JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 03, Issue No. 01, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0301-0027

Automated Retinopathy of Prematurity Screening via Interpretable and Efficient Transfer Learning Architectures

Muhammad Haseeb ur Rehman[1,*], Saqib Dar[2]

[1] Independent Researcher, Lahore, Pakistan
[2] Independent Researcher, Riyadh, Saudi Arabia

*Corresponding author: mehaseeburrehman@gmail.com
Contributing authors: mehaseeburrehman@gmail.com, saqi2926@gmail.com

Received: April 26, 2026 **Accepted:** June 20, 2026

Abstract- Retinopathy of prematurity (ROP) can progress to severe visual loss in preterm infants when treatment-requiring disease is missed. We assessed six transfer-learning CNN configurations for four-class classification of retinal fundus images: VGG19, DenseNet121, MobileNetV2, AlexNet, EfficientNetV2-S, and CBAM-ResNet50. In the last configuration, the Convolutional Block Attention Module (CBAM) is integrated with a ResNet50 backbone. The dataset comprised 2,580 images from four diagnostic categories: Healthy, Type 1 ROP, Type 2 ROP, and retinal detachment (RD), using both publicly available and clinical images. Images were preprocessed and augmented during training, and model performance was assessed with 5-fold cross-validation. CBAM-ResNet50 produced the highest reported accuracy and F1-score, reaching 95.4% and 95.9%, respectively. Its fold-specific macro-AUC ranged from 0.700 in Fold 1 to 0.975 in Fold 5, Fold 4 reached 0.942, and the mean across the five folds was 0.854. The variation between folds suggests that performance depends on the composition of the held-out data.

Keywords- Retinopathy; Machine learning; Early diagnosis; Transfer learning; CNN

1. Introduction

ROP develops in the incompletely vascularized retina of preterm infants and is associated particularly with prematurity and low birth weight. Abnormal vascular growth can follow disruption of normal retinal maturation, while changes in oxygen exposure after birth may further affect this process [1]. Advances in neonatal care have improved survival among extremely preterm infants, increasing the number of infants who require repeated ROP examinations. Providing these examinations is difficult in settings with limited access to trained ophthalmologists and organized screening services. Untreated ROP may progress to retinal detachment (RD) and permanent visual impairment. Screening therefore depends on identifying treatment-requiring disease before advanced retinal damage occurs. Clinical assessment is based on retinal examination and digital fundus imaging. When determining the severity of a disease and the state of therapy, ophthalmologists take into account characteristics such vascular dilation, tortuosity, demarcation lines, and ridges. Practical constraints can have an impact on the screening procedure. Image quality can be lowered by infant movement and media opacity, and grading judgments can be impacted by observer differences [3]. Another limitation is access to expert evaluation, especially when a significant number of preterm babies need to have repeated exams. Clinical environments also differ in practice patterns, such as how ROP is treated and monitored [10]. Automated retinal image analysis has been investigated as a way to assist image interpretation and screening. CNN-based methods can learn discriminative retinal features directly from labeled images, and related work spans diabetic-retinopathy analysis as well as ROP-specific applications [2,5]. The ROP literature also includes studies of associated ophthalmic complications and the training required for reliable clinical

assessment [4, 6]. These strands of work reflect different clinical needs, from image-based classification to specialist interpretation and follow-up. In the present study, automated classification is considered as decision support rather than a replacement for ophthalmic assessment. Transfer learning provides a practical way to adapt deep neural networks when task-specific medical image datasets are smaller than the datasets typically used to train large models from random initialization. A network pretrained on a large image collection such as ImageNet can be adapted to a target task by replacing or modifying its classification layers and fine-tuning part of the network. Transfer learning has been applied to medical classification problems outside ROP, including MRI-based disease classification and diabetes detection [7, 8]. Deep-learning methods have also been studied specifically for ROP screening [9]. In the present work, pretrained networks are fine-tuned on the same ROP dataset. The models are then compared using the same data partitions, preprocessing steps, training procedure, and evaluation measures. Previous reviews of AI methods for ROP report differences in the prediction tasks, datasets, network architectures, and validation procedures used across papers [11, 12]. Both image classification and retinal blood-vessel analysis have been used in ROP research [14]. Papers also vary in the quantity and description of output categories. While the current task distinguishes between Healthy, Type 1 ROP, Type 2 ROP, and RD, other models employ fewer diagnostic classifications. It is challenging to directly compare results from different research due to differences in image sources and dataset makeup. Different clinical concerns are addressed by work on other retinal disorders, which should be separated from ROP-specific categorization research [13]. Because of this, the preprocessing, augmentation, data partitioning, and evaluation criteria used in our trials are the same for all six models. Four groups are included in the categorization task: RD, Type 1 ROP, Type 2 ROP, and Healthy. Using the same experimental setup, we evaluate AlexNet, MobileNetV2, VGG19, DenseNet121, EfficientNetV2-S, and CBAM-ResNet50. The CBAM-ResNet50 configuration incorporates the Convolutional Block Attention Module (CBAM), which applies channel and spatial attention to the feature maps produced by the underlying ResNet50 architecture. Image preprocessing and augmentation are applied before training to account for variation in illumination, retinal appearance, and image acquisition conditions.

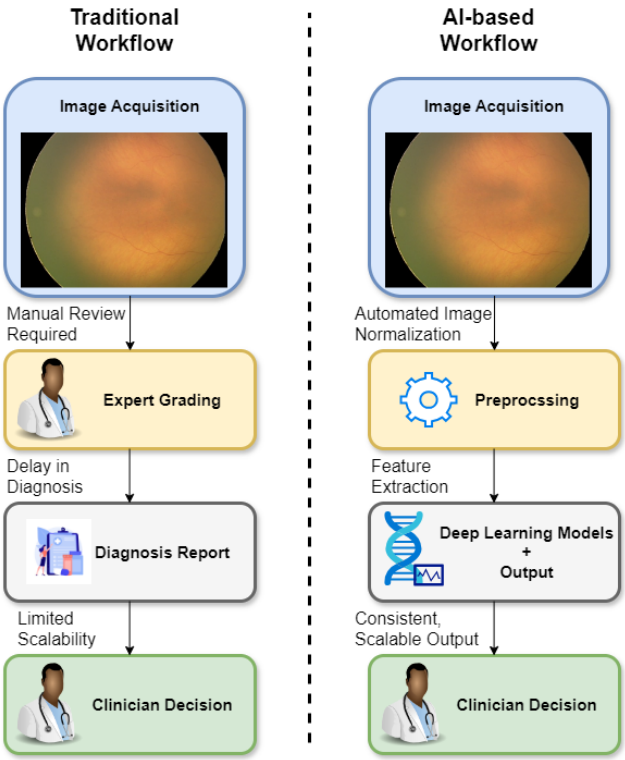


Figure 1: *Traditional and AI-based ROP screening workflows.*

Figure 1 contrasts the main stages of conventional ROP screening with the image-processing process considered in this paper. Conventional screening depends on specialist examination and manual interpretation of retinal findings. In the proposed workflow, preprocessing, feature extraction, and classification are performed computationally, after which the predicted class can be used as decision-support information. The final clinical interpretation remains with the ophthalmologist.

The main contributions of this paper are as follows:

- We formulate a transfer-learning framework for four-class ROP classification using pretrained CNN architectures and retinal fundus images.
- We evaluate a CBAM-ResNet50 configuration in which channel and spatial attention are incorporated into ResNet50 to emphasize retinal features used during classification.

