IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 March 2025, accepted 22 April 2025, date of publication 24 April 2025, date of current version 7 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3564213

RESEARCH ARTICLE

ROPGCViT: A Novel Explainable Vision Transformer for Retinopathy of Prematurity Diagnosis

MUSTAFA YURDAKUL[1], KÜBRA UYAR[2], ŞAKİR TAŞDEMİR[3], AND İRFAN ATABAŞ[1]
[1]Computer Engineering Department, Kırıkkale University, 71450 KÄ±rÄ±kkale, Türkiye
[2]Computer Engineering Department, Alanya Alaaddin Keykubat University, 07425 Antalya, Türkiye
[3]Computer Engineering Department, Selçuk University, 42130 Konya, Türkiye

Corresponding author: Mustafa Yurdakul (mustafayurdakul@kku.edu.tr)

ABSTRACT Retinopathy of Prematurity (ROP) is a severe disease that occurs in premature babies due to abnormal development of retinal vessels and can lead to permanent vision loss. Fundus images are critical in the diagnosis of ROP; however, the examination of fundus images is a subjective, time-consuming, and error-prone process that requires experience. This situation can lead to delayed diagnosis and inaccurate evaluations. Therefore, the need for computer-aided diagnosis (CAD) systems is increasing day by day. Deep learning (DL) methods have a high potential in analyzing such complex images. In this study, a total of 50 DL models, 25 Convolutional Neural Network (CNN), and 25 Vision Transformer (ViT) models were tested to diagnose ROP from fundus images. Furthermore, the ROPGCViT model based on the Global Context Vision Transformer (GCViT) was proposed. GCViT was enhanced with Squeeze-and-Excitation (SE) block and Residual Multilayer Perceptron (RMLP) structures to effectively learn local and global context information. With a dataset of 1099 fundus images, the performance of the model was evaluated in terms of accuracy, precision, recall, f1-score, and Cohen's kappa score. To enhance explainability, the Gradient-Weighted Class Activation Mapping (Grad-CAM) method was utilized to visualize the regions of fundus images the model focused on during classification, providing insights into its decision-making process. ROPGCViT outperformed both 50 DL models and methods in the literature with 94.69% accuracy, 94.84% precision, 94.69% recall, 94.60% f1-score, and Cohen's kappa score of 93.10%. Additionally, the Grad-CAM visualizations demonstrated the ability of the model to focus on clinically relevant regions, enhancing trust and interpretability for experts. The proposed ROPGCViT model provides a robust solution for ROP diagnosis with high accuracy, flexibility, and generalization capacity.

INDEX TERMS Retinopathy of prematurity, vision transformer, convolutional neural network, deep learning, squeeze-and-excitation, grad-CAM.

I. INTRODUCTION

The eyes are one of the most vital sensory organs for perceiving and understanding the environment. The eyes perceive light, convert visual information into electrical signals, and transmit them to the brain, allowing us to see our surroundings [1]. However, many factors such as congenital genetic predisposition, environmental factors, and lifestyle can make the eyes susceptible to certain diseases.

ROP is a dangerous eye disease seen in premature babies. Retinal vascularization, which should be completed before birth under normal conditions, cannot be completed with premature birth [2]. Factors such as high oxygen concentrations used to meet the oxygen needs of babies trigger ROP and lead to abnormal vascular growth [3]. These abnormal vessels can cause tears and hemorrhages in the retinal layer and even lead to loss of vision.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

Studies show that approximately 30-50% of babies born at 28 weeks or earlier and 10-15% of those born between 28-32 weeks have ROP. Worldwide, ROP constitutes about 10% of all visual impairments [4]. Globally, each year an estimated 30000 premature babies suffer from severe vision loss due to ROP. As Jalali et al. [5] stated, early diagnosis of ROP is crucial for preserving visual acuity in newborns and minimizing complications. For the early diagnosis of eye disorders, detailed visualization of the back of the eye has an important role [6], [7]. Although techniques such as optical coherence tomography (OCT) and ultrasound biomicroscopy are used for this purpose, retinal imaging devices that provide clearer and more detailed images are preferred in cases such as ROP, cataract, glaucoma, Diabetic Retinopathy (DR), and macular edema. Among these devices, the fundus camera can provide high-resolution images. The evaluation of fundus images in the diagnosis of ROP is a complex and subjective process and is performed by ophthalmologists. On the other hand, lack of ophthalmologists, high medical consultation fees and labor intensity have led to an increasing need for CAD systems. A CAD system can automatically analyze fundus images for more accurate and faster detection of ROP disease. In this way, patients can be treated more quickly without negative consequences such as vision loss.

In recent years, artificial intelligence has shown great potential in image-based disease diagnosis [8], [9]. DL algorithms have shown very successful results in the diagnosis of many eye diseases such as glaucoma [7], [10], DR [11], age-related macular degeneration [12], and cataracts [13]. Recently, in literature some studies were conducted to diagnose the ROP from medical images.

Zhao et al. [14] presented a large dataset of 1,099 fundus images of 483 premature infants that includes normal, laser scars, stage 1, stage 2, and stage 3 classes for use in artificial intelligence systems for ROP. In the study, the images were classified with ResNet50, ResNet101, ConvNeXt-T, and ViT-B models, and accuracy rates of 87.61%, 79.65%, 83.19%, and 84.07% were obtained, respectively. Among the models, ResNet50 showed the best performance.

Ding et al. [15] proposed a hybrid model combining segmentation and CNN for ROP analysis. The hybrid model performs classification by incorporating the mask obtained from an object segmentation model that detects demarcation lines in retinal images into CNN as an additional channel of the image. In the experiments, the hybrid model improved accuracy by 13% and 20% compared to CNN-only or segmentation-only models, respectively. The model achieved 67% overall accuracy and f1-scores of 78% for ROP stage 1, 61% for stage 2, and 62% for stage 3.

Tan et al. [16] proposed a novel DL model called ROP.AI to recognize the existence of disease on normal and plus disease fundus images. The ROP.AI model is based on the Inception-v3 CNN architecture. They utilized a dataset containing a total of 6974 fundus images. The model achieved an accuracy of 97.3%, recall of 96.6%, and specificity of 98.0% on the internal test dataset (20% hold-out). However, on an external test set, the model's performance declined, achieving 93.9% recall and 80.7% specificity.

Brown et al. [17] used the InceptionV3 model on retinal images to automatically diagnose plus disease in ROP. Each image was assigned to normal, pre-plus disease, or plus disease categories by consensus of three experts. On the test dataset, the recall and specificity of the algorithm were 93% and 94% for plus disease and 100% and 94% for pre-plus disease, respectively.

Tong et al. [18] developed a system to automate ROP diagnosis and enhance diagnostic accuracy. ResNet-101 and Faster-RCNN models were trained to classify ROP severity and detect plus disease. The models achieved an overall accuracy of 90.3% for four-level classification (normal, mild, semi-urgent, urgent), 89.6% for plus disease detection, and 95.7% for ROP stage classification (stages I to V). Furthermore, the performance of the model was compared with evaluations by two retinal experts and demonstrated comparable or superior results.

Zhang et al. [19] proposed a DL model for the diagnosis of Aggressive Posterior Retinopathy of Prematurity (AP-ROP), an aggressive form of ROP. The proposed model was based on ResNet architecture and integrated Hierarchical Bilinear Pooling (HBP) and Squeeze and Excitation (SE) modules to improve performance. In the study, on a dataset with 12230 fundus images belonging to normal, regular ROP, and AP-ROP classes, 95.81% accuracy, 89.81% recall, 97.90% specificity, 93.69% precision, and 91.70% f1-score were obtained.

Sankari et al. [4] used MultiResUNet and Gaussian derived-filters for segmentation of retinal vessels in fundus images. After segmentation, GLCM (Gray Level Co-occurrence Matrix) and contour features were extracted and these features were classified with a Random Forest (RF) classifier. The proposed approach achieved 94.5% accuracy, 94% recall, and 93% specificity.

Salih et al. [20] utilized VGG-19, ResNet-50, EfficientNetB5, and custom CNN models for zone detection in ROP diagnosis. EfficientNetB5 demonstrated the best individual performance with 87.27% accuracy, 87.11% precision, 86.75% recall, and 86.84% f1-score. The voting classifier, combining all models, achieved the highest performance with 88.82% accuracy, 88.6% precision, 88.03% recall, and 88.17% f1-score.

Huang et al. [21] developed a DL model for the prediction of ROP recurrence, called ROPRNet. ROPRNet includes several components such as a cascaded deep network (CDN) designed for fundus images, enhancement head (EH) to reduce color and contrast differences, enhanced ConvNeXt (EnConvNeXt) for feature extraction, and multidimensional multiscale feature fusion (MMFF). In tests with real clinical data, the model demonstrated superior performance with an AUC of 89.4%, accuracy of 81.8%, recall of 82.8%, and specificity of 80%.

Sankari et al. [22] used the SegNet architecture for automatic segmentation of retinal vessels and created feature vectors by combining high-level features such as SURF and SIFT after segmentation. These features were classified using machine learning classifiers (SVM, REP Tree, K-Star, and LogitBoost) and Quantum Support Vector Machine (QSVM). QSVM exhibited higher accuracy (95.5%), recall (93%), specificity (98%), and AUC (97%) compared to other methods.

Attallah [23] proposed the DIAROP model, which is built using four different CNN models, namely ResNet-50, InceptionV3, Xception, and Inception-ResNetV2. From these models, spatial features are extracted using transfer learning and then spectral features are extracted and integrated using Fast Walsh Hadamard Transform (FWHT). The integrated features are reduced using autoencoder (AE), principal component analysis (PCA), and discrete wavelet transform (DWT) techniques. In the final stage, classification is performed using linear SVM, quadratic SVM, and linear discriminant analysis (LDA) classifiers. The model was tested on a dataset of 17801 images. The best performance was achieved by integrating ResNet-50, Inception and InceptionResNet features using DWT with 93.2% accuracy, 89.7% sensitivity, 96.4% specificity, and 92.3% f1-score.

