pairs farther apart. The Contrastive learning-based Optimization algorithm utilized in the proposed study for glaucoma diagnosis is illustrated in detail in Algorithm 1 of Section III.B. Specifically, in this work, positive pairs are generated by applying stochastic data augmentation techniques such as random cropping, rotation, and horizontal flipping to the same retinal fundus images, which results in creation of multiple correlated views. These augmentations encourage the model to focus primarily on clinically relevant structural patterns instead of superficial variations. Further, Negative pairs are constructed using augmented views from different retinal images, which introduce dissimilar anatomical characteristics and disease patterns. In this way, separation between these negative samples helps the model to learn more discriminative embeddings, which results in improved feature representation and classification performance for glaucoma assessment.

**Algorithm 1: Contrastive learning-based optimization algorithm for ViT-B16 model**

**Step 1: Data Collection and Pre-processing:**

Let DB represents the input database such that,

$$DB = \sum_{i=1}^{n} G \cup H \mid i = 1, 2, \ldots, n) \quad (1)$$

where *G,H* indicates Glaucomatous, healthy eyes, and *n* indicates the total no. of the images respectively. The input images are subjected to pre-processing activities including normalization and resizing as described in Section IV.B.

**Step 2: Utilization of Model architecture and Learning Framework:**

In this framework, ViT B16 architecture is utilized, which splits input images into 16×16 patches and processes the patches with transformer encoders. MLP is used as the projection head for mapping the output embeddings to a lower-dimensional space, which is suitable for contrastive learning. Furthermore, Simple Framework for Contrastive Learning of Visual Representations (SimCLR) is employed in this framework due to its less complex nature when compared to Momentum Contrast for Unsupervised Visual Representation Learning (MoCo) framework [22].

**Step 3: Data Augmentation and Creation of Positive, Negative Pairs:**

**Problem:** Positive & Negative pairs generation to implement contrastive learning technique.

**Solution:** The Positive pairs are generated by employing the augmented views of the same retinal image, after applying cropping, rotation and horizontal flipping transformations to the DB images. The Negative pairs are created using augmented views from different retinal images, which introduces dissimilarity in the images.

**Step 4: Computation of Contrastive Loss function:**

**Problem:** Suitable function is needed to compute the similarities between pairs.

**Solution:** The contrastive loss function computes the similarities between all pairs of embeddings in a batch by minimizing the distance between positive pairs and maximizing the distance between negative pairs [25]. We used Normalized Temperature-scaled Cross Entropy (NT-Xent) loss function in this framework.

**Step 5: Selection of the Temperature hyperparameter:**

**Problem:** Selecting appropriate value for the hyper-parameter 'Temperature', which controls the sharpness of the distribution.

**Solution:** After conducting multiple trials, temperature parameter in the Contrastive loss function is set as 0.5.

**Step 6: Entropy-based Fine-tuning:**

In this framework, initially, the model is pre-trained using the contrastive learning framework and then it is fine-tuned on a labelled dataset using cross-entropy loss, in order to adapt the learned representations for the specific glaucoma detection task.

**Step 7: Hyperparameters Optimization:**

In this framework, after implementing multiple trials for hyper-parameters values, the finalized ones are given by:

$$HP_O \subset \{BS = 64,\ Lr = 1e-3,\ T = 0.5\} \quad (2)$$

where $HP_O$ indicates the optimized hyper-parameters, $BS$ indicates the batch size, $Lr$ indicates the learning rate, $T$ indicates the temperature scaling factor and Adam optimizer is employed in this framework.

Contrastive training details:

To support reproducibility, contrastive training consisting of 2 stages are detailed as follows:

During Stage 1 Pre-training, parameters considered are:

- loss: contrastive loss
- epochs: 100
- batch size: 64
- optimizer: AdamW
- learning rate: 3e−4
- weight decay: 1e−4

During Stage 2 Supervised Fine-tuning, parameters are set as:

- loss: binary cross-entropy
- epochs: 50
- learning rate: 1e−5
- early stopping patience: 10

A cosine learning rate scheduler is applied in both stages. In addition, Five-fold cross-validation is performed for validating robustness. This framework is implemented

in PyTorch, trained using CUDA 11.8 on NVIDIA A100 GPU with 24 GB memory.

### C. Database Creation

To evaluate the performance of the proposed glaucoma detection framework, a comprehensive database is created by collecting fundus images from various open-access and private datasets. Specifically, the input database comprising 4500+ images of glaucomatous and normal eye conditions is constructed from 6 benchmark datasets, which are listed in Table I. The snapshot of sample fundus images from the input Database is shown in Fig. 3. In Fig. 3, the different stages of severity of Glaucoma eye disease are clearly indicated in the form of multiple colored (orange, red, gray) sample images.

TABLE I. EXPERIMENTAL DATABASE CONSTRUCTED FROM 6 BENCHMARK DATASETS

| Dataset Name | Description | Total Images |
|---|---|---|
| BEH [26] | Bangladesh Eye Hospital Dataset | 634 |
| FIVES [27] | A Fundus Image Dataset for AI-based Vessel Segmentation Dataset | 800 |
| G1020 [28] Dataset | Kaiserslautern, Germany | 1020 |
| PAPILA Dataset [29] | Fundus images of both eyes | 244 |
| ORIGA [30] | Online Retinal fundus Image database for Glaucoma Analysis | 650 |
| REFUGE [31] | Retinal Fundus Glaucoma Challenge Dataset | 1200 |