- We apply a preprocessing and data-augmentation process designed for neonatal retinal images, including variation associated with illumination, motion artifacts, and retinal appearance.
- We compare six pretrained architectures, namely AlexNet, MobileNetV2, VGG19, DenseNet121, EfficientNetV2-S, and CBAM-ResNet50, using the same four-class dataset and 5-fold cross-validation. Each model was tested on five held-out partitions rather than on one fixed train-test split. The same partitioning and training procedure was used for all six architectures.
- Classification performance is reported using accuracy, sensitivity, specificity, precision, F1-score, AUC, confusion matrices, and ROC curves. The class-wise measures and confusion matrices show where errors occur, while AUC and ROC curves describe discrimination across decision thresholds.

The rest of the paper is arranged as follows. Section 2 discusses previous work on ROP detection, classification, and related prediction tasks. Section 3 gives the dataset, preprocessing and augmentation steps, model architectures, training procedure, and evaluation protocol. The experimental results and their analysis are presented in Section 4. Section 5 closes the paper with the main findings, limitations, and directions for further work.

2. Related Work

Artificial intelligence has been applied to retinal images for a range of ophthalmic diseases, including ROP and other ocular conditions [15, 17]. Within the ROP literature, separate studies have examined visual impairment burden, clinical risk factors, and screening practice [16, 18, 19]. Treatment practice is another strand of this work. Wang and Dai [20] surveyed international treatment preferences and reported differences in the use of laser photocoagulation and intravitreal anti-VEGF therapy across ROP presentations. Sun et al. [21] approached the disease from a genetic perspective and identified factors associated with severe ROP, including variants linked to oxidative stress, inflammation, and angiogenesis. Deep-learning studies have addressed different stages of ROP assessment. ROPRNet predicts recurrence from longitudinal retinal images after treatment [22]. Yang et al. [23] developed an interpretable system for severity screening and used saliency information to relate predictions to regions of the retinal image. Ebrahimi et al. [24] studied the spectral content of color fundus photographs and reported differences in classification performance when different spectral information was used. Luo et al. [25] proposed VGG-ST, which combines a VGG-based convolutional component with a Swin Transformer for ROP diagnosis. A different combination was explored by Sankari et al. [26], who evaluated deep learning together with quantum machine learning for automated ROP detection. Some papers focus on narrower components of the screening process. Peng et al. [27] developed a key-area localization method for automatic retinal zoning, which is relevant to determining ROP zone information. Salih et al. [28] compared deep learning models and a fusion-based classification method for ROP diagnosis in premature infants. Yurdakul et al. [29] proposed ROPGCViT, an explainable vision-transformer model for ROP diagnosis. A broader assessment of the field was provided by Chu et al. [30] through a systematic review and meta-analysis of image-based machine-learning methods for ROP diagnosis. Their review compared reported diagnostic performance across studies that used different datasets and evaluation procedures. Image data are not the only source used for prediction. Durmaz Engin et al. [31] used demographic and clinical variables to predict both ROP development and the need for treatment. Other papers have addressed genetic risk, disease stage, and clinical risk prediction. Young et al. [32] investigated prediction of genetic risk from retinal images. Govind et al. [33] used deep learning to classify ROP stages, whereas Chen et al. [34] developed a model based on multiple clinical risk factors. Deep-learning-based prediction of ROP in premature infants was further examined in [35]. Automated screening may therefore have practical value in environments where specialist ophthalmic assessment is difficult to provide consistently. Ortiz et al. [36] investigated AI-assisted ROP screening in low-resource settings, where access to trained ophthalmologists may restrict conventional screening programs. At the other end of the model-complexity spectrum, Lakshmanan et al. [37] examined conventional machine-learning algorithms for ROP detection. Such methods provide a useful comparison with computationally heavier deep neural networks. AI-based quantification has also been investigated for plus disease, where assessment of retinal vascular abnormalities can vary among observers. Coyner et al. [39] evaluated an AI-generated Vascular Severity Score (VSS) derived using the i-ROP DL system. With VSS assistance, linearly weighted agreement increased from 0.69 to 0.81. The reported area under the ROC curve increased from 0.94 to 0.98, while the area under the precision-recall curve increased from 0.86 to 0.95. These results provide evidence that a continuous AI-derived vascular score can support specialist assessment of plus disease under the conditions evaluated in that paper. Deepthi et al. [40] proposed a Transformer-based Unsupervised Curriculum Learning Model (TUCLM) for automated plus-disease classification. The framework uses a PRes-SE-att-ViT backbone for feature extraction and unsupervised clustering to assign pseudo-labels. Curriculum learning then gives priority to samples according to their proximity to cluster centroids during iterative clustering and fine-tuning. The paper reports an accuracy of 99.3%, precision of 98.8%, recall of 99.1%, and an F1-score of 99%. These results were obtained within the experimental setting of that paper and indicate that reduced dependence on manually assigned labels is feasible for this particular classification task. Table 1 summarizes the principal characteristics of the studies most closely related to the present work. The literature covers recurrence prediction, retinal zoning, plus-disease assessment, severity classification, clinical risk prediction, and low-resource screening. However, these studies differ in their input data, target classes, architectures, and evaluation procedures.

Table 1: *Comparative summary of key studies in automated ROP detection using deep learning.*

Ref.	Model / Framework	Architecture Type	Input Modality	Task	Explainability	Highlights
[22]	ROPRNet	CNN-based DNN	Longitudinal retinal images	Recurrence prediction	No	Models recurrence monitoring
[23]	Custom DNN	CNN + Saliency	Fundus images	Severity classification	Yes (Saliency)	Provides visual interpretability
[24]	Spectral DNN	CNN	Spectral fundus images	Classification	No	Examines the effect of spectral information
[25]	VGG-ST	CNN + ViT	Fundus images	Classification	Implicit	Combines local and global image features
[26]	Quantum Hybrid	CNN + Quantum ML	Fundus images	Detection	No	Combines deep learning and quantum machine learning
[27]	Retina Zoner	CNN	Fundus images	Region localization	No	Supports retinal zoning for ROP assessment
[28]	Ensemble DNN	CNN Ensemble	Fundus images	Classification	No	Uses model fusion for classification
[29]	ROPGCViT	Vision Transformer	Fundus images	Severity classification	Yes (Attention)	Uses attention information for interpretation
[30]	Meta-paper	Traditional ML	Various sources	Review paper	No	Synthesizes reported diagnostic performance
[31]	Clinical Estimator	Traditional ML	Clinical + demographics	Treatment estimation	No	Uses clinical and demographic information
[32]	Genotype Predictor	CNN	Retinal images	Genetic risk stratification	No	Investigates retinal phenotype and genetic risk
[33]	StageNet	CNN	Fundus images	Stage classification	No	Classifies ROP stages from retinal images
[34]	Collab Predictor	Hybrid ML	Blood, oxygen, and perinatal data	Severity prediction	No	Combines multiple clinical risk factors
[35]	Attentive DNN	CNN + Attention	Multimodal inputs	Progression prediction	Yes (Attention)	Uses multimodal information and attention
[36]	AI Screener	CNN	Fundus images	Remote screening	No	Evaluates screening in low-resource settings
[37]	Classical ML	Handcrafted + SVM	Retinal images	Detection	No	Uses conventional machine-learning features
[39]	i-ROP DL + VSS	CNN + Scoring	Retinal images	Plus disease detection	Yes (Score Visuals)	Uses a vascular severity score to support diagnosis
[40]	TUCLM	ViT + Curriculum	Fundus images	Plus disease classification	Yes (ViT Attention)	Combines unsupervised curriculum learning with transformer classification

The present paper therefore evaluates six pretrained architectures under a common preprocessing, cross-validation, and four-class classification protocol.

3. Proposed Methodology

The paper uses a transfer-learning workflow for four-class classification of ROP from retinal fundus images. The workflow covers dataset preparation, expert labeling, image preprocessing, training-time augmentation, model adaptation, and evaluation. Six model configurations are examined under the same data-partitioning procedure. Four established CNN architectures are used as baselines, while EfficientNetV2-S and a CBAM-ResNet50 configuration provide two additional architectural settings. Figure 3 summarizes the overall workflow.