Agrawal et al. [24] developed modified U-Net architectures with Attention Gates (AG) and SE blocks for automatic segmentation of retinal structures in ROP. The researchers used two separate datasets (HVDROPDB-BV and HVDROPDB-RIDGE) consisting of 4000 fundus images from two different imaging devices. Among these models, AG U-Net performed the best in the test data with 96% sensitivity, 89% specificity, and AUC values above 0.94.

Feng et al. [25] developed a semi-supervised DL model for automatic staging of ROP. The model aims to reduce the workload of doctors by combining a limited amount of labeled data with a large amount of unlabeled data. For this purpose, two different loss functions, prediction consistency loss and semantic structure consistency loss, are used in the model. Prediction consistency ensures that the model produces consistent predictions under different data augmentations, while semantic structure consistency helps to preserve the semantic relationships between ROP stages. In the study, 4796 training and 386 test fundus images. According to the test results, the model performed well, achieving 86.3% AUC, 78.95% accuracy, 78.68% sensitivity, and 80.1% specificity. Moreover, even when the amount of labeled data was reduced to 50%, results close to fully supervised models were obtained.

Yang et al. [26] developed an interpretable DL model to determine ROP severity. This model transparently assesses ROP severity by mimicking clinical screening processes. The operation of the model consists of lesion detection, stage determination and plus disease identification. First, lesions are detected with a RetinaNet-based model. This model determines lesion types and locations using the ResNet50 backbone. Then, panoramic images are acquired for each eye. The locations of the lesions are combined to determine the ROP stage. Finally, the presence of plus disease is assessed with an optic disc-centered image. The proposed model is trained on a dataset of 3330 images and achieved an AUC of 95% and an f1-score of 76%.

GabROP, developed by Attallah [27], is a CAD tool and its diagnosis process consists of four main stages. In the first stage, textual features are extracted using the GW method and combined using the Discrete Wavelet Transform (DWT), resulting in a textual-spectral-temporal representation. In the second stage, spatial features extracted from the original fundus images are combined with the features from the first stage. In the third stage, the Discrete Cosine Transform (DCT) is used to reduce the dimensionality of these combined features. Finally, three different classifiers (LDA, SVM, and ESD) are used for classification. GabROP improved its accuracy with three combination stages and achieved the best results in the third combination stage. In this stage, 93.9% accuracy was achieved with the SVM classifier. Other metrics are as follows: f1-score 93.02%, sensitivity 90.12%, and specificity 96.96%.

Ebrahimi et al. [28] investigated the effect of spectral channels in color fundus photography for DL classification of ROP. The study aimed to evaluate how different color channels (red, green, and blue) affect DL classification of ROP stages. The researchers used a dataset of 200 images from 158 patients, divided into four ROP stages (normal, stage 1, stage 2, stage 3), all captured using the RetCam imaging system. The method involved training CNN with separate color channels and multicolor channel fusion architectures, comparing performance between green, red and blue channels. The results revealed that the green channel performed the best with 88% accuracy, 76% sensitivity, and 92% specificity.

Jemshi et al. [29] proposed an artificial neural network (ANN) architecture for classifying ROP Plus disease. Their method combines vessel features (tortuosity, width, number of leaf nodes, and vessel density) and transform-based features derived from wavelet and curvelet transforms to improve classification performance. The dataset consists of 178 images from Narayana Nethralaya Eye Hospital, of which 81 have plus disease and 97 have no plus disease. The authors first preprocessed the retinal images by increasing contrast and reducing noise. Then, thick and thin vessels were extracted using top-line transformation and directed filters to segment the vessels. Feature extraction was done with both vessel features and features from wavelet and curvelet transforms. The study found that the combination of curvelet energy features and vessel features provided the best classification performance. The proposed ANN classifier achieved 100% sensitivity, 92% specificity, and 96% accuracy.

Wang et al. [30] proposed MOCO-MIM, a novel self-supervised learning framework for early detection of ROP. This method aims to overcome the challenges posed by

limited labeled datasets by combining contrast learning and masked image modeling (MIM). A small ROP dataset consisting of 553 labeled fundus images of 89 preterm infants was used to train the model. The proposed approach achieved impressive results with 98.29% test accuracy, 88.30% precision, 88.37% sensitivity, and 88.33% f1-score.

Table 1 provides a summary of the related studies, detailing their methodologies, results, and the limitations identified. When the literature studies in Table 1 are examined, many DL-based approaches have been proposed for ROP diagnosis, and very successful results have been obtained. In addition, in some studies, the performance of the DL model was compared with the performance of experts in this field, and it was observed that the DL model was more successful than the expert.

Existing approaches in the literature as summarized in Table 1 are generally based on CNN-based models and hand-crafted feature extraction. Hand-crafted feature extraction is a challenging and specialized process, especially for disease detection in complex data such as medical images. Detecting the early stages of diseases such as ROP can be challenging because hand-crafted features are often limited and may fail to capture fine details. Therefore, approaches based on hand-crafted features may reduce the generalizability of the models and challenge in scalability and adaptability. On the other hand, while CNN-based approaches are highly successful in learning local features in the image, they suffer from long-range feature extraction. This limitation can prevent accurate detection of diseases where subtle structural changes are important, especially in early stages. While CNNs generally perform better at recognizing advanced diseases, their accuracy may be insufficient in the early stages. At this point, both approaches have difficulties in detecting early-stage diseases and the need for a more flexible, explainable model that can capture long-range relationships and learn local and global features. In addition, integrating multiple architectures into a single framework improves the classification performance, however, these hybrid models often increase complexity, making them less practical for real-world applications. Finally, most of DL models mentioned in literature studies are not transparent and are seen as ''non-transparent models'' that negatively impact the reliability of the model in clinical settings.

From the analysis of literature studies, it is essential to develop flexible, robust, explainable, and more performant models for ROP diagnosis that effectively address the challenges of multi-stage classification. These models should take advantage of state-of-the-art DL techniques, strike the balance between complexity and practicality, and prioritize scalable solutions across diverse clinical and computational environments. Such advances will significantly enrich literature and contribute to the development of reliable, efficient and accessible automated ROP diagnostic systems.

In this context, the contributions and novelties of this paper can be listed as follows:

- The studies in the literature on ROP detection with DL methods have been extensively reviewed and analyzed.
- A total of 50 DL models, 25 CNN and 25 ViT models, are tested on the ROP dataset, making it the most comprehensive and comparative experiment in this field.
- The stem module of the GCViT architecture is enhanced with the SE attention module, and the MLP layers in the GC blocks are replaced with RMLP resulting in the proposed ROPGCViT model.
- The Grad-CAM method that makes the explainability of deep models effective is employed to visualize the regions of fundus images that the model focuses on during classification, providing explainability and helping experts understand the decision-making process of the ROPGCViT model.

The remainder of the paper is organized as follows: The Material and Methodology section explains the characteristics of the dataset, data distribution, and DL models used. The Proposed Model section describes the details of the ROPGCViT architecture. The Experimental Setup section describes the experimental environment, the software and hardware used, and the hyperparameters. In the Results section, the performance results and Grad-CAM analysis of the DL model are presented comparatively. In the Discussion section, the findings are compared with the literature and the strengths and weaknesses are analyzed. Finally, the Conclusion section summarizes the contributions of the study.

II. MATERIAL AND METHODOLOGY

The basic concepts of CNN and ViT, attention modules, and dataset were explained in this section.

A. ROP DATASET

The ROP dataset used in this study is proposed in the study [14]. The dataset was collected within a 19-year period between 2004 and 2023. Before fundus images were taken, the infants' pupils were dilated with tropicamide approximately 30 minutes before the examination. With this method, a total of 512 × 512 pixel fundus images with a total size of 1099, including normal, stage 1, stage 2, stage 3, and laser scars classes were obtained. The distribution of the data for training, validation, and testing as in the original is given in Table 2. Sample images of the dataset are shown in Figure 1. Normal images show healthy premature fundus images. Stage 1, 2, and 3 indicate the degree of retinopathy. Laser class represents images with laser scars after laser treatment for ROP.

Within the scope of the study, all fundus images were resized according to the input of the DL model. During the training process, online data augmentation techniques which are rotation, width shift, height shift, and shear were used. These techniques and parameter values are given in Table 3. In addition, examples of data augmentation techniques on images are shown in Figure 2.

TABLE 1. Comparative and critical summaries of literature studies on the diagnosis of ROP with DL-based methods.