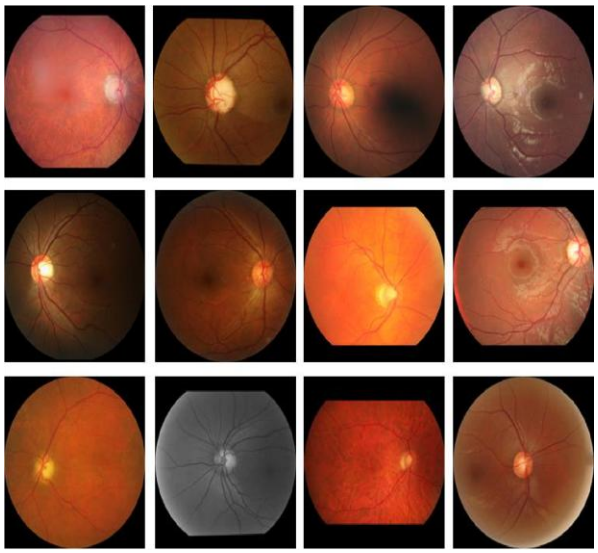


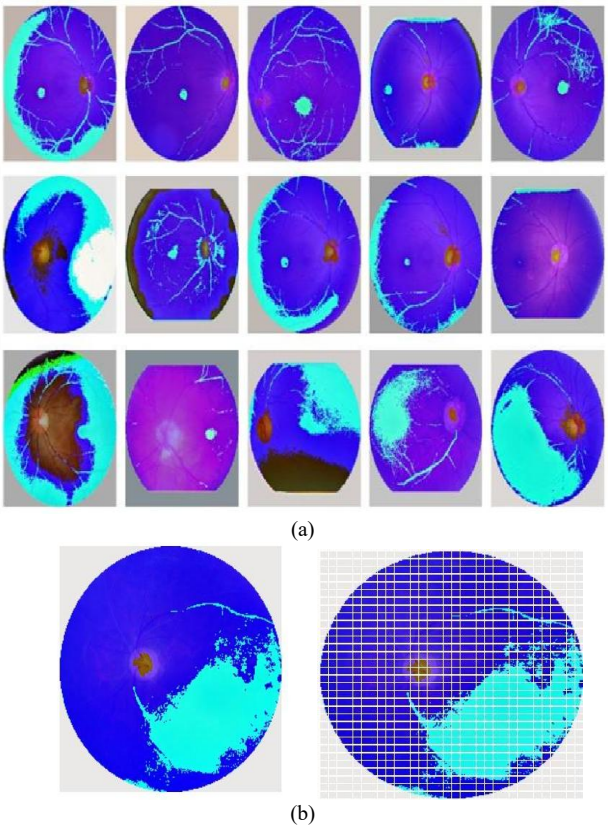
Fig. 3. Snapshot showing Glaucoma DB images with different disease severities.

**Problem:** Need to handle class imbalance issue, which is quite common in glaucoma datasets.

**Solution:** In our study, the combined experimental database is constructed from six public datasets, which is generated as a nearly balanced class distribution after dataset harmonization. Specifically, in the final 80/20 split, the held-out test set contains 951 glaucomatous and 953 normal images. Due to this near-balanced distribution, explicit re-sampling or cost-sensitive loss functions are not required. Instead, we employed stratified train-validation splitting strategy to preserve class proportions during training and validation. Additionally, extensive geometric and photometric data augmentation are applied uniformly across both classes, which increases sample diversity and mitigates potential bias and thereby enhances generalization.

### D. Pre-processing and Creation of Input Image Patches

In the proposed study, initially, the images are collected from the input database as described in Section III.C. In this study, domain shift is implicitly mitigated through a standardized preprocessing and representation learning pipeline, which is applied uniformly across all datasets. Specifically, all fundus images undergo normalization and pixel-level augmentations (brightness, contrast, and saturation adjustments) to reduce device-specific intensity variations. Further, all images are rescaled to a fixed spatial resolution of 224×224, which ensures consistent input dimensionality and thereby reduces resolution-induced domain discrepancies across datasets. Furthermore, after preprocessing, images are partitioned into non-overlapping 16×16 patches before being fed into the Vision Transformer (ViT). This patch-based tokenization enables the model to focus more on localized structural patterns such as optic disc and cup characteristics instead of global image statistics, which are more susceptible to domain-specific variations.



(a)

(b)

Fig. 4. Snapshot of sample input images, (a) after pre-processing; (b) after creating image patches.

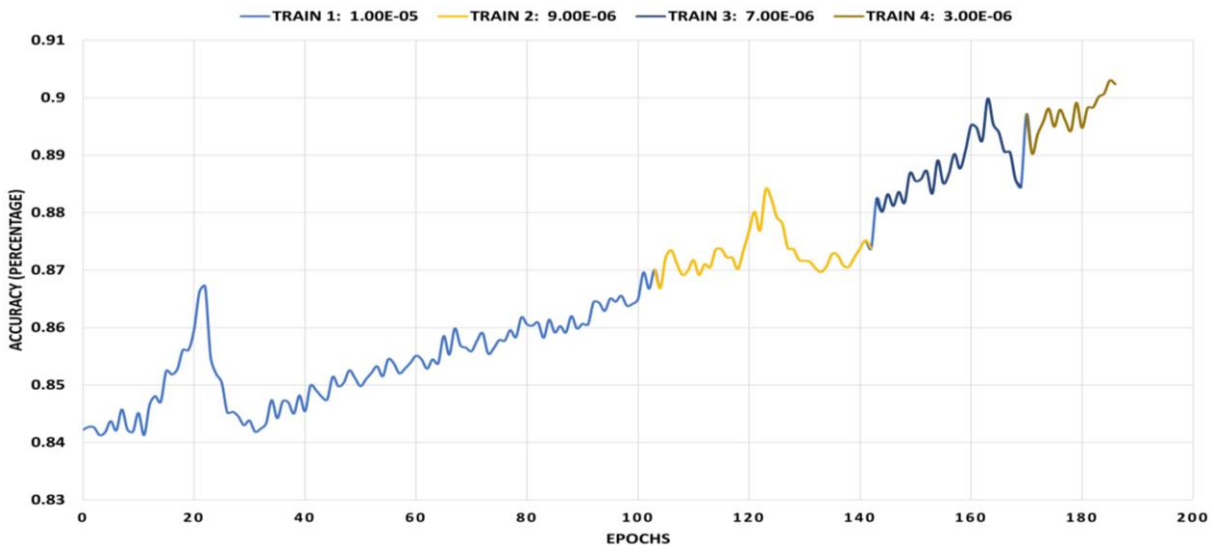
Fig. 4(a) shows snapshot of sample images after completing pre-processing activities. As indicated in Fig. 4(a), these preprocessing steps enhance visual

consistency across datasets, and thereby facilitate more reliable discrimination between glaucomatous and normal eyes. Although no explicit domain adaptation module is employed in this work, the combination of uniform preprocessing, aggressive data augmentation, and patch-level representation learning significantly helps to reduce inter-dataset variability and thereby improves the robustness of learned features across heterogeneous glaucoma datasets. Once the pre-processing operations are completed, the resultant input images are divided into non-overlapping 16×16 patches. Specifically, to enable the proposed ViT architecture to concentrate on smaller regions and quickly capture local features, the input image is divided into multiple patches. More specifically, Fig. 4(b) shows the snapshot of sample image before and after achieving patching in left, light portions respectively. From Fig. 4(b), it can be observed that, as the result of patching, the image elements of corresponding eye anomalies are clearly visible and thereby it facilitates the accurate glaucoma diagnosis.