3.1. Dataset Description and Preprocessing

Retinal fundus images were obtained from two sources. The first was a publicly available dataset containing approximately 6,500 images from 188 premature infants [38]. Three classifications were applied to the images from this source utilized in this paper: Type 1 ROP (627 images), Type 2 ROP (650 images), and Healthy (653 images). For the RD category, another dataset was utilized. It included 650 clinical fundus images taken at Pakistan's General Hospital in Lahore.

Prior to the images being utilized in the paper, any patient-identifying information was eliminated. The images were examined for diagnostic usefulness prior to model training. Images with significant motion blur, excessive exposure, or retinal areas without information were not included. Four diagnostic groups were created from the preserved images: RD, Type 1 ROP, Type 2 ROP, and Healthy. Before being sent to the CNN models, each picture was scaled to $224 \times 224 \times 3$ pixels. Figure 2 shows images from the four categories. Table 2 reports the class counts and the training and test allocation stated for the 5-fold cross-validation procedure.

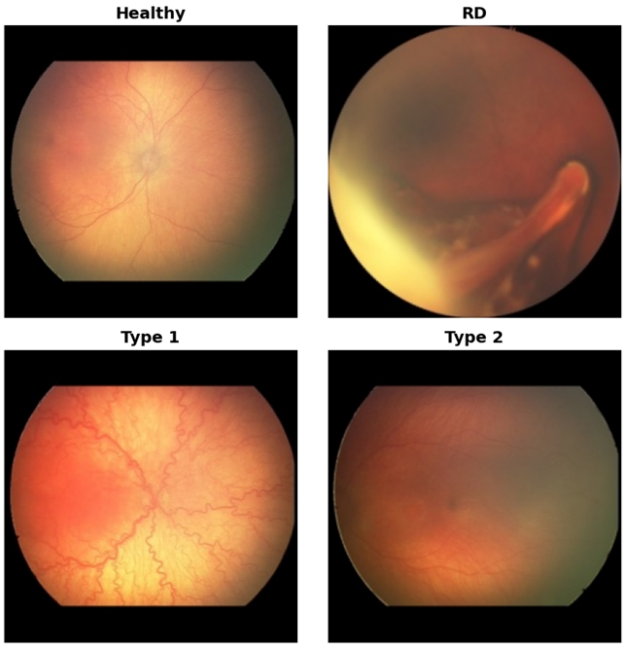


Figure 2: *Sample retinal images from the four diagnostic categories in the ROP dataset: Healthy, RD, Type 1 ROP, and Type 2 ROP.*

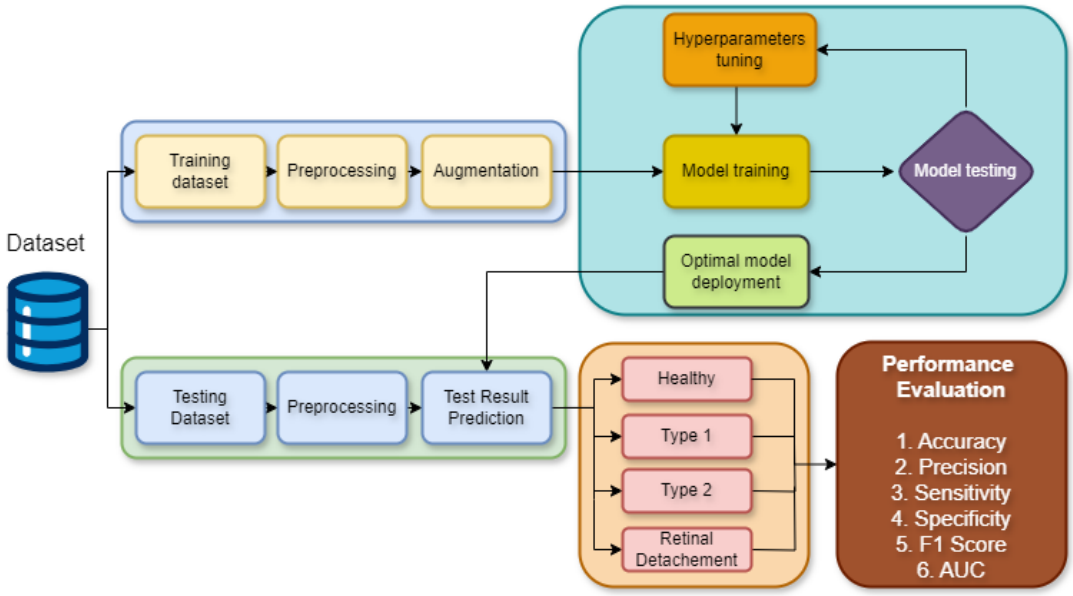


Figure 3: *Proposed ROP classification framework. The framework includes preprocessing, augmentation, model training, and evaluation using multiple CNN architectures.*

Table 2: *Distribution of ROP images across the training and test folds.*

Category	Total Images	Training Set (4 Folds)	Test Set (1 Fold)
Healthy	653	522	131
Type 1	627	501	126
Type 2	650	520	130
Retinal Detachment (RD)	650	520	130
Total	**2580**	**2064**	**516**

As shown in Figure 3, the images first undergo preprocessing and training-time augmentation. The processed images are then used to fine-tune the selected CNN models. Model evaluation is performed using the cross-validation procedure described later in this section.

3.2. Image Labeling

Two experienced ophthalmologists evaluated the retinal images. Each specialist initially assessed the images independently according to the International Classification of ROP and assigned them to the diagnostic categories used in this paper. Healthy images were treated as the normal category, while Type 1 ROP, Type 2 ROP, and RD represented the three abnormal categories. The independently assigned labels were subsequently compared to identify disagreements between the two assessments. These expert assessments provided the reference labels used during model development and evaluation.

3.3. Image Normalisation

All retinal images were processed using the same input process before model training. Each image was resized to $224 \times 224 \times 3$ pixels and converted using the PyTorch `transforms` module. Pixel intensities represented as 8-bit values from 0 to 255 were converted to floating-point tensors and scaled to the interval $[0, 1]$. Channel-wise normalisation was then applied to the red, green, and blue channels using a mean of 0.5 and a standard deviation of 0.5 for each channel. This operation maps the scaled pixel values to the interval $[-1, 1]$. The same normalisation procedure was used for the training and evaluation images.

3.4. Data Augmentation

Instead of making a separate enhanced copy of the dataset, image augmentation was performed dynamically during training. Images allocated to the test fold remained unaltered and only those in the training folds were enhanced. Color jittering, rotation, rescaling, and random horizontal flipping with a chance of 0.5 were all part of the augmentation process. The stated values for brightness, contrast, saturation, and hue were all set to 0.5. During model training, these modifications introduced variance in the orientation, size, color, and lighting of the images. In every cross-validation step, augmentation was implemented independently. Consequently, augmentation was restricted to the training portion of the corresponding fold, while the held-out test images were evaluated without augmentation.

3.5. Classification Model Architectures

Six model configurations were evaluated for four-class ROP classification. They were selected to represent different CNN design strategies, including sequential convolutional networks, dense connectivity, lightweight feature extraction, network scaling, residual learning, and attention-based feature weighting. AlexNet, MobileNetV2, VGG19, and DenseNet121 were used as baseline architectures. EfficientNetV2-S and CBAM-ResNet50 were evaluated as additional configurations.

3.5.1. Baseline Architectures

The four baseline architectures were AlexNet, MobileNetV2, VGG19, and DenseNet121. ImageNet-pretrained weights were used to initialize the corresponding networks before fine-tuning on the ROP dataset. Their original classification layers were replaced with output layers for the four target classes: Healthy, Type 1 ROP, Type 2 ROP, and RD. The four architectures differ in how image features are propagated through the network. VGG19 has a deeper stack of convolutional processes, whereas AlexNet offers an earlier sequential CNN approach. DenseNet121 reuses intermediate characteristics by utilizing dense connections between network blocks. Inverted residual structures are used by MobileNetV2 in order to lower computing demands. A consistent experimental foundation for comparison is provided by evaluating these models using the same dataset and training process.