Study	Methodology	Results	Limitations
Zhao et al. [14]	ResNet50	Accuracy:87.61% Precision:82.64% Recall:83.31% F1-score: 87.81%	The classification performance is limited.
Ding et al. [15]	Hybrid Model (Object Segmentation + CNN)	Overall Accuracy: 67% F1-scores: 78% for ROP stage 1, 61% for stage 2, 62% for stage 3.	The classification performance of the hybrid model is limited, and it has high computational complexity.
Tan et al. [16]	Proposed ROP.AI, a novel DL model for ROP diagnosis	Recall: 93.90% Specificity: 80.7%	The proposed ROP.AI suffered performance loss on test data.
Brown et al. [17]	Used InceptionV3 for diagnosing plus and pre-plus disease via 5-fold cross-validation	Plus disease: Recall: 93%, Specificity: 94% Pre-plus disease: Recall: 100%, Specificity: 94%	The study focused only on plus and pre-plus disease. The model could be extended to other levels of ROP.
Tong et al. [18]	Developed ResNet-101 and Faster-RCNN for ROP severity classification and plus detection	Four-level classification: Accuracy: 90.3% Plus disease detection: Accuracy: 89.6% ROP stage detection: Accuracy: 95.7%	The study achieved successful results in four-level classification, but by combining multiple methods, a complex structure was created and flexibility was lost.
Zhang et al. [19]	Proposed a ResNet-based model with HBP and SE modules for diagnosing AP-ROP	Accuracy: 95.81% Recall: 89.81% Specificity: 97.90% Precision: 93.69% F1-score: 91.70%	The study focused only on AP-ROP disease. The model could be extended to other levels of ROP.
Sankari et al. [4]	Developed a hybrid DL model using MultiResUNet and Random Forest for classifying Normal vs. ROP	Accuracy: 94.5% Recall: 94% Specificity: 93%	Binary classification (Normal and ROP) achieved success but suffers from the detection of different stages of ROP.
Salih et al. [20]	tilized VGG-19, ResNet-50, EfficientNetB5, and custom CNNs with ensemble voting classifier for retinal zone detection.	Voting classifier: Accuracy: 88.82% Precision: 88.6% Recall: 88.03% F1-score: 88.17%	There is an imbalance between performance and computational complexity. Limited performance despite increased computational complexity with the ensemble.
Huang et al. [21]	Proposed a novel model called ROPRNet for the detection of recurrence of ROP	Accuracy: 81.8% Recall: 82.8% Specificity: 80% AUC: 89.4%	The classification accuracy of the model developed for the detection of ROP recurrence is moderate. And its applicability to real-life application is a more difficult task.
Sankari et al. [22]	Proposed a novel approach combining SegNet-based segmentation with high-level feature extraction (SIFT and SURF) and classification using QSVM	Accuracy: 95.5% Recall: 93% Specificity: 98% AUC: 97%	The results of the study depend on features such as SIFT and SURFACE, and therefore the ability of feature-dominated models to generalize to different datasets is limited.
Attallah [23]	Proposed DIAROP, a hybrid model integrating features from ResNet-50, Inception-v3, and Inception-ResNet using FWHT and AE	Accuracy: 93.2% Recall: 93.0% Precision: 93.2% F1-score: 93.1%	The model does not classify ROP severity stages, limiting its use for comprehensive disease management.
Agrawal et al. [24]	Proposed modified U-Net architectures with AG and SE blocks for segmenting retinal structures in ROP.	Sensitivity: 96%, Specificity: 89% AUC: 94%	The model focused solely on segmentation tasks without classifying different stages of ROP, limiting its comprehensive diagnostic utility.
Feng et al. [25]	Semi-supervised DL model (ResNet-50 based)	Accuracy: 78.95%, Sensitivity: 78.68%, Specificity: 80.1 AUC: 86.3%	The model used semi-supervised learning to increase learning from unlabeled data, but still did not achieve optimal performance compared to fully supervised methods.
Yang et al. [26]	Using a RetinaNet-based model, lesion detection was performed, stage determination was performed with panoramic images, and plus disease was evaluated with optic disc-centered images.	AUC: 95% F1-score: 76	There are inconsistencies between the AUC and f1-score metrics, which may cause the model to over-identify or under-identify in some cases.
Attallah [27]	GabROP uses GW and a three-stage combination process (DWT and DCT) to extract textual and spatial features from fundus images and combines these features with three different classifiers such as LDA, SVM, and ESD for classification.	Accuracy: 93.9 F1-score 93.02% Sensitivity 90.12% Specificity 96.96%	The limitations of GabROP include increased model complexity due to the three-step fusion process, inference speed and high computational requirements, and the resulting increased cost, which pose challenges for rapid and efficient use in the clinical setting.

TABLE 1. *(Continued.)* **Comparative and critical summaries of literature studies on the diagnosis of ROP with DL-based methods.**

Ebrahimi et al. [28]	CNN model using monochrome channel images and multicolor channel fusion strategies	Accuracy: 88%, Sensitivity: 76% Specificity: 92%	A limited dataset of 200 color fundus images, which may have impacted the classification performance, potentially limiting the model's generalizability.
Jemshi et al. [29]	ANN that combines vessel features and features from wavelet and curvelet transforms	Accuracy: 96% Sensitivity: 100% Specificity: 92%	A limited dataset of 178 color fundus images, which may have impacted the classification performance, potentially limiting the model's generalizability.
Wang et al. [30]	The methodology involves a novel self-supervised learning approach, MOCO-MIM, which combines CL and MIM to enhance feature extraction from limited-labeled ROP fundus images for early-stage detection and classification.	Accuracy: 98.29% Precision: 88.30%, Sensitivity: 88.37%, F1-score: 88.33% AUC: 97.6%.	The imbalance between the metrics indicates that the model may be overfitting or underfitting.

TABLE 2. **Class-based distribution of training, validation, and test image data.**

Category	Training	Validation	Test	Total
Laser scars	274	34	35	343
Normal	188	24	24	236
Stage 1	75	9	10	94
Stage 2	132	16	17	165
Stage 3	208	26	27	261
Total	877	109	113	1099

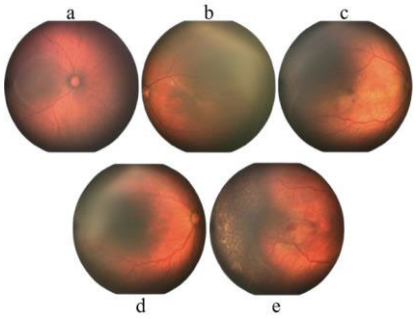

FIGURE 1. **Sample fundus images from the dataset (a: Normal, b: Stage 1, c: Stage 2, d: Stage 3, e: Laser scars).**

TABLE 3. **Data augmentation techniques and parameters.**

Method	Rotation	Width Shift	Height Shift	Shear
Parameter	0.3	0.4	0.3	0.2

FIGURE 2. **The dataset images illustrate the application of different data augmentation techniques: Rotation, width shift, height shift, and shear.**

B. CNN

In the realm of computer vision, the past decades have witnessed remarkable advances in various algorithms and architectures. These advances have significantly increased the capabilities of computer vision systems, enabling them to handle an ever-expanding range of tasks with impressive accuracy and efficiency. CNN is an artificial intelligence algorithm that has revolutionized the field of computer vision and achieved extraordinary success, especially in areas such as image classification, object recognition, and segmentation. CNN can automatically learn complex features from inputs through convolutional filters between layers, making them superior to traditional machine learning methods. CNN is highly successful in understanding spatial hierarchy by capturing local features of images, allowing them to move from simple edge and corner detection to more complex shape and object recognition. Furthermore, CNN architectures use pooling layers to reduce the size of the datasets, thereby reducing the computational burden while increasing the generalization capability of the model. Over the years, a number of variants of CNN architecture have been proposed, each offering unique strengths and addressing specific challenges. The DenseNet [31] architecture provides a more efficient learning process by using the feature maps of each layer from all previous layers. The Inception [32] architecture is a structure that can perform multiple feature extraction by using convolutional filters of different dimensions together. MobileNet [33] offers lightweight and efficient architecture using depth-wise separable convolution. ResNet [34] architecture enables training and more robust feature extraction of deep networks using skip connection. Xception [35] is an architecture based on depth separable convolutions and is designed to achieve higher accuracy rates. CSPNeXt [36] is an architecture that can efficiently learn high and low-frequency features using parallel large-core convolution and simple average pooling. GhostNet [37] uses a ghost module to enhance the feature extraction capability. ResNeSt [38] is a variant of ResNet that stacks split-Attention blocks.

C. ViT

In recent years, transformers have achieved very successful results in the field of NLP. Transformers' success is directly related to the self-attention in their architecture. Inspired by

the Transformer's success in NLP, transformers were used in image analysis and the first model named ViT was proposed in 2020 by Dosovitskiy et al [39]. The ViT architecture has three main components: the input embedding mechanism, the transformer encoder layers, and the output classification head. The input embedding mechanism adapts images to the transformer structure by segmenting and embedding them into fixed-size patches. The encoder layers learn complex representations with multi-headed self-attention mechanisms and feed-forward neural networks, providing an advantage over CNNs in understanding global context. The output head generates predictions using the output of the last transformer layer and usually includes a SoftMax function. The ViT model achieved 88.36% accuracy on ImageNet. After this impressive success, many other ViT models have been proposed.

D. GCViT

ViT models are a robust DL approach for learning short and long-range contexts of images. However, the high computational cost and low inductive bias of ViT have limited its practical use by increasing its dependence on large datasets. GCViT [40], which was developed as a solution to these problems, has a hierarchical structure that combines local and global attention mechanisms. With its innovative ''global query token'' mechanism, it collects contextual information from different image regions and combines it with local features to efficiently learn both short and long-range relationships. Due to its parameter-efficient design, GCViT has achieved highly successful results in tasks such as classification, object detection, and semantic segmentation while reducing computational cost in high-resolution images. Figure 3 shows a schematic diagram of the GCViT architecture. The input image is first processed in a stem layer and feature maps are generated by compressing the spatial dimensions. Local and global attention modules are then applied at each stage of the model. While local attention modules process short-range spatial relations, global attention modules capture global context information over the entire image. Global context information is provided by global token generation and these tokens are generated once at each stage and shared with local modules. The mathematical formula of the attention mechanism of GCViT is expressed in Equation 1.

$$Attention(q_g, k, v) = Softmax(\frac{q_g}{\sqrt{d}} + b)v \qquad (1)$$

where q_g is the global query token, k is the key token, v is the value token, d is the scaling factor, and b is the location bias. The formulation makes it possible to capture information over a wider area by combining both local and global context information. Furthermore, in each stage, the spatial dimensions are reduced by dividing by two and the number of channels is doubled. In the final stage, the feature maps are passed through the Global Average Pooling (GAP) layer and forwarded to the classification head. GCViT achieved 85.7% top-1 accuracy on the ImageNet-1K dataset, better than models such as Swin Transformer and ConvNeXt. With both high accuracy rates and low computational costs, GCViT offers a robust solution for modern computer vision tasks.