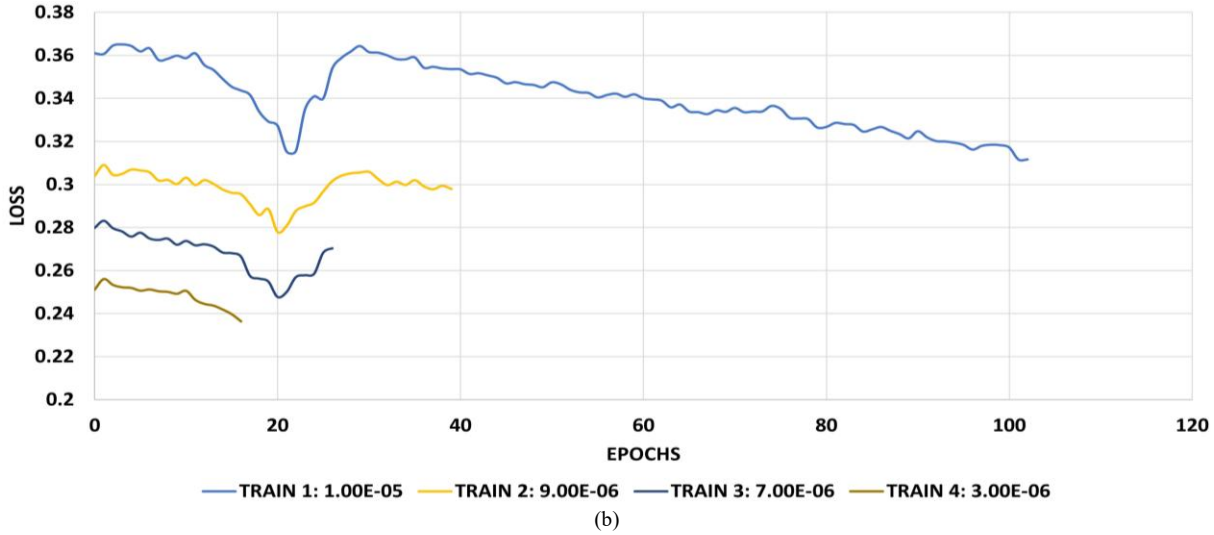
### *E. Hyperparameter Tuning*

Problem: The fine-tuning of hyperparameters such as 'Learning Rate' plays a crucial role in determining accuracy of the proposed model.

Solution: The optimal learning rate, which facilitates faster convergence, lowest loss and stable training is needed. To identify the optimal learning rate, in the proposed study, we evaluated the model in 4 different phases such that, in each cycle it is initialized and a complete training session is executed. Specifically, Fig. 5(a) and 5(b) indicates the model loss and accuracy values during the four different phases of training with 200 epochs. Similarly, Fig. 6(a) and 6(b) show loss and accuracy of model evaluation during four different testing runs with 200 epochs respectively. It can be observed from Figs. 5 and 6 plots indicating the loss and accuracy evaluations of the model during multiple training, testing phases that for Learning rate = 1.00E⁻05, the performance is better compared to other rates. In this study, this optimized learning rate is employed in the glaucoma diagnosis problem.




(a)




(b)

Fig. 5. Evaluation of Testing accuracy and Loss over four training cycles with varied learning rates, highlighting their influence on convergence behavior and performance consistency. (a) accuracy vs learning rates; (b) loss vs learning rates.