3.5.2 Advanced Architectures

Two additional configurations were included in the comparison:

- EfficientNetV2-S: EfficientNetV2-S uses a scaling strategy that jointly considers network depth, width, and image resolution. It was included as an alternative CNN design to the four baseline architectures.
- CBAM-ResNet50: This configuration combines a ResNet50 backbone with the Convolutional Block Attention Module (CBAM). CBAM applies channel attention followed by spatial attention to intermediate feature maps. The configuration was evaluated alongside the other architectures for the same four-class ROP classification task.

The CNN backbones were initialized with ImageNet-pretrained weights and adapted to the ROP dataset through fine-tuning. Earlier convolutional layers were kept frozen during the reported training procedure, whereas deeper layers were made trainable. The final classification layer was adapted to produce predictions for the four ROP categories. Figure 4

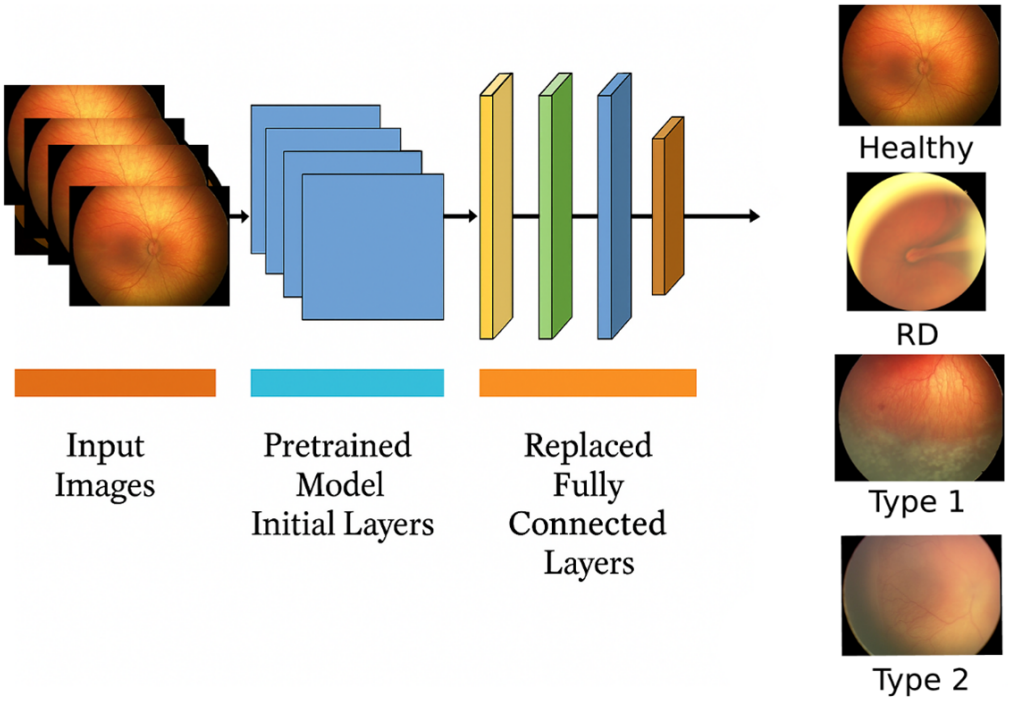


Figure 4: *Fine-tuning workflow for the CNN architectures used for ROP classification. The pretrained network is adapted through layer freezing, output-layer replacement, and retraining on the ROP dataset.*

summarizes the fine-tuning procedure, including pretrained initialization, layer freezing, replacement of the output layer, and training on the ROP images. Table 3 reports the convolutional-layer counts stated for the six configurations.

Table 3: *Pretrained CNN architectures used and their reported convolutional depth.*

Model	Number of Convolutional Layers
AlexNet	8
MobileNetV2	40
VGG19	18
DenseNet121	57
EfficientNetV2-S	23
CBAM-ResNet50	28

The selected models cover several CNN design families rather than variants of a single architecture. AlexNet and VGG19 represent sequential convolutional designs; DenseNet121 uses dense feature connectivity; MobileNetV2 uses inverted residual blocks; EfficientNetV2-S applies a scaling-based design; and CBAM-ResNet50 combines residual learning with channel and spatial attention. All six configurations are evaluated using the same ROP classification task and data-partitioning procedure.

3.6. Model Evaluation

To thoroughly evaluate the efficacy of CNN models fine-tuned via transfer learning in the multi-class classification of ROP, we implemented a range of established performance metrics and validation methods. Considering the consequences of missed or incorrect ROP classifications, model evaluation was conducted using both overall and class-specific performance measures. Type 1 ROP and RD received special attention as mistakes in these categories might impact later clinical evaluation. Instead of employing a single train-test split, each model's performance was evaluated across many held-out data divisions using 5-fold cross-validation. Throughout the cross-validation process, performance metrics were calculated for each fold and then compiled. Accuracy, sensitivity, specificity, precision, F1-score, and AUC are among the metrics that have been published. In order to analyze the distribution of predictions across the four categories Healthy, Type 1 ROP, Type 2 ROP, and RD—confusion matrices were also created for each model. Each factor was taken into consideration in a one-versus-rest way for class-specific evaluation in the four-class context. True negatives (TN) are all remaining images classified outside of a class, false negatives (FN) are images of that class assigned to another category, false positives (FP) are images from other categories incorrectly assigned to that class and true positives (TP) are images of that class

predicted correctly. This formulation allows sensitivity, specificity, and precision to be calculated separately for each diagnostic category. Confusion between Type 1 and Type 2 ROP was also examined because the two categories can show overlapping retinal characteristics. The evaluation metrics used in this study are defined below.

- **Accuracy** measures the proportion of all test images that are assigned to the correct class. For the multi-class classification task, it is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\%. \tag{1}$$

- **Sensitivity (Recall)** measures the proportion of samples belonging to a particular class that are correctly identified. For each class under the one-vs-rest formulation, it is calculated as:

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \times 100\%. \tag{2}$$

  A lower sensitivity indicates that a larger proportion of samples from the target class are being assigned to other categories.
- **Precision** measures the proportion of predictions assigned to a particular class that are correct. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\%. \tag{3}$$

  Precision therefore reflects the extent to which positive predictions for a class correspond to images that actually belong to that class.
- **Specificity** measures the proportion of samples outside the target class that are correctly identified as negatives. For each class, it is calculated as:

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\%. \tag{4}$$

  Higher specificity corresponds to fewer images from other categories being incorrectly assigned to the target class.
- F1-Score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between false positives and false negatives:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{5}$$

- In addition to macro-averaged metrics that treat all classes equally, we also compute weighted averages to account for class imbalance. For any metric M_i, the weighted average is:

$$\text{Weighted-}M = \sum_{i=1}^{C} w_i \cdot M_i \quad \text{where} \quad w_i = \frac{n_i}{\sum_{j=1}^{C} n_j} \tag{6}$$

  Here, n_i is the number of samples in class i, and w_i is its corresponding weight.
- Although the ROC and AUC results are presented in a separate section, we briefly formalise AUC here as a measure of class separability. For each class i, the AUC is defined as:

$$\text{AUC}_i = \int_0^1 \text{TPR}_i(t) \cdot \frac{d}{dt}\text{FPR}_i(t)\, dt \tag{7}$$

  Here, TPR_i and FPR_i represent the true positive and false positive rates at threshold t. The macro-AUC is the unweighted mean of all class-wise AUCs.