E. RMLP

The MLP structure used in the GCViT architecture first transmits the input features to a linear (dense) layer. The linear layer is implemented with the Gaussian Error Linear Unit (GELU) activation function. GELU is a non-linear activation function that allows the features to learn more complex relationships. After the activation process, a dropout layer is added to prevent the model from overfitting and to increase its generalization ability. This process is repeated with a second linear layer, allowing the MLP structure to learn deeper features. The MLP structure can be described mathematically by Equation 2, where Y represents the output, X is the input characteristics, W_1 is the weights of the first linear layer, B_1 is the bias term of the first linear layer, p is the dropout rate, W_2 is the weights of the second linear layer, B_2 is the bias term of the second linear layer.

$$Y = W_2.Dropout(GELU(W_1.X + B_1), p) + B_2 \qquad (2)$$

The proposed Residual Multilayer Perceptron (RMLP) structure includes a skip connection in addition to the standard MLP. Skip connection conveys the input features directly to the output and contributes to the learning process of the MLP. The use of skip connection plays a critical role in avoiding the vanishing gradient problem that is often encountered in deep models. In addition, it makes the model more stable during the training process. RMLP has the same components (Linear, GELU, Dropout) as the standard MLP, but it allows the model to learn deeper and more complex features thanks to the skip connection. In the GCViT architecture, it is an important contribution to improving the performance and generalization capacity of the model. The RMLP structure can be described mathematically by Equation 3. Schematic diagrams of the GCViT original MLP and RMLP are shown in Figure 4.

$$Y = W_2.Dropout(GELU(W_1.X + B_1, p) + B_2 + X) \qquad (3)$$

F. SE BLOCK

After the success of AlexNet over ImageNet, there have been ongoing efforts to improve model performance. In order to improve performance, the number of model layers has been increased and deep networks such as AlexNet with 8 layers, VGG with 16 layers, GoogleNet with 22 layers, and ResNet with 152 layers have been developed. Various approaches such as CNN-based feature extraction, hybrid architecture, advanced convolutions, and pooling methods have also been proposed. Another approach, the use of attention mechanisms, has also been an important milestone in improving performance. Attention mechanisms can be defined as structures that work by weighting the input tensor to emphasize the important regions that must be focused on.

One of the most frequently used attention mechanisms is SE block. The SE block is a key component introduced in the

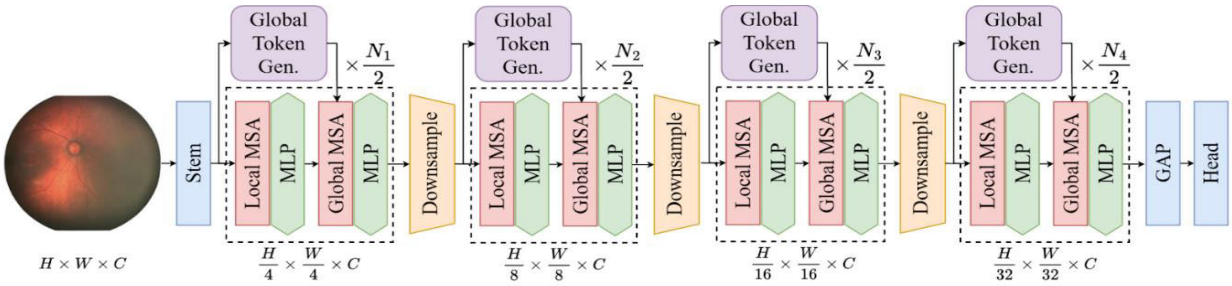

FIGURE 3. Schematic diagram of GCViT architecture.

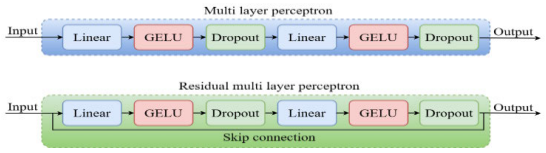

FIGURE 4. Baseline GCViT MLP layer and proposed residual MLP.

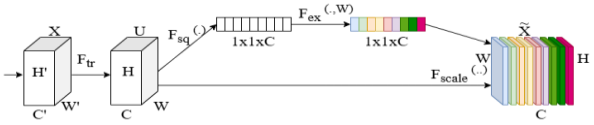

FIGURE 5. Conceptual diagram of the SE block.

Squeeze-and-Excitation Network (SENet) that significantly increases the representational capacity of CNNs by adaptively recalibrating channel-based feature responses. The SE block works in two phases: ''squeezing'' and ''excitation''. The squeezing phase involves applying global average pooling to the feature maps to summarize the overall information of each channel. In this way, it preserves important information while reducing the size of the data. The excitation phase determines the importance level for each channel by processing the feature map obtained at the end of the squeezing stage with fully connected layers and a sigmoid activation function. These important ratings ensure that more useful features are emphasized, and less useful ones are suppressed. Integrating the SE block into existing networks has led to significant performance improvements in many studies. A schematic diagram of the SE block is presented in Figure 5.

G. TRANSFER LEARNING

Transfer Learning, across different disciplines, refers to the reuse of knowledge acquired in one domain for problem-solving in another domain. In DL, it refers to the reuse of models trained on large datasets for small or different datasets. The model usually learns basic features on a large dataset and then is retrained for a new task. By reusing previously learned knowledge without having to train the model from scratch, both time and resources are saved, especially in the case of data constraints. TL provides effective results in areas such as classification, object detection, and natural language processing and plays a critical role for in a faster, more efficient training process. All DL models used in this study have weights trained on the ImageNet dataset. Since ImageNet has 1000 classes, the classifier layer in the last layers of the models is organized in such a way that the last layer can classify 5 classes in accordance with the purpose of the study.

III. ROPGCViT

The proposed model is based on GCViT architecture and enhanced with several improvements to make GCViT's local and global feature extraction more robust and effective. In the GCViT model, the first feature extraction from the image is performed in the stem block, which is the fundamental component of the model, and the features extracted from this block are used as input to the subsequent stages. Therefore, highlighting the important features extracted from this layer is crucial to improve the overall performance of the model. Therefore, highlighting the most critical features extracted at this layer is crucial to optimize the overall performance of the model. To address this need, an SE block has been integrated at the end of the stem layer to enable adaptive recalibration of channel-based features, effectively prioritizing salient visual information. Furthermore, the MLP layers in the model are redesigned as residual MLPs by inserting a skip connection. With this approach, the gradient vanishing problem, which is common in DL, is reduced and improves the model's ability to learn complex hierarchical features, ultimately contributing to more robust and efficient training dynamics. The model is named ROPGCViT since it is trained and tested on the ROP dataset. A schematic diagram of ROPGCViT is presented in Figure 6.

IV. EXPERIMENTAL SETUP

This section presents a comprehensive evaluation of advanced DL models for classification tasks of ROP stages. The experiments were conducted on a computer with the specifications listed in Table 4.

The data set for training the DL models was used as 80% train, 10% validation and 10% test and the data distribution is given in Table 2. In addition, for the training of the DL

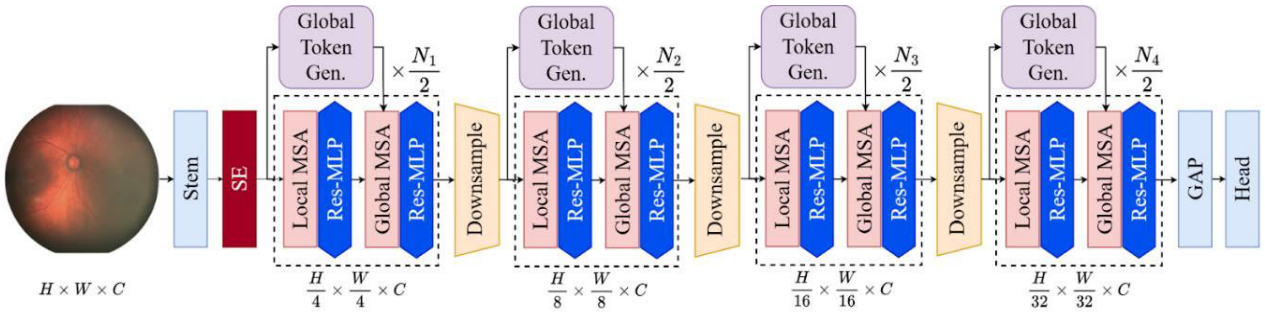

FIGURE 6. Schematic diagram of the proposed ROPGCViT architecture.

TABLE 4. Experiment environment configurations.

Configuration	Parameter
Operating System	Windows 11
Programming Language	Python version 3.11.4
Frameworks	Tensorflow 2.14.0, Keras version 2.11.4, Matplotlib 3.7.1
GPU	2 x Nvidia RTX 3090 24GB
CPU	Intel(R) Core (TM) i9-10920X CPU @ 3.50GHz
RAM	128 GB
CUDA	v12.7

TABLE 5. DL model hyperparameters.

Batch size	Epoch	Learning rate	Momentum	Optimizer	Weight decay
64	50	1e-3	0.9	SGD	1e-4

models, the hyperparameters in Table 5 were used. The same hyperparameters are adopted for all DL models. The reason for choosing these parameters is to make a fair comparison with the recent work [14].