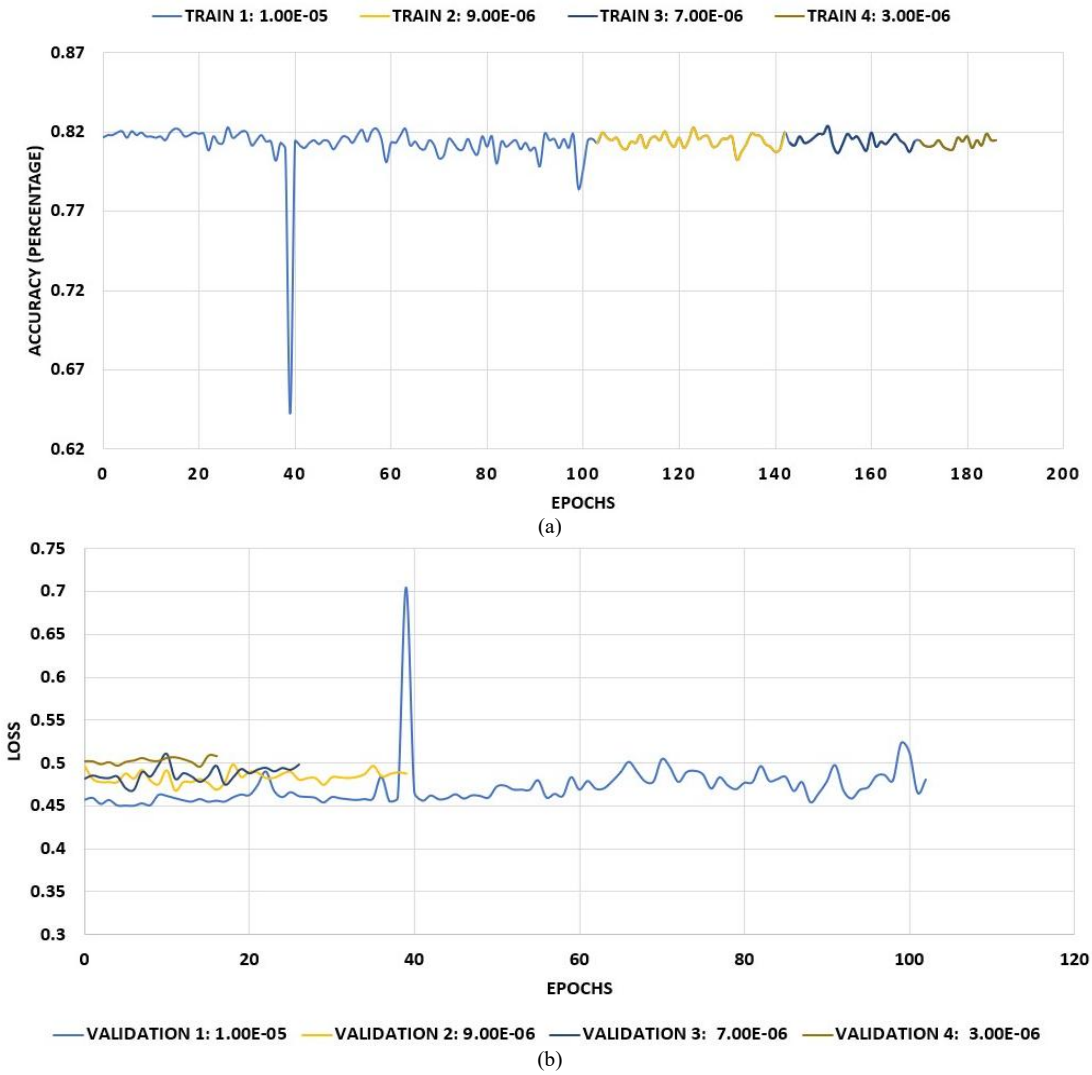



Fig. 6. Testing accuracy and loss trends across four training cycles with varying learning rates, illustrating the impact of learning-rate selection on model convergence stability and generalization performance. (a) accuracy vs learning rates; (b) loss vs learning rates.

In this study, in addition to learning rate, the following key hyperparameters are optimized during the contrastive pre-training and supervised fine-tuning stages, in order to achieve stable and effective representation learning, as given by: a) batch size, b) number of epochs with early stopping, c) optimizer choice, d) temperature parameter in the contrastive NT-Xent loss, e) dense head layer sizes, f) dropout rate, and label smoothing factor. During contrastive pretraining, the learning rate and temperature parameter of the NT-Xent loss were tuned, since they directly influence embedding separation and training stability. We empirically found that, the learning rate of $1 \times 10^{-3}$ and a temperature of 0.5, combined with a batch size of 64, provided the best trade-off between convergence speed and positive, negative pairs discriminations. The optimized temperature improved inter-class separability by appropriately scaling similarity scores, while the selected learning rate ensured stable gradient updates during contrastive optimization.

During supervised fine-tuning, the ViT-B16 backbone kept unchanged, and only the classification head (130,665 trainable parameters out of 85,929,873 total parameters) was optimized to prevent overfitting on the limited medical dataset. Further, hyper parameters including optimizer choice, learning rate, batch size, regularization, and loss formulation were carefully tuned. Specifically, Rectified Adam optimizer with a learning rate of $1 \times 10^{-4}$ and batch size 16 enabled stable convergence of the classification head. The use of GELU activations, batch normalization, and a dropout rate of 0.5 improved generalization, while label smoothing (0.1) reduced overconfidence in predictions and enhanced robustness to label noise. In this way, this targeted hyperparameter tuning strategy significantly contributed to performance gains by improving feature discrimination ability during contrastive pre-training and enhancing generalization during supervised fine-tuning, and thereby resulted in a reliable glaucoma classification.

### *F. Architecture of the Proposed ViT-Based Model*

Table II describes the different configuration parameters utilized in the proposed model, including Layer roles, activation functions, No.of units and no.of parameters used in each of the layers. As indicated in Table II, initially in Layer 1, the total parameters trained are 85,798,656 parameters using 768 units/filters.

However, as the layer advances, the no.of units employed and the no.of parameters trained is getting reduced, which eventually results in binary classification output (i.e. Glaucoma Vs Normal images).

TABLE II. LAYER-WISE ARCHITECTURAL CONFIGURATION AND PARAMETER DISTRIBUTION

| Layer | Role | Activation Function | No. of Units/ Filters | No. of Parameters |
|---|---|---|---|---|
| Layer 1 | Functional (vit-b16) | GeLU | 768 | 85798656 |
| Layer 2 | Flatten | - | 768 | - |
| Layer 3 | Dense | GeLU | 148 | 113812 |
| Layer 4 | Normalization | - | 148 | 592 |
| Layer 5 | Dropout | - | 148 | - |
| Layer 6 | Dense | GeLU | 84 | 12516 |
| Layer 7 | Normalization | | 84 | 336 |
| Layer 8 | Dropout | - | 84 | - |
| Layer 9 | Dense | GeLU | 44 | 3740 |
| Layer 10 | Normalization | - | 44 | 176 |
| Layer 11 | Dense 3 (output) | Sigmoid | 1 | 45 |

## IV. RESULTS AND DISCUSSION

In this study, the proposed ViT-based model is initially assessed on each individual benchmark datasets (listed in Table I) to evaluate its effectiveness in the glaucoma detection task. Subsequently, the model is evaluated on the comprehensive database, which combines images from all six benchmark datasets, and the detection results are analysed. Finally, a comparative analysis of proposed model Vs existing baseline methods is carried out, as discussed in the following subsections.

### *A. Glaucoma Diagnosis Results on Individual Benchmark Datasets*

Table III shows the glaucoma detection results of the proposed framework for each of the individual datasets including: a) BEH dataset [26]; b) FIVES dataset [27]; c) G1020 dataset [28]; d) PAPILA dataset [29]; e) ORIGA dataset [30]; and f) REFUGE dataset [31] respectively, which are listed in Table I. Specifically, in this study, we evaluated our model across multiple datasets by considering two different training vs testing ratios: 80:20, 70:30 and presented the respective prediction results in Table III. For instance, the proposed glaucoma assessment model scored 59.82% accuracy for the BEH dataset [26], while the training Vs testing ratio of 80:20 is considered, as indicated in Table III. The specificity performance is slightly lower for this BEH dataset [26], which might be due to the introduction of inductive biases in the ViT architecture, when tested on the smaller datasets. The proposed model scored 63% accuracy for the G1020 dataset [28] for the training Vs testing dataset ratio of 80:20, as indicated in Table III. It can be observed from Table III results that, the proposed framework achieves better results of 71.8% accuracy for the ORIGA dataset [30], while using 80:20 dataset ratios for evaluating the model. Similarly, the proposed ViT-based model attained an accuracy of 66.07% on the PAPILA dataset [29] and demonstrated slightly better accuracy of 92.5% on the REFUGE dataset [31] respectively, while 80:20 dataset ratios are utilized for evaluations. The enhanced performance of the ViT model can be characterized as, its ability to capture global features more effectively, when the dataset size increases and in turn enhances its detection capability.