Models that demonstrate increased AUC values are considered more reliable in differentiating between the stages of ROP. Class-specific metrics were used to examine performance across all four diagnostic categories rather than relying only on an overall accuracy value. Because the dataset contains similar numbers of images for Healthy, Type 1 ROP, Type 2 ROP, and RD, the evaluation focuses on differences in classification behavior among the categories rather than on majority- and minority-class performance. Fold-to-fold variability was also examined during cross-validation. Where reported, the standard deviation of accuracy and F1-score across the five folds was used to quantify variation in model performance between data partitions.

3.7. Cross-Validation Strategy

A 5-fold cross-validation procedure was used to evaluate the six model configurations. The dataset was divided into five folds while preserving the class proportions as closely as possible across the partitions. In each iteration, four folds were used for model training and the remaining fold was held out for testing. The process was repeated five times so that each fold served once as the test set. Table 2 summarizes the corresponding four-fold training and one-fold test allocation. Using multiple held-out partitions reduces dependence on the outcome of a single train-test split and permits the six models to be evaluated under the same partitioning procedure. Performance metrics were calculated for each cross-validation iteration and used to examine both overall classification performance and variation across folds.

4. Results and Discussion

= This section reports the experimental configuration and the classification results for the six model configurations. Performance is assessed using accuracy, sensitivity, specificity, precision, F1-score, AUC, confusion matrices, and receiver operating characteristic (ROC) curves.

4.1. Experimental Setup

The experiments were conducted on a system equipped with an Intel Core i5 processor, an NVIDIA RTX 3080 GPU, and 32 GB of RAM. Model training and evaluation followed the 5-fold cross-validation procedure described in the 3.7. Cross-Validation Strategy subsection. In each iteration, approximately 80% of the images were contained in the four training folds, while the remaining 20% formed the held-out test fold. The CNN models were initialized using pretrained weights and subsequently fine-tuned for the four-class ROP classification task. The pretrained networks were adapted to produce four output scores corresponding to Healthy, Type 1, Type 2, and RD. Class probabilities were obtained using the softmax function. Model parameters were optimized using Adam with a learning rate of 1×10^{-4} and a batch size of 32. Categorical cross-entropy was used as the loss function. Training was conducted for up to 50 epochs, with early stopping based on validation loss.

4.2. Classification Performance and Evaluation Metrics

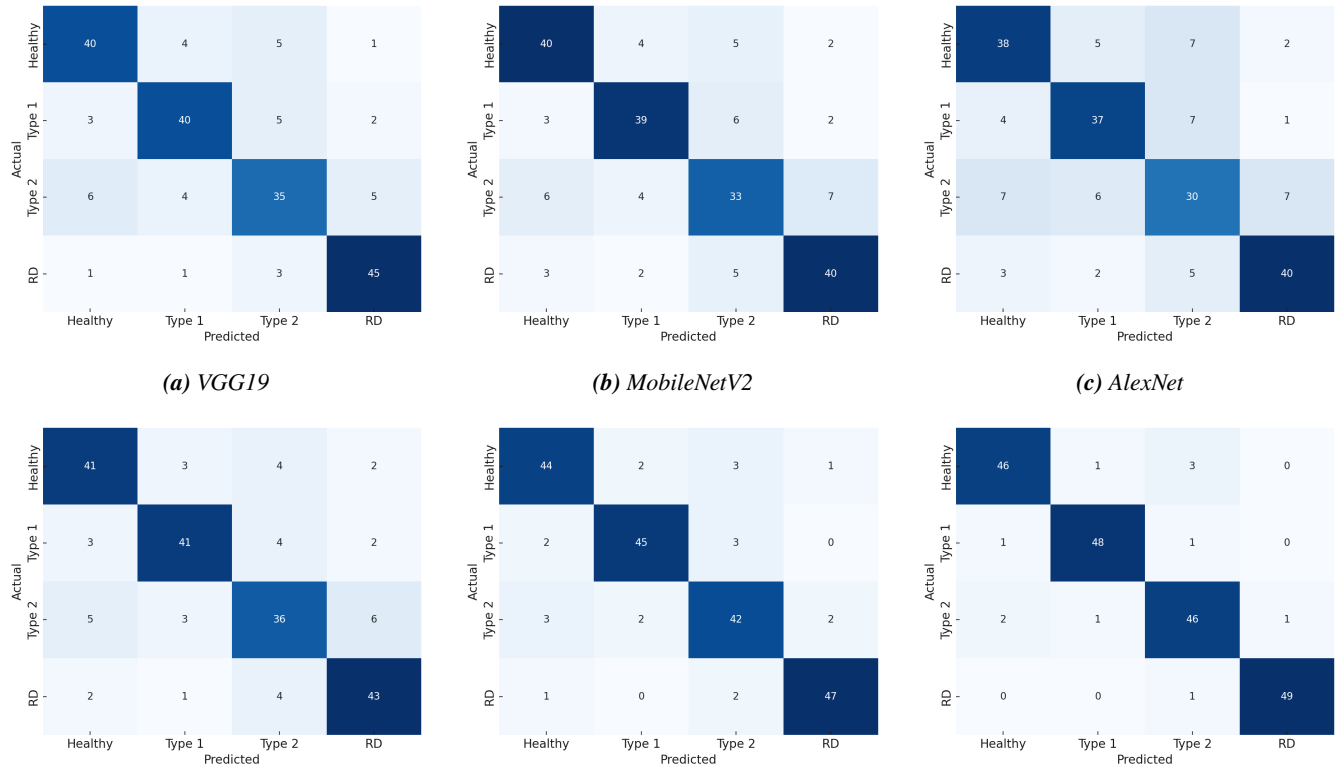
Model performance was evaluated using accuracy, precision, sensitivity, specificity, and F1-score. Confusion matrices and ROC curves were also examined to assess class-level prediction behavior. All six architectures were evaluated after transfer-learning-based fine-tuning rather than training from random initialization. Table 4 reports the summary classification results. CBAM-ResNet50 has the highest values among the six evaluated models for all five reported metrics, with an accuracy of 95.4%, sensitivity of 96.1%, specificity of 94.8%, precision of 95.7%, and F1-score of 95.9%. EfficientNetV2-S provides the second-highest accuracy at 93.2%, followed by VGG19 at 87.4%, DenseNet121 at 84.6%, MobileNetV2 at 82.7%, and AlexNet at 79.3%. These results establish the relative performance of the models under the evaluation protocol used in this paper and they do not by themselves isolate which architectural component is responsible for the observed differences.

Table 4: *Performance evaluation of the DNN models on ROP classification.*

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
CBAM-ResNet50	**95.4**	**96.1**	**94.8**	**95.7**	**95.9**
EfficientNetV2-S	93.2	92.8	93.6	93.0	92.9
VGG19	87.4	85.6	89.1	86.8	86.2
DenseNet121	84.6	82.1	86.5	83.9	83.0
MobileNetV2	82.7	80.5	84.9	82.2	81.3
AlexNet	79.3	77.6	81.1	78.4	78.0

Figure 5 presents the confusion matrices obtained for the six architectures. For the evaluation represented in the figure, CBAM-ResNet50 correctly classified 49 of 50 RD instances and showed relatively few errors between Type 1 and Type 2 ROP. EfficientNetV2-S correctly classified at least 45 instances in each of the four categories shown in its confusion matrix. Its displayed matrix contains more confusion between Healthy and Type 2 images than the CBAM-ResNet50 matrix.

DenseNet121 shows more Type 2-RD classification errors in the displayed evaluation. VGG19, MobileNetV2, and AlexNet also contain more off-diagonal predictions than CBAM-ResNet50. Because the architectures differ in several design characteristics, these differences cannot be attributed specifically to the presence or absence of an attention mechanism without a separate ablation experiment.