A. EVALUATION METRICS

The metrics used to evaluate the performance of DL models have a crucial role in determining the accuracy and reliability of the model. Within the scope of this study, accuracy, precision, recall, f1-score, and Cohen's Kappa score metrics were used. Accuracy is the percentage of all correct predictions to total samples, while precision is the percentage of true positives to total positive predictions. Recall is the percentage of true positives to total true positives. The f1-score is the harmonic mean of precision and recall, which helps to assess the overall performance of the model in unbalanced data sets. Cohen's Kappa score is statistical metric that measures the difference between observed and expected accuracy; it is used to understand how much better the model performs than random predictions. The performance metrics use terms such as true positive (TP), false positive (FP), false negative (FN), and true negative (TN) calculated from the complexity matrix. These terms allow the model's predictions to be compared with actual results and are critical for assessing the overall success of the model. The formulas for all these metrics are given in Equations 4-8.

$$\text{Accuracy} = (\text{TP+TN})/(\text{TP+TN+FP+FN}) \quad (4)$$

$$\text{Precision} = \text{TP}/(\text{TP+FP}) \quad (5)$$

$$\text{Recall} = \text{TP}/(\text{TP+FN}) \quad (6)$$

$$\text{F1-score} = 2(\text{Precision x Recall})/(\text{Precision}+\text{Recall}) \quad (7)$$

$$\text{Cohen's Kappa} = (p_0 - p_e)/(1 - p_e) \quad (8)$$

In Eq. 8 p_0 represents the actual positive proportion of the positive class predicted by the model. p_e represents the expected proportion of the positive class if the model makes random predictions.

V. RESULTS

A. CNN RESULTS

Within the scope of the study, CSPNeXt, DenseNet, GhostNet, Inception, MobileNet, ResNeSt, ResNet, and Xception models were tested at different scales. The accuracy, precision, recall, f1-score, and Cohen's kappa score were used to objectively evaluate the classification performance of the models and the results are listed in Table 5.

Table 6 shows that the classification accuracy values of the models vary in a wide range of 78.76% - 91.15%. Among the CSPNeXt variants, CSPNeXt_M performed the best, achieving 87.61% accuracy and 87.69% f1-score. The CSPNeXt_T model performed similarly high results, with a kappa score of 83.91%.

In the DenseNet models, the DenseNet121 and DenseNet201 models achieved high accuracy rates. DenseNet121 was one of the most successful models with 89.38% accuracy, 90.06% precision, and 88.44% f1-score. The DenseNet201 model showed similarly strong performance, while the DenseNet169 model performed relatively poorly.

GhostNet variants generally achieved balanced results, with the GhostNet50 model showing the best overall performance. It performed well, with 88.49% accuracy, 89.61% precision, and 88.73% f1-score.

TABLE 6. Comparative classification results of CNN models based on performance metrics.

Model	Accuracy	Precision	Recall	F1-score	Kappa
CSPNeXt_T	87.61	87.85	87.61	87.29	83.91
CSPNeXt_S	84.95	87.37	84.95	85.33	80.64
CSPNeXt_M	87.61	89.45	87.61	87.69	83.99
CSPNeXt_L	86.72	88.53	86.72	87.02	82.92
CSPNeXt_XL	83.18	84.89	83.18	83.62	78.36
DenseNet121	89.38	90.06	89.38	88.44	86.10
DenseNet169	81.41	83.73	81.41	79.79	75.79
DenseNet201	88.49	88.56	88.49	88.18	85.06
GhostNet50	88.49	89.61	88.49	88.73	85.13
GhostNet100	86.72	88.26	86.72	87.10	82.85
GhostNet130	87.61	86.65	87.61	86.69	83.86
GhostNetV2100	85.84	86.30	85.84	85.25	81.60
GhostNetV2130	87.61	87.15	87.61	87.15	83.88
GhostNetV2160	85.84	89.41	85.84	86.47	81.75
InceptionV3	84.07	83.87	84.07	83.86	79.31
MobileNet	87.61	87.78	87.61	87.64	83.96
ResNeSt50	89.38	88.87	89.38	89.02	86.18
ResNeSt101	**91.15**	**91.10**	**91.15**	**90.87**	**88.50**
ResNet50	78.76	82.94	78.76	79.47	72.64
ResNet50V2	81.41	80.73	81.41	80.34	75.63
ResNet101	83.18	82.57	83.18	82.77	78.09
ResNet101V2	86.72	87.81	86.72	86.99	82.83
ResNet152	83.18	85.36	83.18	80.70	77.95
ResNet152V2	86.72	87.01	86.72	86.74	82.80
Xception	84.95	87.93	84.95	83.73	80.46

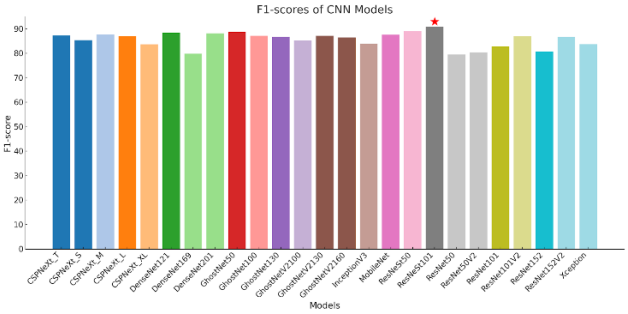

FIGURE 7. F1-score values of CNN models.

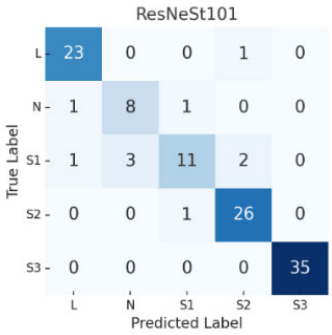

FIGURE 8. Confusion matrix of ResNeSt101 which is best performing among the CNN models (L: laser scars, N: Normal, S1: Stage 1, S2: Stage 2, S3: Stage 3).

In the ResNeSt family, the ResNeSt101 model performed the best in all metrics. This model outperformed other CNN architectures with 91.15% accuracy, 91.10% precision, and 90.87% f1-score.

The overall performance of ResNet variants is lower than other models. However, ResNet101V2 and ResNet152V2 models achieved better results compared to other variants. ResNet101V2 performed the best with 86.72% accuracy, 87.81% precision, and 86.99% f1-score. In contrast, ResNet50 was the model with the lowest performance. MobileNet and Xception models were also used in the study. MobileNet had a balanced performance with 87.61% accuracy and 87.64% f1-score, but it was not among the best models. The Xception model performed moderately well with 84.95% accuracy and 87.93% precision.

Considering the overall results of the present study, ResNeSt101 was the most successful CNN model for ROP diagnosis with 91.15% accuracy, 91.10% precision, and 90.87% f1-score. In addition, DenseNet121, DenseNet201, and GhostNet50 models were also among the strong alternatives with high accuracy. On the other hand, the ResNet50 model showed the poorest performance, exhibiting low values of 78.76% accuracy, 82.94% precision, and 79.47% f1-score. It indicates that the depth and design features of the model are insufficient compared to other CNN architectures. In conclusion, this analysis confirms that the ResNeSt101 model is the best choice for ROP diagnosis, while emphasizing the need to improve the lower performing models.

In Figure 7, the f1-score values of CNN models are visualized in a bar graph. The general trend in the graph reveals that models with more advanced architectures (e.g. ResNeSt, DenseNet, and GhostNet) are more successful in ROP diagnosis. This finding shows that both the architecture and the attentional mechanisms used play a critical role in classification tasks. The confusion matrix where class-based predictions can be seen on the test data of ResNeSt101 is shown in Figure 8.

In the Laser Scars (L) class, 23 of 24 images were correctly classified (TP = 23), 1 was predicted as false negative (FN = 1) and 2 as false positive (FP = 2). In the Normal (N) class, 8 of 10 images were correctly classified (TP = 8), 2 were predicted as false negative (FN = 2) and 3 as false positive (FP = 3). In the Stage 1 (S1) class, 11 of 17 images were correctly classified (TP = 11), 6 were estimated as false negative (FN = 6) and 2 images were estimated as false positive (FP = 2). Since the features in Stage 1 represent the early stages of ROP development, their low discriminative features negatively affected the performance of the model. In Stage 2 (S2) class, 26 of 27 images were correctly classified (TP = 26), 1 was false negative (FN = 1) and 3 images were false positive (FP = 3). All 35 images in the Stage 3 (S3) class were correctly classified (TP = 35), with no misclassifications (FP = 0, FN = 0). The model performed generally well across classes. Although the overall correct classification rate is high in Laser Scars (L), Normal (N), Stage 1 (S1), Stage 2 (S2) and Stage 3 (S3) classes, the number of false negatives and positives observed especially in Stage 1 (S1) and Stage 2 (S2)

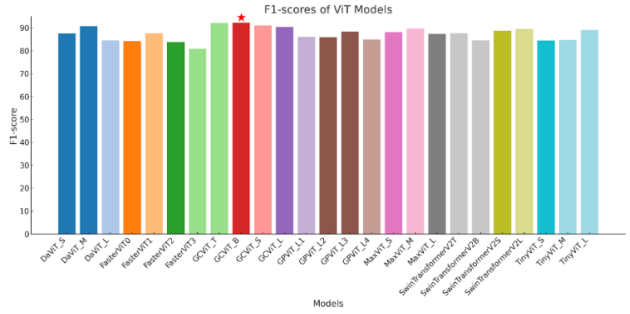

FIGURE 9. F1-score values of ViT models.

classes shows that the discriminative power of the model in these classes needs to be improved.

B. ViT RESULTS

In the study, ViT models such as DaViT, FasterViT, GCViT, GPViT, MaxViT, SwinTransformer, and TinyViT are analyzed comprehensively with their different variants. Classification performance of the ViT models and the results are listed in Table 7.

Among the DaViT models, DaViT_M performed the best. The model achieved 91.15% accuracy, 92.04% precision and 90.70% f1-score. DaViT_S (87.61% accuracy) and DaViT_L (86.72% accuracy) also performed at reasonable levels.

Among the FasterViT models, FasterViT1 performed the best, with 87.61% accuracy and 87.63% f1-score. FasterViT0 and FasterViT2 performed similarly, both achieving 84.07% accuracy.

In the GCViT series, GCViT_T and GCViT_B achieved the highest accuracy of 92.03% among all models. However, the GCViT_B model gave the best results with an f1-score of 92.32%. The GCViT_S and GCViT_L models also performed well with 91.15% accuracy respectively. In general, GPViT models exhibited accuracy values ranging between 84% and 88%. GPViT_L3 was the most successful model in this series with 88.49% accuracy.