TABLE III. PERFORMANCE RESULTS OF PROPOSED MODEL ON DIFFERENT DATASETS (IN %)

| Dataset | Ratio | Accuracy | Precision | Recall | F1-Score | Specificity |
|---|---|---|---|---|---|---|
| BEH [26] | 70:30 | 57.39 | 74.04 | 61.6 | 67.24 | 47.05 |
| | 80:20 | 59.82 | 73.24 | 66.67 | 69.80 | 44.11 |
| FIVES [27] | 70:30 | 51.79 | 72.41 | 31.34 | 43.74 | 82.22 |
| | 80:20 | 70.0 | 84.21 | 64.0 | 72.72 | 80.0 |
| G1020 [28] | 70:30 | 62.83 | 69.73 | 84.26 | 76.30 | 11.02 |
| | 80:20 | 63.02 | 70.13 | 81.2 | 75.26 | 22.03 |
| PAPILA [29] | 70:30 | 66.07 | 86.36 | 66.28 | 74.99 | 65.38 |
| | 80:20 | 65 | 83.02 | 69.84 | 75.86 | 47.05 |
| ORIGA [30] | 70:30 | 70.31 | 79.31 | 80.99 | 80.14 | 40.02 |
| | 80:20 | 71.88 | 79.8 | 83.16 | 81.44 | 39.39 |
| REFUGE [31] | 70:30 | 91.67 | 93.36 | 97.69 | 95.47 | 37.5 |
| | 80:20 | 92.5 | 92.31 | 100 | 96 | 25.12 |

As indicated in Table III, on BEH Dataset, Recall improves with 80:20 ratio, which specifies that more true glaucoma cases are correctly detected. However, Specificity is low (44–47%), which indicates that, the model struggles to correctly identify non-glaucoma cases. In case of FIVES dataset, Recall increases drastically (i.e., 31.34% to 64%), which specifies the better detection of glaucoma cases with more training data. For G1020 dataset, the Accuracy of the model stays similar (63%), Recall is very high (81–84%), indicating detection of most of the glaucoma cases, whereas Specificity is extremely low (11–22%), showing a high false-positive rate. In case of PAPILA dataset, Precision and recall are balanced, with F1-Scores around 75%, whereas Specificity drops with 80:20 (from 65.38% to 47.05%), which indicates occurrence of more no.of false positives. On the ORIGA dataset, the model Accuracy slightly improves with more training data (70.31% to 71.88%), where both the Precision and recall are remaining consistent and eventually result in stable F1-Scores (81%). In case of REFUGE dataset, very high accuracy in both splits (91.67% to 92.5%) is achieved and Recall reaches 100% in the 80:20 split, which clearly specifies the correct identification of all glaucoma cases. However, Specificity is low (25–37%), which indicates the tendency of misclassification in the model. The 80:20 split generally

yields better accuracy and recall on all 6 datasets, since more training data allows the ViT model to learn richer representations.

It can be observed from Table III results that specificity remains comparatively low (11–25%) in two datasets (i.e., G1020 and REFUGE), which clearly indicates the tendency towards overprediction of glaucoma. This behavior reflects a sensitivity-oriented learning bias, which prioritizes the detection of glaucoma cases to minimize false negatives. This kind of trade-off is usually observed in glaucoma screening systems, in which missing a positive case poses greater clinical risk than false alarms. However, as seen in Table III, other datasets scored better specificity, which is due to the presence of higher image quality and more consistent labelling. This findings highlight the importance of data heterogeneity and labelling variability on specificity performance. In addition, findings emphasize the need for future work on threshold calibration and class-balanced optimization, in order to resolve over-diagnosis related issues.

### *B. Glaucoma Prediction Performance for the Entire Input Database*

In general, ViT architecture performs better on larger size datasets, since their ability to capture comprehensive features enhances, if the dataset size increases [11]. Based on these aspects, we evaluated the proposed ViT-based glaucoma diagnosis model on the complete input database, which combines the data from all the 6 benchmark datasets, which are listed in Table I.

Specifically, the prediction results of the proposed framework on the combined dataset is indicated in Table IV, by means of Accuracy, Precision, Recall and F1-Score, Specificity and Sensitivity metrics for different training Vs testing dataset ratios respectively. For instance, for 80:20 dataset ratio, the prediction results of the model are: accuracy is 76.1%, recall is 70 %, precision is 79.6%, F1-Score is 74.5%, specificity is 82.1% and sensitivity is 70% respectively, which indicates the better performance, when compared to other dataset ratios. Table IV results demonstrate that the proposed glaucoma assessment framework achieves reasonably better performance, even though the database size is high (i.e. combination of 6 different datasets). The reason for this better performance, is the incorporation of contrastive learning algorithm in the model along with hyperparameter fine-tuning, which helped ViT to learn useful feature representations.

In the context of real-world glaucoma screening, the balance between sensitivity and specificity is critical for ensuring clinical safety. In general, Glaucoma screening systems are designed to favor higher sensitivity to reduce the risk of missed glaucoma cases, as delayed detection may lead to irreversible vision loss. Table IV results indicate that, the proposed ViT-based framework demonstrates a balanced operating point on the combined dataset, by means of achieving 70% sensitivity and 82.1% specificity at 80:20 train-test split ratio. This trade-off suggests suitability of the proposed model for population-level screening or referral support, where false positives can be addressed through secondary ophthalmic evaluation. The integration of contrastive learning and hyperparameter fine-tuning enables robust feature learning across heterogeneous datasets, which contributes to stable performance even under increased dataset diversity.