(a) *VGG19* ***(b)*** *MobileNetV2* ***(c)*** *AlexNet*

(d) *DenseNet121* ***(e)*** *EfficientNetV2-S* ***(f)*** *CBAM-ResNet50*

Figure 5: *Confusion matrices for the six architectures evaluated on the ROP classification task.*

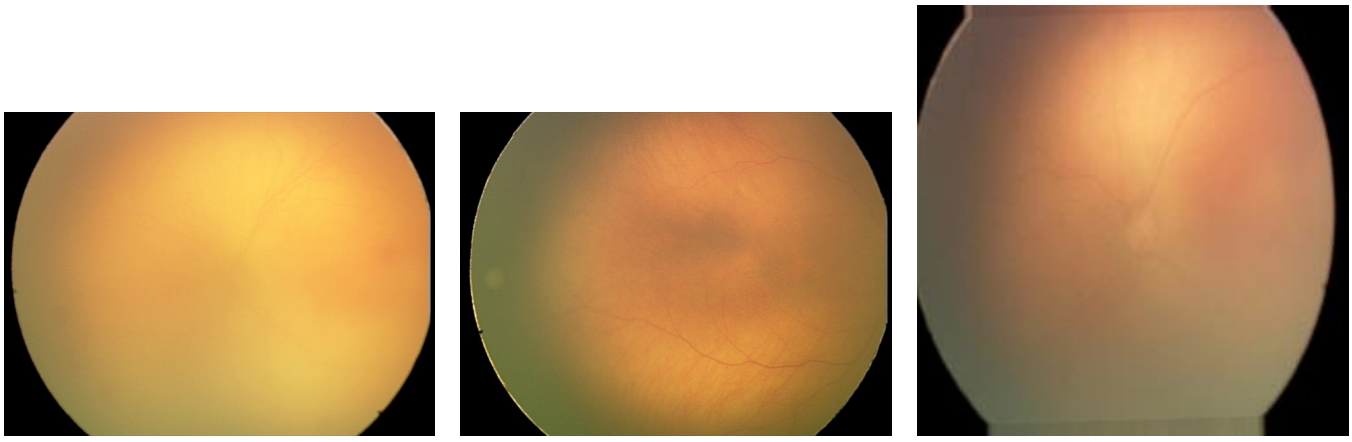
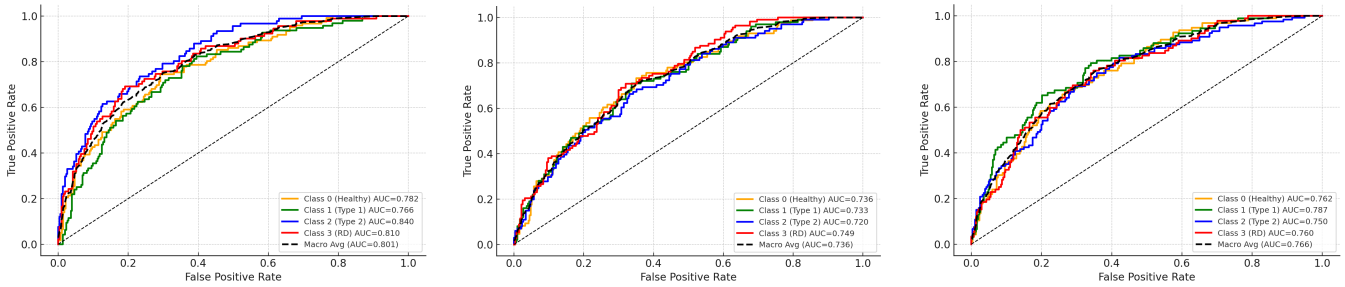
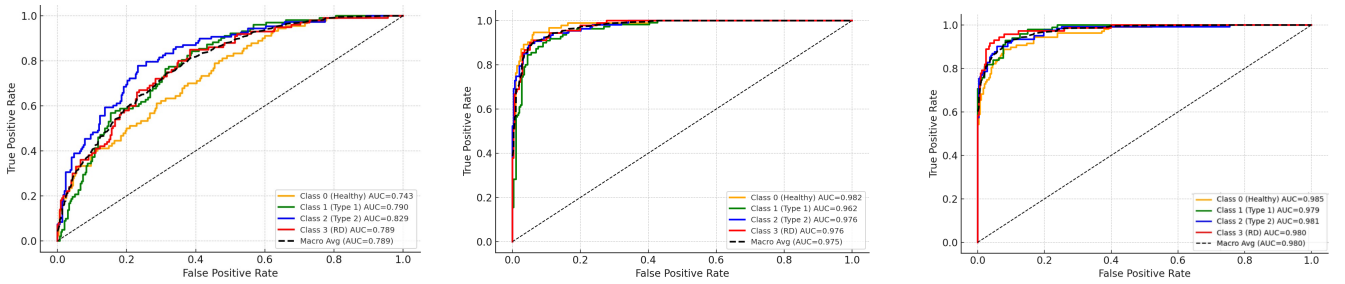


Figure 6: *Examples of misclassified retinal images.*

Figure 6 shows representative images that were assigned to an incorrect class. Some of these examples contain low contrast or visually ambiguous retinal structures, which may have made discrimination more difficult. Healthy images were occasionally assigned to the Type 1 category. Misclassification was also observed for Type 2 images with CBAM-ResNet50. These errors may reflect overlap in the retinal appearance represented by adjacent diagnostic categories, although the present experiments do not separately quantify the effect of image quality or visual ambiguity on prediction errors.

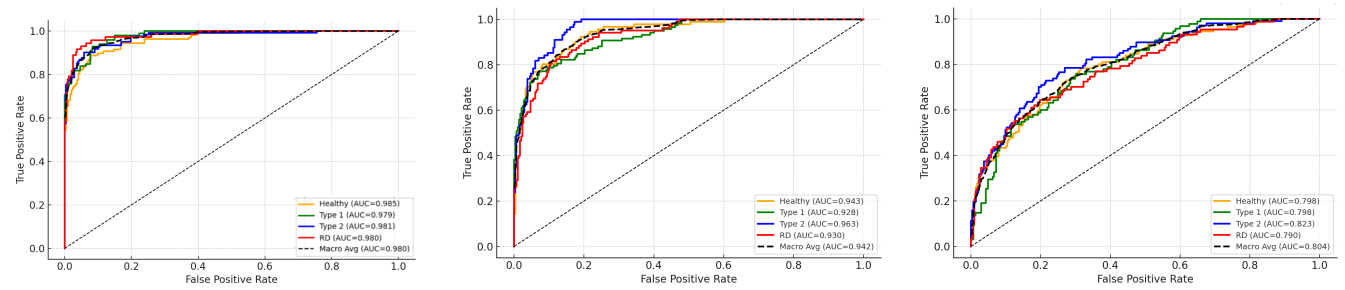


(a) VGG19 *(b) MobileNetV2* *(c) AlexNet*

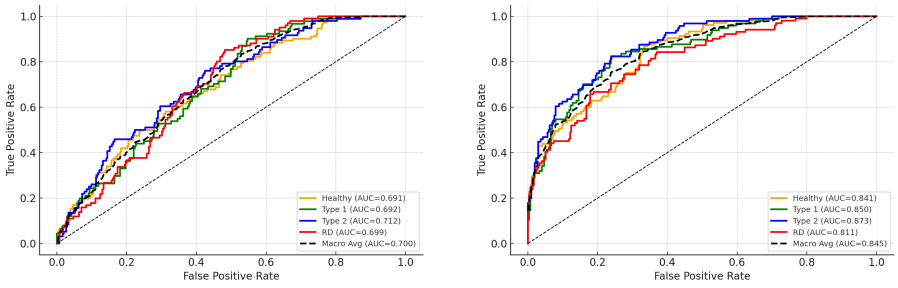


(d) DenseNet121 *(e) EfficientNetV2-S* *(f) CBAM-ResNet50*

Figure 7: *ROC curves for the six pretrained models evaluated on the ROP classification task.*