In the MaxViT series, MaxViT_B stands out from the other models with an accuracy of 90.26%. SwinTransformer models achieved accuracy rates ranging from 84% to 89%. SwinTransformerV2L was the most successful SwinTransformer variant with 89.38% accuracy. In the TinyViT series, TinyViT_L outperformed the other TinyViT models with 89.38% accuracy.

Figure 9. shows a bar chart displaying the comparative f1-scores of ViT models, which reveals that ViT-based models offer overall higher performance compared to CNN models due to their strong global context extraction. However, these performance differences between different ViT models highlight the critical impact of design and optimization on performance. These results also point to potential areas of improvement for the development of ViT architecture.

Among all ViT models, GCViT_B was the most successful model with 92.03% accuracy, 92.32% f1-score, and 89.72% Cohen's Kappa. DaViT_M and GCViT_T also performed well. On the other hand, the lowest-performing models were FasterViT3 with 83.18% accuracy and FasterViT0 and FasterViT2 with 84.07% accuracy.

As a result, among the ViT models, GCViT_B achieved the highest classification performance and provided an effective solution for ROP diagnosis. The lower-performing FasterViT models need to be improved. Among the CNN models, ResNeSt101 performed the best with 91.15% accuracy, while GCViT_B outperformed the ViT models with 92.03% accuracy. It indicates that ViT models are more robust than CNN in local and global feature extraction. The complexity matrix of the GCViT_B model is shown in Figure 10.

In the Laser Scars (L) class, 23 of 24 images were correctly classified (TP = 23) and 1 was evaluated as false negative (FN = 1). There were no false positives in this class (FP = 0). In the Normal (N) class, 9 of 10 images were correctly

TABLE 7. Comparative classification results of ViT models based on performance metrics.

Model	Accuracy	Precision	Recall	F1-score	Kappa
DaViT_S	87.61	87.88	87.61	87.59	83.95
DaViT_M	91.15	92.04	91.15	90.70	88.47
DaViT_L	86.72	85.06	86.72	84.61	82.54
FasterViT0	84.07	86.16	84.07	84.28	79.47
FasterViT1	87.61	87.73	87.61	87.63	83.95
FasterViT2	84.07	87.65	84.07	83.90	79.44
FasterViT3	83.18	83.34	83.18	80.98	77.84
GCViT_T	92.03	92.36	92.03	92.15	89.72
GCViT_B	**92.03**	**93.39**	**92.03**	**92.32**	**89.72**
GCViT_S	91.15	91.10	91.15	91.08	88.52
GCViT_L	91.15	92.39	91.15	90.44	88.45
GPViT_L1	86.72	87.17	86.72	86.02	82.69
GPViT_L2	86.72	86.11	86.72	85.95	82.65
GPViT_L3	88.49	88.66	88.49	88.41	85.08
GPViT_L4	84.95	86.34	84.95	84.91	80.63
MaxViT_S	88.49	88.54	88.49	88.21	85.06
MaxViT_M	90.26	90.13	90.26	89.77	87.32
MaxViT_L	87.61	88.61	87.61	87.45	83.91
SwinTV2T	87.61	87.90	87.61	87.70	83.97
SwinTV2B	84.95	85.71	84.95	84.59	80.50
SwinTV2S	89.38	89.03	89.38	88.78	86.13
SwinTV2L	89.38	90.85	89.38	89.64	86.28
TinyViT_S	84.95	84.28	84.95	84.46	80.35
TinyViT_M	84.95	84.97	84.95	84.75	80.47
TinyViT_L	89.38	88.89	89.38	89.09	86.19

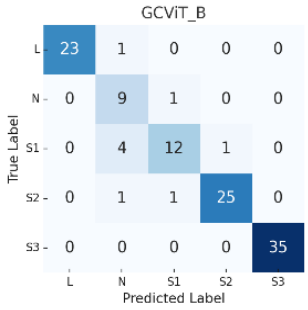

FIGURE 10. Confusion matrix of GCViT_B model which is best performing among the ViT models (L: laser scars, N: Normal, S1: Stage 1, S2: Stage 2, S3: Stage 3).

classified (TP = 9), 1 was false negative (FN = 1) and 6 were false positive (FP = 6). In Stage 1 (S1) class, 12 out of 17 images were correctly classified (TP = 12), 5 were false negative (FN = 5) and 2 were false positive (FP = 2). In Stage 2 (S2) class, 25 of 27 images were correctly classified (TP = 25), 2 were false negative (FN = 2) and 1 was false positive (FP = 1). All 35 images in the Stage 3 (S3) class were correctly classified (TP = 35). There were no misclassifications in this class (FP = 0, FN = 0). The ViT model provided significant improvements compared to the CNN model. In the Laser Scars (L) class, the number of false positives decreased from 2 to 0. In the Normal (N) class, the number of true positives increased from 8 to 9, while the number of false negatives decreased from 2 to 1. In the Stage 1 (S1) class, the number of true positives increased from 11 to 12. In the Stage 2 (S2) class, the number of false negatives decreased from 3 to 2 and the number of false positives decreased from 3 to 1, providing a more accurate classification. Overall, the ViT model was very effective in reducing false negatives and provided a better discrimination capacity between classes.

C. ROPGCViT RESULTS

The GCViT_B model was the most successful model among ViT models with an accuracy of 92.03%. Considering the GCViT_B model as the baseline model, the proposed enhancements were implemented. First, the SE block was added to the stem module and experiments were also performed with Efficient Channel Attention (ECA) and Convolutional Block Attention Module (CBAM) attention mechanisms to compare the effect of the SE block. Table 8 shows the results of all experiments using the attention modules. When Table 8 is analyzed, it is seen that attention mechanisms have positive impacts on the performance of the model. In computer-aided systems in healthcare, every improvement, even if minimal, is critical. ECA and CBAM showed almost the same performance with 92% accuracy and recall rates. The SE model yielded 93.80% accuracy, 94.15% precision, 93.80% recall, 93.80% recall, 93.67% f1-score and 91.96% Cohen's kappa score. Comparing the baseline model and the SE-enhanced model, there is an increase of 1.92% for accuracy, 0.81% for precision, 1.92% for recall, 1.46% for f1-score and 2.50% for kappa score. The fact that the SE module outperforms CBAM and ECA can be explained by two reasons: the spatial attention added by CBAM does not add any additional contribution to the specificity of the dataset or problem structure, but rather makes the model more complex; and ECA, although it offers low overhead with its lightweight channel attention approach, does not capture in-channel interactions as deeply as SE.

Complexity matrices visualizing the class-based predictions of attention mechanisms on the test data are shown in Figure 11. CBAM, ECA and SE models showed high accuracy in all classes. In the Laser Scars (L) class, CBAM 23, ECA and SE provided 24 correct classifications (TP). In the Normal (N) class, the SE model performed the best with 9 correct classifications (TP), while the other models achieved 8 correct classifications. In Stage 1 (S1) class, CBAM provided 14, ECA 13 and SE 12 correct classifications. In Stage 2 (S2) class, the SE model achieved the best result with 26 correct classifications. In Stage 3 (S3) class, all models achieved excellent results with 35 correct classifications. In general, the SE model stands out with fewer false negatives (FN) and false positives (FP). The SE model performed better than the GCViT_B model, especially in Stage 1 (S1) and Stage 2 (S2) classes. While it reduced the number of false negatives (FN) in S1 class, it provided more correct classifications (TP) in S2 class. In the other classes, both models gave similar results, but the SE model provided more accurate and reliable classification in S1 and S2.

The final modification to the GCViT_B architecture is the replacement of the MLP blocks of the model enhanced with stem block SE with RMLP blocks and this final model is named ROPGCViT. The residual connections in the RMLP blocks alleviate the vanishing gradient problem by strengthening the weakened gradients in the backpropagation process and thus the training process of the model becomes more stable. The test results in this context are presented in Table 8. ROPGCViT achieved 94.69% accuracy, 94.84% precision, 94.69% recall, 94.60% f1-score and 93.10% Cohen's kappa score. When GCViT_B (SE) and ROPGCViT (SE+RMLP) models are compared, an increase of 0.89% for accuracy, 0.69% for precision, 0.89% for recall, 0.93% for f1-score and 1.14% for kappa score is achieved.

TABLE 8. Comparative classification results of GCViT models with improved stem blocks using various attention mechanisms.

Attention	Performance measurement metrics (%)				
	Accuracy	Precision	Recall	F1-score	Kappa
GCViT_B	92.03	93.39	92.03	92.32	89.72
ECA	92.92	92.74	92.92	92.76	90.79
CBAM	92.92	92.91	92.92	92.89	90.80
SE	**93.80**	**94.15**	**93.80**	**93.67**	**91.96**

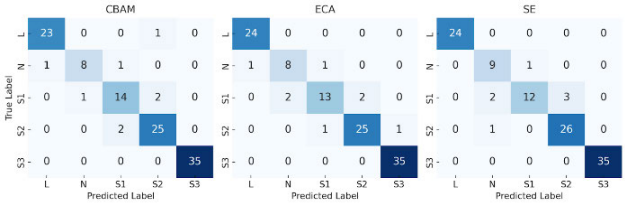

FIGURE 11. Confusion matrices of the GCViT model where the stem block is improved with different attention mechanisms (L: laser scars, N: Normal, S1: Stage 1, S2: Stage 2, S3: Stage 3).

TABLE 9. The classification performance of SE-Enhanced Stem Block and RMLP layer on the GCViT architecture.