TABLE IV. PERFORMANCE EVALUATIONS ON COMBINED INPUT DATABASE (IN %)

| Dataset ratio | Accuracy | Recall | Precision | F1-Score | Specificity | Sensitivity |
|---|---|---|---|---|---|---|
| 70:30 | 59.52 | 49.03 | 73.8 | 51.6 | 63.1 | 59.8 |
| **80:20** | **87.91** | **70.03** | **79.6** | **74.5** | **82.1** | **70.3** |
| 90:10 | 70.97 | 74.57 | 69.2 | 71.8 | 70.7 | 63.9 |

### *C. Performance Comparisons with Baseline Methods*

The detection performance of the proposed model are compared with the ViT-based existing Glaucoma detection techniques, which are indicated as ViT [8], Light ViT [13] and ResNet50 [24] respectively in Table V, which correspond to CNN-based and hybrid deep learning models including ResNet50 used for glaucoma assessment. Specifically, the performance comparisons of proposed Vs all three models for different training, testing dataset ratios are illustrated in Table V. Performance improvements were quantified using standard classification metrics including accuracy, precision, recall, and F1-Score, which are evaluated under multiple train-test splits. As evident in Table V experimental results, the proposed contrastive based ViT-architecture consistently outperforms all baselines, achieving a maximum accuracy of 87.91% (80:20 split), compared to 71.67–74.56% for existing methods. Table V prediction results clearly demonstrate that, our model is scoring an enhanced accuracy of 87.91% for 80:20 ratio, when compared to accuracy of reference methods, which are 71.67%, 74.56% and 86.1% respectively. Although ResNet50 models are slightly scoring better [24], they are evaluated with limited datasets and absence of XAI methods. The integration of contrastive learning-based ViT architecture with hyperparameter optimization is the reason for the improved performance results of proposed model. In particular, the proposed method achieves a maximum accuracy of 87.91% (80:20 split), whereas existing ViT-based and lightweight transformer baselines score in the range of 71.67–74.56% under identical experimental settings. These improvements indicate that the proposed contrastive pre-training model enables more discriminative feature representations and better generalization for glaucoma classification. The proposed framework demonstrates competitive or superior performance while additionally incorporating explainability through Layer wise Relevance Propagation (LRP), which provides clinically meaningful heatmaps highlighting optic disc and nerve head regions. Unlike several CNN-based approaches that are evaluated on limited datasets without interpretability analysis, our method combines multi-dataset evaluation, improved

accuracy, and explainable predictions, thereby offering a more robust and clinically reliable glaucoma screening solution.

The detailed Glaucoma assessment results of the proposed contrastive-learning based model are indicated in Table VI, in terms of Accuracy, Recall, Precision, F1-Score, Specificity and Sensitivity scores. It can be observed from Table VI results that, the proposed model is achieving better performance scores in terms of accuracy and other metrics. The reason behind the better prediction accuracy of the proposed model is, the utilization of two optimization strategies, which are illustrated as given by:

- GELU Activation: We used Gaussian Linear Error Unit (GELU) activation function, instead of using the traditional Rectified Linear Unit (ReLU), since GELU generates smoother gradients and prevents vanishing gradients.
- Batch Normalization and Dropout: We utilized batch normalization and dropout aspects in layers to prevent overfitting issues, so that activation variance is reduced and thereby the proposed model can learn robust features.

Further, explainable AI technique-LRP is employed in this study, to generate LRP explanations, which helps to interpret the predictions of the proposed model. The LRP outputs in the form of Heatmaps, demonstrate that the proposed model can be successfully employed in clinical applications for validation purposes.

TABLE V. ACCURACY COMPARISON OF PROPOSED VS BASELINE TECHNIQUES

| Training Testing | ViT [8] (in %) | Light ViT [13] (in %) | ResNet50 [24] (in %) | Proposed (in %) |
|---|---|---|---|---|
| 70:30 | 51.04 | 63.06 | 71.08 | 69.52 |
| **80:20** | 71.67 | 74.56 | 86.1 | **87.91** |
| 90:10 | 68.56 | 69.01 | 79.03 | 75.97 |
| **Explainable AI** | - | - | - | **LRP Explanations** |

TABLE VI. GLAUCOMA DETECTION RESULTS OF PROPOSED CONTRASTIVE-LEARNING BASED MODEL(IN %)

| Accuracy | Recall | Precision | F1-Score | Specificity | Sensitivity |
|---|---|---|---|---|---|
| 86.1 | 70.03 | 79.6 | 74.5 | 82.1 | 70.3 |

### *D. Enhancement of Training Efficiency*

The training efficiency related issues of baseline techniques are addressed in this study as follows: All experiments are conducted on a single NVIDIA A100 GPU. The training of the proposed contrastive pre-trained ViT-B/16 model on the combined six-dataset benchmark consumed approximately 4.5 h for 63 epochs (app. 4.2 min per epoch). Further, computational efficiency is improved by freezing the entire ViT-B/16 backbone (with 85.8M parameters) and fine-tuning only a lightweight classification head with 130,665 parameters. This strategy significantly reduced gradient computation, memory usage, and training time when compared to end-to-end ViT training.

### *E. Layer-Wise Relevance Propagation (LRP) Explanations*

In this study, we used a popular Explainable AI (XAI) technique, Layer-wise Relevance Propagation (LRP) [14], which facilitates to interpret the predictions of the proposed ViT-based model. LRP represents the details of input features, which primarily contributed to arrive at a decision. Specifically, LRP characterizes the proposed model's prediction, by moving backward through the layers to the input features and distributing the prediction score back to the input pixels and thereby emphasizes the "relevance" of each feature in decision making process [14]. LRP utilizes the conservation principle, which specifies that the total relevance is preserved as it's propagated from output to input, as given by,

$$\sum_j R_j^{(l+1)} = \sum_i R_i^{(l)} \tag{3}$$

where $R$ indicates relevance and $l$ indicates the layer index.

