Automating ROP Diagnosis and Severity with Deep Learning

Muhammad Mansoor Alam
School of Science and Engineering
Habib University
Karachi, Pakistan
ma08322@st.habib.edu.pk

Ahsan Siddiqui
School of Science and Engineering
Habib University
Karachi, Pakistan
as08155@st.habib.edu.pk

Zohaib Aslam
School of Science and Engineering
Habib University
Karachi, Pakistan
za08134@st.habib.edu.pk

***Abstract*—Retinopathy of Prematurity (ROP) is a condition affecting premature infants, which can lead to blindness if not detected and treated early. However, timely intervention remains a challenge due to a global shortage of trained specialists, particularly in resource-limited settings. This study explores a deep learning-based approach to detect and classify the severity of ROP from retinal images. We evaluated five models: a custom Convolutional Neural Network (CNN), ResNet, EfficientNet, Inception, and VGG. Among these, the custom CNN achieved the highest performance with a test accuracy of 98.48% and a test loss of 0.06 on a dataset of 6,004 retinal images. EfficientNet followed with a test accuracy of 96.07% and a loss of 0.13, while ResNet achieved an accuracy of 95.35% with a loss of 0.14. Inception and VGG performed comparatively lower, with test accuracies of 91.83% and 88.86%, respectively. The proposed deep learning-based approach demonstrates strong potential to enhance diagnostic accuracy and improve accessibility, aiding healthcare professionals in reducing the burden of ROP screening and improving patient outcomes.**

***Index Terms*—Retinopathy of Prematurity, Deep Learning, Convolutional Neural Networks, CNN, ResNet, EfficientNet, Medical Imaging, Machine Learning, ROP Classification**

I. Introduction

Retinopathy of Prematurity (ROP) is a serious condition that affects the eyes of premature infants and can lead to permanent blindness if left untreated. It is characterized by abnormal vascular development in the retina, which can progress through various stages, eventually causing retinal detachment in severe cases. ROP remains a major cause of childhood blindness worldwide, with its incidence closely tied to factors such as low birth weight, premature birth, oxygen therapy, and other medical interventions.

ROP primarily affects infants born before 32 weeks of gestation and is categorized into five stages, ranging from mild (Stage 1) to severe (Stage 5). Early detection is critical as the progression of the disease can often be halted or reversed through timely interventions. Screening programs typically involve retinal fundus imaging, which is reviewed by ophthalmologists to determine the severity of the condition. Treatments such as laser surgery or anti-VEGF (vascular endothelial growth factor) therapy have shown significant success in reducing the risk of permanent vision loss. However, the increasing number of ROP cases has far outpaced the availability of qualified specialists, particularly in remote or underserved areas.

The shortage of accessible ROP specialists highlights the need for alternative diagnostic methods, such as telemedicine and computer-aided diagnosis (CAD). These technologies hold great potential to improve the early detection of ROP, especially in regions with limited medical resources. In recent years, artificial intelligence has emerged as a transformative tool in medical diagnostics, including for conditions like ROP. Using deep learning techniques and image analysis, AI models can assist in identifying ROP and evaluating its severity based on retinal fundus images. This not only reduces the burden on medical professionals but also enhances the accuracy and efficiency of diagnosis, ultimately improving patient outcomes.

II. Research Problem

The primary objective of this study is to develop and validate a deep learning-based system capable of predicting both the presence and severity of Retinopathy of Prematurity (ROP) from retinal images. The research problem can be formulated as:

> *"Given a retinal image of an infant, identify the presence and severity of Retinopathy of Prematurity using a deep learning model."*

The proposed deep learning model will process retinal images as input and output a classification of ROP severity based on the vascular severity score (VSS). The VSS serves as a reference to determine whether the infant's condition requires medical intervention, where a zero score indicates that there is no ROP present. Alongside designing custom neural networks, this study will also evaluate the performance of existing deep learning architectures in addressing this classification task. Classification, a type of supervised learning, involves input data being associated with corresponding objectives. Models can adapt and learn to perform classification tasks, enabling a better understanding of diverse data. Deep learning-based classification models have wide applications in various fields, such as credit approval, disease diagnosis, and target marketing. In the

context of this study, our model will analyze retinal images and classify them into severity classes of ROP, providing a structured diagnosis based on the VSS score.