Technique	Performance measurement metrics (%)				
	Accuracy	Precision	Recall	F1-score	Kappa
GCViT_B (SE)	93.80	94.15	93.80	93.67	91.96
ROPGCViT (SE+RMLP)	94.69	94.84	94.69	94.60	93.10

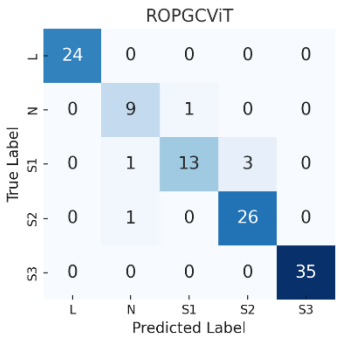

FIGURE 12. Confusion matrix of ROPGCViT (L: Laser scars, N: Normal, S1: Stage 1, S2: Stage 2, S3: Stage 3).

TABLE 10. Comparison of classification performances of proposed models.

Comparison Models	Performance measurement metrics (%)				
	Accuracy	Precision	Recall	F1–score	Kappa
ROPGCViT vs ResNeSt101	+3.54 (↑)	+3.74 (↑)	+3.54 (↑)	+3.73 (↑)	+4.60 (↑)
ROPGCViT vs GCViT_B	+2.89 (↑)	+1.55 (↑)	+2.89 (↑)	+2.47 (↑)	+3.77 (↑)
ROPGCViT vs GCViT_B (SE)	+0.95 (↑)	+0.73 (↑)	+0.95 (↑)	+0.99 (↑)	+1.24 (↑)

The complexity matrix of the prediction results of the ROPGCViT model on the test data is shown in Figure 12. In class Laser Scars (L), all 24 samples were correctly predicted (TP = 24, FP = 0, FN 0). In class Normal (N), 9 samples were correctly predicted (TP = 9), while 1 sample was incorrectly classified as S1 (FN = 1). In the Stage 1(S1) class, 13 correct predictions were made (TP = 13) and 4 instances (FN = 4) were misclassified to N and S2 classes. In class Stage 2(S2), 26 samples were correctly classified (TP = 26), but 3 samples from S1 were incorrectly predicted as S2 (FP = 3). In class Stage(S3), all 35 samples were correctly predicted (TP = 35, FP = 0, FN = 0). Overall, the model performed well with high accuracy in all classes. Compared to ROPGCViT, GCViT_B and GCViT_B (SE) models, it was found to be the most successful method with an increase in TP rates and a decrease in FP rates. A detailed comparison of the classification performance of the models is given in Table 10.

The proposed ROPGCViT model showed higher classification success compared to both ResNeSt101, the most successful CNN-based architecture, and ViT-based GCViT_B and GCViT_B (SE) models. According to the results in Table 10, ROPGCViT showed an increase of +3.54, +3.74, +3.54, +3.73 and +4.60 points in accuracy, precision, recall, f1- score and kappa metrics, respectively, compared to ResNeSt101. In addition, ROPGCViT achieved +2.89, +1.55, +2.89, +2.89, +2.47-, and +3.77-points higher results in the same metrics compared to GCViT_B, and provided a performance improvement of +0.95, +0.73, +0.95, +0.95, +0.99, and +1.24 points compared to the GCViT_B (SE) model. These data show that the ROPGCViT approach significantly improves classification performance.

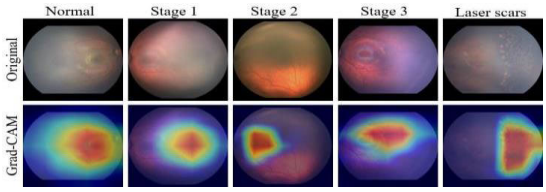

FIGURE 13. Heat maps visualizing the regions considered by the ROPGCViT model on randomly selected fundus images from the test data using the Grad-CAM method.

D. EXPLAINABLE AI RESULTS

Although DL models are highly successful in medical imaging, they are seen as a "black box" because they cannot explain the decision mechanisms. Such a situation can affect confidence in the diagnosis of eye diseases and the speed of adoption in clinical applications. In this study, the Grad-CAM method was used to see which regions the model considers when classifying fundus images. Grad-CAM helps specialists better understand the decision process and detect errors by visualizing the areas that the model pays attention to as a heatmap. Thus, specialists can evaluate not only the result of the model, but also how it arrived at that result. Figure 13 shows the Grad-CAM outputs of ROPGCViT on randomly selected images from the test data.

In normal fundus images, Grad-CAM output is usually concentrated around the macula and optic disc, as these are the main reference points in the retinal anatomy where healthy vascular and nerve layers are present. At Stage 1, Grad-CAM attention is focused on the borders of the retinal vasculature, as microvascular changes such as dilation and tortuosity of retinal vessels are observed at this stage. At Stage 2, attention is focused on the peripheral regions where vascular changes are more pronounced. In Stage 3, the focus is on areas of fibro-vascular proliferation, as at this stage pathological deposits become more pronounced and abnormalities on the retinal surface increase. In the case of laser scars, attention is focused on these scar areas because the model perceives these areas as pathological findings and highlights them in the decision mechanism.

E. LITERATURE COMPARISON

In order to make a fair comparison, we compared the performance of ROPGCViT with the results of studies in the literature that use the same dataset. The comparison of classification performances is presented in Table 11. There is only one study in the literature on the dataset used in this study. As mentioned in Table 1 in the introduction, Zhao et al. [14] tried different DL models and achieved the highest performance with the ResNet50 model. The authors achieved 87.61% accuracy, 82.64% precision, 83.31% recall, 82.81% f1-score and 83.97% kappa score. On the other hand, ROPGCViT outperforms existing methods in all metrics on this dataset. The results presented in Table 11 indicates substantial improvements over the baseline model, with notable increases in accuracy (+7.08%), precision (+12.20%), recall

TABLE 11. Comparative classification performance results between literature study and ROPGCViT model.

Ref.	Performance measurement metrics (%)				
	Accuracy	Precision	Recall	F1-score	Kappa
Zhao et al. [14]	87.61	82.64	83.31	82.81	83.97
Proposed model	94.69	94.84	94.69	94.60	93.10

(+11.38%), f1-score (+11.79%), and Kappa (+9.13%). The experimental findings prove that the ROPGCViT model offers enhanced classification capability, potentially due to its advanced architectural design and improved local and global feature extraction mechanisms.

VI. DISCUSSION

ROP is a serious retinal disease in premature babies that can lead to permanent vision loss if left untreated. Diagnosis with traditional methods is time-consuming and may contain inconsistencies as it relies on expert knowledge. In this study, a total of 50 DL models, 25 CNN and 25 ViT models were tested for ROP diagnosis. ResNeSt101, the most successful model among CNN models, achieved 91.15% accuracy, 91.10% precision, 91.15% recall, and 90.87% f1-score. Among the ViT models, GCViT_B was the most successful model with 92.03% accuracy, 92.36% precision, 92.03% recall and 92.32% f1-score. Furthermore, the ROPGCViT model is proposed to improve both accuracy and overall performance.

First, the SE block was added to the stem block of the GCViT_B model, which improved the accuracy of model to 93.80%, precision to 94.15%, recall to 93.80%, and f1-score to 93.67%. It is observed that the SE block improves the class separation ability of the model by rescaling the channel-based features. The performance of the SE block is also higher than that of CBAM and ECA. The CBAM-enhanced GCViT_B model achieved 92.92% accuracy, 92.91% precision, 92.92% recall, and 92.89% f1-score. The model improved with ECA achieved 92.92% accuracy, 92.74% precision, 92.92% recall, and 92.76% f1-score. These results show that the SE block outperforms both mechanisms. In the final stage, the final ROPGCViT model was created by adding RMLP layers to the GCViT_B model. The ROPGCViT model outperformed all previous models with 94.69% accuracy, 94.84% precision, 94.69% recall, and 94.60% f1-score. This success is based on the channel-based rescaling capacity of the SE blocks and the ability of the RMLP structure to improve the training dynamics. Moreover, a detailed analysis of the performance of the model showed significant improvements in TP, TN, FP, and FN rates. The ROPGCViT model significantly reduced the FN and FP rates, while maintaining a balanced discrimination between all classes. As a result, the proposed ROPGCViT model provided a significant superiority over both the 50 DL model and the methods in the literature. In addition, the inclusion of Grad-CAM integration in the analysis process increased the explainability of the ROPGCViT model, visualizing regions of interest in fundus images. This not only increased confidence in the model's decision-making process, but also provided important insights for clinical experts, highlighting the model's areas of focus for classification, further supporting its potential for real-world clinical applications.

The limitations of this study are; first, segmentation techniques have not been applied to analyze the eye disease in a detailed way. Transformer-based deep model can be used for segmentation task for detailed analysis of ROP. Also, the real ROP diagnosis includes aggressive posterior ROP, acute ROP, plus disease ROP, pre-threshold ROP, and ROP Zones, however, they have not been addressed. Future work will address these limitations.

VII. CONCLUSION

In this study, ROPGCViT, a DL-based model for ROP diagnosis, is proposed and achieved superior achievements in terms of both performance and explainability compared to existing methods. It is developed based on the GCViT architecture and enhanced with SE blocks and MLP configurations. These improvements allow for effective learning of both local and global features, improving the classification and overall performance of the model. In extensive experiments, the ROPGCViT model was compared with a total of 50 different DL models based on both CNN and ViT and outperformed them in all performance metrics. ROPGCViT also outperformed in accuracy, precision, recall, f1-score, and Cohen's kappa values in comparisons with studies conducted on the same dataset in the literature. In particular, the channel-based rescaling capacity of the SE blocks and the ability of the RMLP structure to improve model training dynamics played a critical role in this success. The explainability of the model was ensured by the Grad-CAM method. Grad-CAM visualized the regions that the model takes into account during classification, allowing experts to understand the decision mechanism and increasing the reliability of the model in clinical applications. The resulting heat maps showed how the model distinguished between healthy and pathological regions in retinal images.

In conclusion, the ROPGCViT model offers an innovative solution that provides high accuracy, reliability and clarity in ROP diagnosis. The success of the model is not only a significant contribution to the existing literature, but also a valuable step in developing a tool that can be used in clinical practice.