Fig. 7 indicate the LRP interpretations of the proposed model for the respective Glaucoma as well as Non-Glaucoma fundus retinal images. Specifically, the left Panel illustrates the original retinal fundus image, which shows the the optic disc (bright circular area at the centreleft), where the optic nerve exits the eye and the Retinal blood vessels are branching from the disc. This is a sample input image considered by the proposed model for glaucoma diagnosis. The right panel indicates the LRP interpretations for the Glaucoma positive case, in which the LRP-generated heatmap is overlaid on the original image. The LRP outputs clearly emphasize the regions which mainly contributed to the model's decision. The Red/Yellow areas indicate the High relevance, which strongly influenced the model's decision in glaucoma prediction, whereas blue areas show the Low or negative relevance, which contributed less in the decision-making process.

It is evident in Fig. 7 that, the optic disc and its nearby areas are highlighted in yellow-red regions. This emphasizes that, the proposed model is correctly focusing on the optic nerve head, which is clinically relevant, since the glaucoma affects the optic nerve and cup-to-disc ratio, as per the literature studies. As per the LRP outputs, Sum(R) = 8555.1582, which indicates the total relevance distributed across all pixels. It is evident from Fig. 7 outputs that, the LRP explanation is valid, since the model is mainly relying on the optic disc region, which is appropriate for the glaucoma disease. The LRP outputs also prove that, the proposed model is useful for Clinical validations, as the visualization helps to validate that the model is not relying on irrelevant areas such as corners or edges. In this way, the LRP-based explanations

demonstrate that the proposed model mainly focuses on the optic disc and surrounding optic nerve head regions, which are clinically meaningful for glaucoma assessment. It also demonstrates that, the model's predictions are driven by anatomically relevant structures instead of spurious background features, and thereby enhances trustworthiness and supports its potential for clinical integration [32].

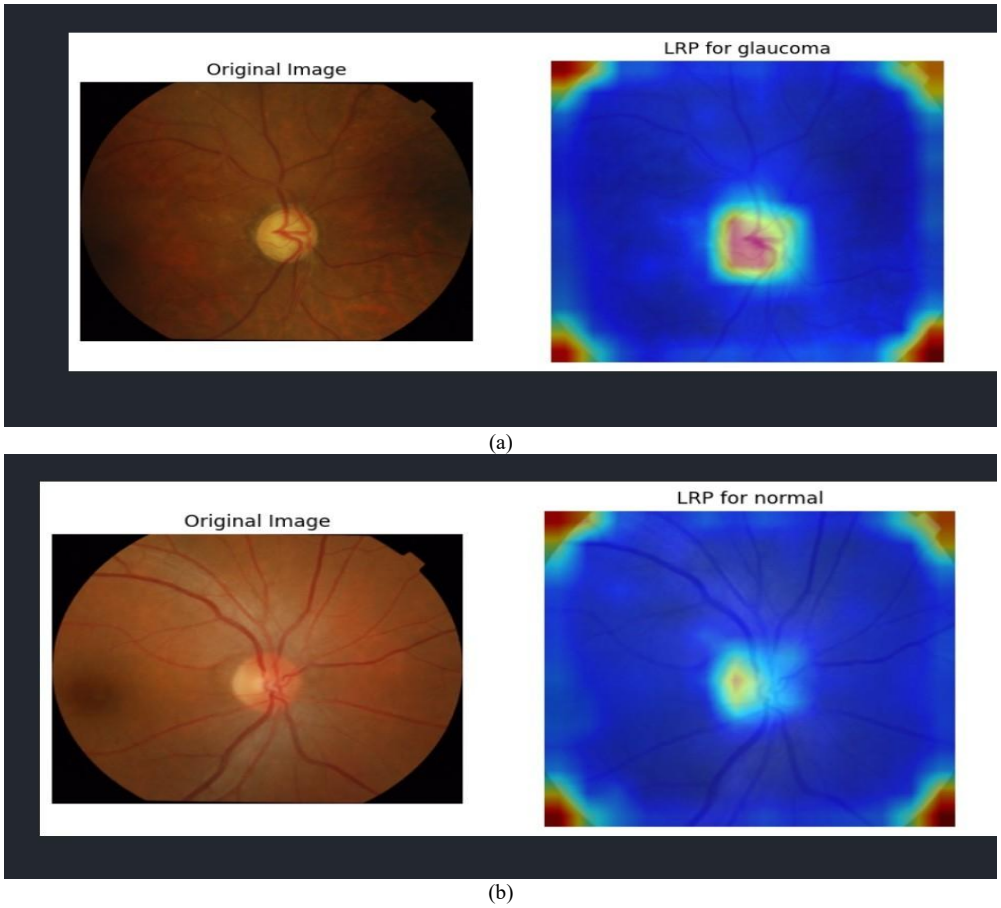



Fig. 7. LRP explanations for representative fundus images- highlighting high-relevance regions around the optic disc and peri-papillary areas, which indicates that the model focuses on clinically meaningful structures associated with glaucoma. (a) Sample glaucoma image; (b) Sample non-glaucoma image.

## V. Conclusion

Early detection of glaucoma is critical to prevent irreversible vision loss. This study proposes a contrastive learning-optimized ViT-B16 framework with hyperparameter tuning for automated glaucoma diagnosis. The proposed model consistently outperforms CNN and baseline ViT methods, achieving up to 87.91% accuracy across multiple evaluation metrics.

This framework is validated on the combined benchmark comprising six public datasets, collected from different devices and acquisition settings for ensuring robustness and generalizability. This multi-dataset evaluation demonstrates strong cross-domain performance of the proposed model. Further, Layer-wise Relevance Propagation (LRP) provides clinically interpretable heatmaps, which enables transparent and trustworthy predictions. While formal regulatory validation and prospective clinical testing are beyond the scope of this study, these aspects are considered as important directions for future work.

## Conflicts of Interest

The authors declare no conflict of interest.

## Author Contributions

R. Roopalakshmi: Concept finalization, dataset collection, Result analysis, original manuscript draft, review and correspondence of the manuscript. Sambhav Nath Jain and Ayush Amarnath Bhakat: Concept finalization, dataset collection and Result analysis. All authors have read and approved the final version of the manuscript.

## References

[1] F. Chincholi and H. Koestler, "Transforming glaucoma diagnosis: Transformers at the forefront," *Front. Artif. Intell.*, vol. 7, 1324109, 2024. https://doi.org/10.3389/frai.2024.1324109

[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, 2014.