(a) Fold 1 *(b) Fold 2* *(c) Fold 3*



(d) Fold 4 *(e) Fold 5*

Figure 8: *ROC curves for CBAM-ResNet50 across the five cross-validation folds.*

Figure 7 presents the ROC curves for the six models across the Healthy, Type 1, Type 2, and RD classes. The figure includes class-specific ROC curves and the corresponding macro-average AUC. CBAM-ResNet50 achieved a

macro-average AUC of approximately 0.98, followed by EfficientNetV2-S at approximately 0.97. For both models, the class-specific AUC values shown in the figure are above 0.96. The remaining models produced lower macro-average AUC values. VGG19 reached 0.801, DenseNet121 0.789, AlexNet 0.766, and MobileNetV2 0.736. These networks differ in several architectural aspects, so the observed performance differences cannot be assigned to a single factor such as depth or the use of attention. The fold-wise ROC results for CBAM-ResNet50 are shown in Figure 8. The macro-AUC was 0.975 in Fold 5, 0.942 in Fold 4, and 0.845 in Fold 3. The reported mean across the five folds was 0.854. Fold 1 gave the lowest macro-AUC, at 0.700, with poorer separation for RD and Type 1 ROP. The experiments did not isolate the cause of these differences between folds. Variations in image quality, class composition, or image characteristics may have contributed, but this was not tested separately.

5. Conclusion and Future Work

We compared six pretrained CNN configurations for four-class ROP classification from retinal fundus images. The four classes were Healthy, Type 1, Type 2, and RD. Each model was trained with transfer learning and evaluated using 5-fold cross-validation with training-time image augmentation. CBAM-ResNet50 recorded the highest values in Table 4, with 95.4% accuracy, 96.1% sensitivity, 94.8% specificity, 95.7% precision, and a 95.9% F1-score. EfficientNetV2-S achieved the second-highest accuracy at 93.2%. The confusion-matrix results also show fewer classification errors for CBAM-ResNet50 in the displayed evaluation than for the other models. These comparisons indicate that CBAM-ResNet50 performed best under the experimental conditions used in this paper. However, the present experiments do not isolate the contribution of the attention module from other architectural differences. The fold-specific ROC analysis also shows that performance was not uniform across data partitions. CBAM-ResNet50 reached a macro-AUC of 0.975 in Fold 5 and 0.942 in Fold 4, while the lowest reported value was 0.700 in Fold 1. The mean macro-AUC across the five folds was 0.854. This variation indicates that performance remains sensitive to the composition of the evaluation fold and should therefore be examined on additional external clinical datasets before conclusions about generalisation are made. The current results support further investigation of transfer-learning-based ROP classification as a decision-support approach. Clinical deployment, real-time operation, and reduction of screening workload were not evaluated directly in this paper and therefore remain subjects for future validation.

Future work will examine longitudinal retinal image sequences for disease-progression modelling and early-stage risk prediction. Additional clinical datasets with greater variation in patient characteristics and acquisition conditions will be considered to assess generalisation across sites. Domain adaptation will also be investigated to address differences between imaging environments. Finally, model compression and quantisation will be studied for deployment on mobile and edge devices and for possible integration with teleophthalmology systems.

Author Declarations

Author Contributions (CRediT): Muhammad Haseeb Ur Rehman: Conceptualization, Methodology, System Design, Investigation, Visualization, and Writing - Original Draft. Saqib Dar: Software, Code implementation, Writing - Original Draft, Writing - Review & Editing. All authors have read and approved the submitted version of the paper.

Funding: This research received no external funding.

Data Availability Statement: All datasets and source code used in this paper are publicly available in the following GitHub repository: https://github.com/mehaseeburrehman/RetinopathyOfPrematurity.

Acknowledgments: The authors would like to thank the Center for Emerging Networks & Technologies (CENT) Lab at the University of Central Punjab (UCP), Lahore, Pakistan, for technical support and constructive feedback.

Conflicts of Interest: The authors declare no competing interests.

Ethics Statement: Not applicable.

Declaration of Generative AI: During the preparation of this work, the authors used generative AI tools only for language editing and proofreading, where applicable. The authors reviewed and edited the content and take full responsibility for the published paper.

References

[1] O. Dammann, M. E. Hartnett, and A. Stahl, "Retinopathy of prematurity," *Developmental Medicine and Child Neurology*, vol. 65, no. 5, pp. 625-631, 2023, doi: 10.1111/dmcn.15468.

[2] M. W. Nadeem, H. G. Goh, M. Hussain, S.-Y. Liew, I. Andonovic, and M. A. Khan, "Deep learning for diabetic retinopathy analysis: A review, research challenges, and future directions," *Sensors*, vol. 22, no. 18, p. 6780, 2022, doi: 10.3390/s22186780.

[3] E. Koc and A. Y. Bas, "Retinopathy of prematurity (ROP): From the perspective of the neonatologist," *Global Pediatrics*, vol. 8, p. 100159, 2024, doi: 10.1016/j.gpeds.2024.100159.

[4] N. Shaikh *et al.*, "Glaucoma in retinopathy of prematurity: A review," *Survey of Ophthalmology*, 2025, doi: 10.1016/j.survophthal.2025.03.006.

[5] V. M. R. Sankari, U. Snekhalatha, A. Chandrasekaran, and P. Baskaran, "Automated diagnosis of retinopathy of prematurity from retinal images of preterm infants using hybrid deep learning techniques," *Biomedical Signal Processing and Control*, vol. 85, p. 104883, 2023, doi: 10.1016/j.bspc.2023.104883.

[6] A. Song *et al.*, "Retinopathy of prematurity training and education: A systematic review," *Survey of Ophthalmology*, 2025, doi: 10.1016/j.survophthal.2025.06.007.

[7] S. S. Hussain, P. M. Shah, H. Dawood, and et al., "A swin transformer and CNN fusion framework for accurate parkinson disease classification in MRI," *Scientific Reports*, vol. 15, p. 15117, 2025, doi: 10.1038/s41598-025-93671-5.

[8] F. Rustam, A. S. Al-Shamayleh, R. Shafique, and et al., "Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model," *Scientific Reports*, vol. 14, p. 23274, 2024, doi: 10.1038/s41598-024-74357-w.

[9] V. Kumar, H. Patel, K. Paul, and S. Azad, "Deep learning-assisted retinopathy of prematurity (ROP) screening," *ACM Transactions on Computing for Healthcare*, vol. 4, 2023, doi: 10.1145/3596223.

[10] R. Shemesh, M. Chiang, R. V. P. Chan, and et. al, "An international survey on retinopathy of prematurity practice patterns during the COVID-19 pandemic and lessons for future management," *International Ophthalmology*, vol. 44, p. 407, 2024, doi: 10.1007/s10792-024-03290-8.

[11] R. H. Gensure, M. F. Chiang, and J. P. Campbell, "Artificial intelligence for retinopathy of prematurity," *Current Opinion in Ophthalmology*, vol. 31, no. 5, pp. 312-317, Sep. 2020, doi: 10.1097/ICU.0000000000000680.

[12] A. Jafarizadeh, S. F. Maleki, P. Pouya, and et. al, "Current and future roles of artificial intelligence in retinopathy of prematurity," *Artificial Intelligence Review*, vol. 58, p. 188, 2025, doi: 10.1007/s10462-025-11153-6.

[13] A. D'Angelo, F. Lixi, L. Vitiello, V. Gagliardi, A. Pellegrino, and G. Giannaccare, "The role of diet and oral supplementation for the management of diabetic retinopathy and diabetic macular edema: A narrative review," *BioMed Research International*, vol. 2025, p. 6654976, 2025, doi: 10.1155/bmri/6654976.