DATA AVAILABILITY

The dataset used in this study can be found at
https://doi.org/10.6084/m9.figshare.25514449

CODE AVAILABILITY

The ROPGCViT model code can be accessed from the following link;
https://github.com/ymyurdakul/ROPGCViT/tree/main

REFERENCES

[1] R. L. Gregory, *Eye and Brain: The Psychology of Seeing*. Torrossa, Online digital Bookstore, 2015.

[2] A. S. Tsai, H.-D. Chou, X. C. Ling, T. Al-Khaled, N. Valikodath, E. Cole, V. L. Yap, M. F. Chiang, R. V. P. Chan, and W.-C. Wu, "Assessment and management of retinopathy of prematurity in the era of anti-vascular endothelial growth factor (VEGF)," *Prog. Retinal Eye Res.*, vol. 88, May 2022, Art. no. 101018.

[3] C. Dai, J. Xiao, C. Wang, W. Li, and G. Su, "Neurovascular abnormalities in retinopathy of prematurity and emerging therapies," *J. Mol. Med.*, vol. 100, no. 6, pp. 817–828, Jun. 2022.

[4] V. Raja 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," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104883.

[5] S. Jalali, R. Azad, H. Trehan, M. Dogra, L. Gopal, and V. Narendran, "Technical aspects of laser treatment for acute retinopathy of prematurity under topical anesthesia," *Indian J. Ophthalmology*, vol. 58, no. 6, p. 509, 2010.

[6] K. Uyar, M. Yurdakul, and Ş. Taşdemir, "Abc-based weighted voting deep ensemble learning model for multiple eye disease detection," *Biomed. Signal Process. Control*, vol. 96, Oct. 2024, Art. no. 106617.

[7] M. Yurdakul, K. Uyar, and S. Tasdemir, "MaxGlaViT: A novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images," 2025, *arXiv:2502.17154*.

[8] I. Pacal, "MaxCerVixT: A novel lightweight vision transformer-based approach for precise cervical cancer detection," *Knowl.-Based Syst.*, vol. 289, Apr. 2024, Art. no. 111482.

[9] G. Sun, H. Shu, F. Shao, T. Racharak, W. Kong, Y. Pan, J. Dong, S. Wang, L.-M. Nguyen, and J. Xin, "FKD-med: Privacy-aware, communication-optimized medical image segmentation via federated learning and model lightweighting through knowledge distillation," *IEEE Access*, vol. 12, pp. 33687–33704, 2024.

[10] M. T. Islam, S. T. Mashfu, A. Faisal, S. C. Siam, I. T. Naheen, and R. Khan, "Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation," *IEEE Access*, vol. 10, pp. 2828–2841, 2022.

[11] S. Gayathri, V. P. Gopi, and P. Palanisamy, "A lightweight CNN for diabetic retinopathy classification from fundus images," *Biomed. Signal Process. Control*, vol. 62, Sep. 2020, Art. no. 102115.

[12] W. Wang, X. Li, Z. Xu, W. Yu, J. Zhao, D. Ding, and Y. Chen, "Learning two-stream CNN for multi-modal age-related macular degeneration categorization," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4111–4122, Aug. 2022.

[13] T. Pratap and P. Kokil, "Computer-aided diagnosis of cataract using deep transfer learning," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101533.

[14] X. Zhao, S. Chen, S. Zhang, Y. Liu, Y. Hu, D. Yuan, L. Xie, X. Luo, M. Zheng, R. Tian, Y. Chen, T. Tan, Z. Yu, Y. Sun, Z. Wu, and G. Zhang, "A fundus image dataset for intelligent retinopathy of prematurity system," *Sci. Data*, vol. 11, no. 1, p. 543, May 2024.

[15] A. Ding, Q. Chen, Y. Cao, and B. Liu, "Retinopathy of prematurity stage diagnosis using object segmentation and convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.

[16] Z. Tan, S. Simkin, C. Lai, and S. Dai, "Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease," *Transl. Vis. Sci. Technol.*, vol. 8, no. 6, p. 23, Dec. 2019.

[17] J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. V. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, J. Kalpathy-Cramer, and M. F. Chiang, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," *JAMA Ophthalmol.*, vol. 136, no. 7, p. 803, Jul. 2018.

[18] Y. Tong, W. Lu, Q.-Q. Deng, C. Chen, and Y. Shen, "Automated identification of retinopathy of prematurity by image-based deep learning," *Eye Vis.*, vol. 7, no. 1, pp. 1–12, Dec. 2020.

[19] R. Zhang, J. Zhao, H. Xie, T. Wang, G. Chen, G. Zhang, and B. Lei, "Automatic diagnosis for aggressive posterior retinopathy of prematurity via deep attentive convolutional neural network," *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115843.

[20] N. Salih, M. Ksantini, N. Hussein, D. Ben Halima, A. Abdul Razzaq, and S. Ahmed, "Prediction of ROP zones using deep learning algorithms and voting classifier technique," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 86, May 2023.

[21] P. Huang, Y. Xie, R. Wu, Q. Lin, N. Cai, H. Chen, and S. Feng, "ROPRNet: Deep learning-assisted recurrence prediction for retinopathy of prematurity," *Biomed. Signal Process. Control*, vol. 100, Feb. 2025, Art. no. 107135.

[22] 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.

[23] O. Attallah, "DIAROP: Automated deep learning-based diagnostic tool for retinopathy of prematurity," *Diagnostics*, vol. 11, no. 11, p. 2034, Nov. 2021.

[24] R. Agrawal, S. Kulkarni, R. Walambe, M. Deshpande, and K. Kotecha, "Deep dive in retinal fundus image segmentation using deep learning for retinopathy of prematurity," *Multimedia Tools Appl.*, vol. 81, no. 8, pp. 11441–11460, Mar. 2022.

[25] W. Feng, Q. Huang, T. Ma, L. Ju, Z. Ge, Y. Chen, and P. Zhao, "Development and validation of a semi-supervised deep learning model for automatic retinopathy of prematurity staging," *IScience*, vol. 27, no. 1, Jan. 2024, Art. no. 108516.

[26] W. Yang, H. Zhou, Y. Zhang, L. Sun, L. Huang, S. Li, X. Luo, Y. Jin, W. Sun, W. Yan, J. Li, J. Deng, Z. Xie, Y. He, and X. Ding, "An interpretable system for screening the severity level of retinopathy in premature infants using deep learning," *Bioengineering*, vol. 11, no. 8, p. 792, Aug. 2024.

[27] O. Attallah, "GabROP: Gabor wavelets-based CAD for retinopathy of prematurity diagnosis via convolutional neural networks," *Diagnostics*, vol. 13, no. 2, p. 171, Jan. 2023.

[28] B. Ebrahimi, D. Le, M. Abtahi, A. K. Dadzie, A. Rossi, M. Rahimi, T. Son, S. Ostmo, J. P. Campbell, R. V. Paul Chan, and X. Yao, "Assessing spectral effectiveness in color fundus photography for deep learning classification of retinopathy of prematurity," *J. Biomed. Opt.*, vol. 29, no. 7, Jun. 2024, Art. no. 076001.

[29] K. M. Jemshi, G. Sreelekha, P. S. Sathidevi, P. Mohanachandran, and A. Vinekar, "Plus disease classification in retinopathy of prematurity using transform based features," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 861–891, Jan. 2024.

[30] D. Wang, W. Qiao, W. Guo, and Y. Cai, "Applying novel self-supervised learning for early detection of retinopathy of prematurity," *Electron. Lett.*, vol. 60, no. 14, Jun. 2024, Art. no. e13267.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[33] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[36] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.

[37] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1577–1586.

[38] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2735–2745.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[40] A. Hatamizadeh, H. Yin, J. Kautz, and P. Molchanov, "Global context vision transformers," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2022, pp. 12633–12646.

MUSTAFA YURDAKUL was born in Kayseri, in 1997. He received the bachelor's degree in computer engineering and the master's degree from SelÃ§uk University, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree. In 2022, he started his academic career as a Research Assistant with the Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Kırıkkale University. He continues his research and studies in the fields of computer vision, artificial intelligence, and signal processing. He has many articles published in national and international journals and papers presented at conferences.

KÜBRA UYAR was born in Antalya, in 1991. She received the master's and Ph.D. degrees from the Institute of Computer Engineering, Selçuk University, in 2022. In 2014, she started her academic career as a Research Assistant with Selçuk University. She continues her research and studies at Alanya Alaaddin Keykubat University, since 2023, in the fields of computer vision and artificial intelligence. She has many articles published in national and international journals and papers presented at conferences.

ŞAKİR TAŞDEMİR was born in Sinop, in 1971. He received the bachelor's degree from Gazi University, in 1994, and the master's and Ph.D. degrees from the Institute of Science and Technology, Selçuk University. In 2013, he received an Associate Professor title in the field of computer-information sciences and engineering. He started his academic career in various institutions affiliated to the Ministry of National Education and held administrative positions, such as the Directorate, the Department head, and the Department Head in universities. He was also a member of various commissions and boards of directors in higher education institutions. He has published more than 80 national and international journal articles, more than 100 articles, more than ten books or book chapters, and received more than 1000 citations. He also has patents, projects, and scientific awards. He has graduated 11 master's and five Ph.D. students until now. His research interests include artificial intelligence, image processing, mobile programming, fuzzy logic, and virtual/augmented reality.

İRFAN ATABAŞ was born in Kırıkkale, in 1978. He received the bachelor's degree from the Faculty of Engineering, Department of Computer Engineering, Erciyes University, in 2001, and the master's degree in mechanical engineering and the Ph.D. degree in computer engineering from the Institute of Science and Technology, Kırıkkale University, in 2006 and 2020, respectively. In 2003, he started his academic career as a Research Assistant with the Department of Computer Engineering, Kırıkkale University, and in 2022, he received the title of a Dr. Faculty Member. He continues his research and studies in the fields of computer vision, hardware design, and signal processing.