[3] P. Rewri, "Towards better management of glaucoma in India," *Indian J. Ophthalmol.*, vol. 71, no. 3, pp. 686–688, Mar. 2023. https://doi.org/10.4103/IJO.IJO37923
[4] Directorate General of Health Services. (2023). National Blindness and Visual Impairment Survey India 2015-2019-A Summary Report. *National Programme for Control of Blindness & Visual Impairment*. [Online]. Available: https://npcbvi.gov.in/writeread data/mainlinkfile/file341.pdf
[5] GLAUCOMA Research Foundation. Types of glaucoma. [Online]. Available: https://glaucoma.org/types
[6] M. Wassel, A. M. Hamdi, N. Adly, and M. Torki, "Vision transformers based classification for glaucomatous eye condition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Montréal, QC, Canada, Aug. 2022, pp. 21–25.
[7] G. An *et al.*, "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," *J. Healthc. Eng.*, vol. 2019, 4061313, 2019. https://doi.org/10.1155/2019/4061313
[8] P. B. Singh *et al.*, "Glaucoma classification using light vision transformer," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 9, 2023.
[9] M. Kalra, and A. Zahoor, "Advancements in deep learning for glaucoma detection from fundus images: A comprehensive analysis," *J. Transform. Technol. Sustain. Dev.*, vol. 9, no. 9, pp. 1–20, 2025. https://doi.org/10.1007/s41314-025-00073-6
[10] R. Fan *et al.*, "Detecting glaucoma from fundus photographs using deep learning without convolutions transformer for improved generalization," *Amer. Acad. Ophthalmol.*, 2022. https://doi.org/10.1016/j.xops.2022.100233
[11] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv Preprint, arXiv:2010.11929, 2020.
[12] S. Mallick, J. Paul, N. Sengupta, and J. Sil, "Study of different transformer based networks for glaucoma detection," in *Proc. TENCON 2022–2022 IEEE Region 10 Conf. (TENCON)*, 2022. https://doi.org/10.1109/TENCON55691.2022.9977730
[13] I.-E. Haouli, W. Hassina, and S.-Bouchelague, "Exploring vision transformers for automated glaucoma disease diagnosis in fundus images," in *Proc. 2023 Int. Conf. Decision Aid Sci. Appl. (DASA)*, 2023. https://doi.org/10.1109/DASA59624.2023.10286714
[14] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, 2015. https://doi.org/10.1371/journal.pone.0130140
[15] R. Praveena and T. Ganeshbabu, "Determination of cup to disc ratio using unsupervised machine learning techniques for glaucoma detection," *Molec. Cell. Biomech.*, vol. 18, 69, 2021. https://doi.org/10.32604/mcb.2021.014622
[16] R. Fan, C. Bowd, M. Christopher *et al.*, "Detecting glaucoma in the ocular hypertension treatment study using deep learning: Implications for clinical trial endpoints," *Authorea Preprints*, 2022.
[17] D. R. Nayak, D. Das, B. Majhi, S. V. Bhandary, and U. R. Acharya, "Ecnet: An evolutionary convolutional network for automated glaucoma detection using fundus images," *Biomed. Signal Process. Control*, vol. 67, 102559, 2021.
[18] M. Christopher *et al.*, "Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018.
[19] S. Phan, S. Satoh, Y. Yoda, K. Kashiwagi, and T. Oshika, "Evaluation of deep convolutional neural networks for glaucoma detection," *Jpn. J. Ophthalmol.*, vol. 63, no. 3, pp. 276–283, 2019.
[20] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, "Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs," *Ophthalmology*, vol. 125, no. 8, pp. 1199–1206, 2018.
[21] S. Yu *et al.*, "MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Springer, 2021, pp. 45–54.
[22] M. Juneja, N. Thakur, S. Thakur, A. Uniyal, A. Wani, and P. Jindal, "GC-NET for classification of glaucoma in the retinal fundus image," *Mach. Vis. Appl.*, vol. 31, pp. 1–18, 2020. https://doi.org/10.1007/s00138-020-01091-4
[23] X. Chen *et al.*, "Detecting glaucoma in highly myopic eyes from fundus photographs using deep convolutional neural networks," *Clin. Experiment. Ophthalmol.*, 2025. https://doi.org/10.1111/ceo.14498
[24] A. Shoukat, S. Akbar, S. A. Hassan, S. Iqbal, A. Mehmood, and Q. M. Ilyas, "Automatic diagnosis of glaucoma from retinal images using deep learning approach," *Diagnostics*, vol. 13, no. 10, 1738, 2023. https://doi.org/10.3390/diagnostics13101738
[25] P. H. Le Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.
[26] GitHub Repository. [Online]. Available: https://github.com/mirtanvirislam/Deep-Learning-Based-Glaucoma-Detection-with-Cropped-Optic-Cup-and-Disc-and-Blood-Vessel-Segmentation/tree/master/Dataset
[27] K. Jin, X. Huang, J. Zhou *et al.*, "FIVES: A fundus image dataset for artificial intelligence based vessel segmentation," *Sci. Data*, vol. 9, 475, 2022. https://doi.org/10.1038/s41597-022-01564-3
[28] M. N. Bajwa *et al.*, "G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection," in *Proc. 2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–7.
[29] O. Kovalyk, J. Morales-Sánchez, R. Verdú-Monedero, I. Sellés-Navarro, A. Palazón-Cabanes, and J.-L. Sancho-Gómez. (2022). PAPILA figshare dataset. [Online]. Available: https://doi.org/10.6084/m9.figshare.14798004.v1
[30] Z. Zhang *et al.*, "ORIGA-light: An online retinal fundus image database for glaucoma analysis and research," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2010, pp. 3065–3068. https://doi.org/10.1109/IEMBS.2010.5626137
[31] H. Fu *et al.*, "REFUGE: Retinal fundus glaucoma challenge," *IEEE Dataport*, 2019. https://ieee-dataport.org/documents/refuge-retinal-fundus-glaucoma-challenge
[32] H. Wang, S. Toumaj, A. Heidari, A. Souri, N. Jafari, and Y. Jiang, "Neurodegenerative disorders: A holistic study of the explainable artificial intelligence applications," *Eng. Appl. Artif. Intell.*, vol. 153, 2025. https://doi.org/10.1016/j.engappai.2025.110752