[14] M. K, S. M. John, A. Joseph, and B. Abraham, "A study on retinopathy of prematurity(ROP) screening using deep learning approaches and a blood vessel segmentation framework for ROP affected eyes," in *2024 second international conference on emerging trends in information technology and engineering (ICETITE)*, 2024, pp. 1-10. doi: 10.1109/ic-ETITE58242.2024.10493798.

[15] A. H. Khan, H. Malik, W. Khalil, S. K. Hussain, T. Anees, and M. Hussain, "Spatial correlation module for classification of multi-label ocular diseases using color fundus images," *Computers, Materials and Continua*, vol. 76, no. 1, pp. 133-150, 2023.

[16] J. Chen, Y. Zhu, L. Li, and et. al, "Visual impairment burden in retinopathy of prematurity: Trends, inequalities, and improvement gaps," *European Journal of Pediatrics*, vol. 183, pp. 1891-1900, 2024, doi: 10.1007/s00431-024-05450-5.

[17] H. Hashemian *et al.*, "Application of artificial intelligence in ophthalmology: An updated comprehensive review," *Journal of Ophthalmic and Vision Research*, vol. 19, pp. 354-367, 2024, doi: 10.18502/jovr.v19i3.15893.

[18] S. Mushriff and et. al, "Risk factors associated with retinopathy of prematurity," *International Journal of Health Sciences*, vol. 6, no. S9, pp. 4895-4909, 2022, doi: 10.53730/ijhs.v6nS9.14198.

[19] Y. Zhong, X. Lin, Y. Yang, and et. al, "Screening for retinopathy of prematurity in china: A five-year cohort study in seven screening centers," *BMC Ophthalmology*, vol. 25, p. 3, 2025, doi: 10.1186/s12886-024-03836-5.

[20] A. T. Wang and S. Dai, "Preferred treatment patterns of retinopathy of prematurity: An international survey," *Pediatric Reports*, vol. 16, no. 3, pp. 816-822, 2024, doi: 10.3390/pediatric16030069.

[21] H. Sun *et al.*, "Identification of genetic factors underlying severe retinopathy of prematurity in preterm infants," *Molecular Vision*, vol. 31, pp. 33-43, 2025.

[22] P. Huang *et al.*, "ROPRNet: Deep learning-assisted recurrence prediction for retinopathy of prematurity," *Biomedical Signal Processing and Control*, vol. 100, p. 107135, 2025, doi: 10.1016/j.bspc.2024.107135.

[23] W. Yang *et al.*, "An interpretable system for screening the severity level of retinopathy in premature infants using deep learning," *Bioengineering*, vol. 11, no. 8, p. 792, 2024, doi: 10.3390/bioengineering11080792.

[24] B. Ebrahimi *et al.*, "Assessing spectral effectiveness in color fundus photography for deep learning classification of retinopathy of prematurity," *Journal of Biomedical Optics*, vol. 29, no. 7, p. 076001, 2024, doi: 10.1117/1.JBO.29.7.076001.

[25] X. Luo, S. Zhao, Y. Chen, G.-S. Ying, and L. He, "VGG-ST: A VGG-swin transformer-based model for ROP disease diagnosis," in *2024 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2024, pp. 4987-4994. doi: 10.1109/BIBM62325.2024.10822551.

[26] V. M. R. Sankari, U. Umapathy, S. Alasmari, and S. M. Aslam, "Automated detection of retinopathy of prematurity using quantum machine learning and deep learning techniques," *IEEE Access*, vol. 11, pp. 94306-94321, 2023, doi: 10.1109/ACCESS.2023.3311346.

[27] Y. Peng *et al.*, "Automatic zoning for retinopathy of prematurity with a key area location system," *Biomedical Optics Express*, vol. 15, no. 2, pp. 725-742, 2024, doi: 10.1364/BOE.506119.

[28] N. Salih, M. Ksantini, N. Hussein, D. Ben Halima, A. Abdul Razzaq, and S. Ahmed, "Deep learning models and fusion classification technique for accurate diagnosis of retinopathy of prematurity in preterm newborn," *Baghdad Science Journal*, vol. 21, no. 5, 2024, doi: 10.21123/bsj.2023.8747.

[29] M. Yurdakul, K. Uyar, Ş. Taşdemir, and İ. Atabaş, "ROPGCViT: A Novel Explainable Vision Transformer for Retinopathy of Prematurity Diagnosis," *IEEE Access*, vol. 13, pp. 77064-77079, 2025, doi: 10.1109/ACCESS.2025.3564213.

[30] Y. Chu *et al.*, "Image analysis-based machine learning for the diagnosis of retinopathy of prematurity: A meta-analysis and systematic review," *Ophthalmology Retina*, vol. 8, no. 7, pp. 678-687, 2024, doi: 10.1016/j.oret.2024.01.013.

[31] C. Durmaz Engin, T. Ozturk, O. Ozkan, and et al., "Prediction of retinopathy of prematurity development and treatment need with machine learning models," *BMC Ophthalmology*, vol. 25, p. 194, 2025, doi: 10.1186/s12886-025-04025-8.

[32] B. Young *et al.*, "Deep learning-based prediction of genetic risk in retinopathy of prematurity using retinal images," *Investigative Ophthalmology & Visual Science*, vol. 65, no. 7, p. 3764, 2024.

[33] A. S. Govind, A. Gadad, D. J. Garodia, A. Gupta, A. Vinekar, and G. Srinivasa, "Deep learning for classifying stages of retinopathy of prematurity," in *2024 international conference on advances in modern age technologies for health and engineering science (AMATHE)*, Shivamogga, India, 2024, pp. 1-6. doi: 10.1109/AMATHE61652.2024.10582102.

[34] S. Chen, X. Zhao, Z. Wu, and et. al, "Multi-risk factors joint prediction model for risk prediction of retinopathy of prematurity," *EPMA Journal*, vol. 15, pp. 261-274, 2024, doi: 10.1007/s13167-024-00363-7.

[35] G. Hubert and S. Silvia Priscila, "Comprehensive prediction of retinopathy in preterm infants using deep learning approaches," in *Advancing intelligent networks through distributed optimization*, S. S. Rajest and others, Eds., Hershey, PA: IGI Global Scientific Publishing, 2024, pp. 353-370. doi: 10.4018/979-8-3693-3739-4.ch018.

[36] A. Ortiz, S. Patino, J. Torres, *et al.*, "AI-enabled screening for retinopathy of prematurity in low-resource settings," *JAMA Network Open*, vol. 8, no. 4, p. e257831, 2025, doi: 10.1001/jamanetworkopen.2025.7831.

[37] D. Lakshmanan, N. R. S. Raaj, S. Ravi, M. Ayyaswamy, and S. Ragunathan, "Retinopathy detection of premature using machine learning algorithms," in *AIP conference proceedings*, AIP Publishing, 2025, p. 030001. doi: 10.1063/5.0271997.

[38] J. Timkoviç, J. Nowakovą, J. Kubịçek, and et. al, "Retinal image dataset of infants and retinopathy of prematurity," *Scientific Data*, vol. 11, p. 814, 2024, doi: 10.1038/s41597-024-03409-7.

[39] A. S. Coyner *et al.*, "Use of an artificial intelligence-generated vascular severity score improved plus disease diagnosis in retinopathy of prematurity," *Ophthalmology*, vol. 131, no. 11, pp. 1290-1296, 2024, doi: 10.1016/j.ophtha.2024.06.006.

[40] K. Deepthi, M. S. Josephine, and V. Jayabala Raja, "Automated diagnosis of plus disease in retinopathy of prematurity based on transformer-based unsupervised curriculum learning," *Biomedical Signal Processing and Control*, vol. 104, p. 107521, 2025, doi: 10.1016/j.bspc.2025.107521.