The motivation behind this research stems from the substantial global burden of ROP, which demands frequent and expert screening, particularly for premature infants. Current manual screening methods are often resource-intensive, posing challenges in low-resource or rural areas where specialists are scarce. Therefore, an automated, AI-based diagnostic tool has the potential to significantly alleviate the strain on healthcare systems, facilitate early diagnosis, and ultimately enhance patient outcomes. Additionally, while deep learning has been widely applied to other eye diseases, its use in predicting ROP severity, particularly in infants, remains underexplored—highlighting a critical gap in the existing body of research.

III. Literature Review

The papers reviewed for this study focus on deep learning algorithms for predicting and classifying Retinopathy of Prematurity (ROP). These studies address various aspects of ROP detection, including disease occurrence and severity classification, using different neural network architectures. The primary goal of these models is to support early diagnosis and intervention, with many prioritizing high sensitivity over specificity to minimize missed ROP cases. Below is a summary of the key findings from the reviewed literature.

One study [1] explored the use of deep learning to predict both the occurrence and severity of ROP. The dataset consisted of 7,033 retinal images from 725 infants for training and 763 images from 90 infants for validation. Retinal images were preprocessed, and deep features were extracted using a pretrained ResNet-50 model. These features were combined with clinical data such as gestational age and birth weight, creating a 558-dimensional feature vector. Two models were developed: OC-Net for predicting ROP occurrence and SE-Net for severity classification (mild vs. severe). A voting scheme based on thresholds was employed for final predictions. The results showed that OC-Net achieved an AUC of 0.90 with an accuracy of 52.8%, sensitivity of 100%, and specificity of 37.8%. SE-Net demonstrated better performance in severity prediction with an AUC of 0.87, accuracy of 68.0%, sensitivity of 100%, and specificity of 46.6%. While the models exhibited high sensitivity, the lower specificity resulted in more false positives, ensuring early detection of ROP cases.

Another research [2] focused on diagnosing plus disease, a severe form of ROP, using the ROP.AI algorithm. The algorithm was trained on 6,974 fundal images, with preprocessing steps including cropping and image augmentation. The Inception-v3 CNN was employed for classification. Internal validation* demonstrated high performance, with a sensitivity of 96.6%, specificity of 98.0%, and an accuracy of 97.3%. External validation, however, showed a slight drop in performance, with a sensitivity of 93.9%, specificity of 80.7%, and an AUROC of 0.977. Despite concerns about generalizability due to reliance on images from a single institution, the model's optimization for high sensitivity and negative predictive value (NPV) made it valuable for screening applications.

A third study [3] utilized a VGG19 model for ROP detection and severity classification. Transfer learning was applied to a dataset of 6,500 images, augmented to 18,808 images through techniques such as rotation, shifting, and zooming. The VGG19 model achieved remarkable accuracy, with 98.8% in severity prediction, 100% sensitivity, and 98.41% specificity. Another model, VGG16, also performed well, albeit with a slightly lower accuracy of 96.5%. This study emphasized the importance of accurate and timely diagnosis to prevent blindness, identifying VGG19 as the superior model for ROP severity prediction.

Additional research [4] investigated various AI models, including ResNet-50, CNNs, and DNNs, for ROP detection. In this approach, a deep neural network (DNN) implemented two models—OC-Net for occurrence and SE-Net for severity classification—while other models like GR-Net focused on grading disease severity. These models were evaluated using ROC curves, demonstrating their effectiveness in medical imaging applications. The study highlighted the extension of these models to other medical conditions, such as tumor detection and skin lesion analysis.

Moreover another study [5] combined support vector machine (SVM) techniques with deep learning to detect ROP. Clinical risk factors, such as low gestational age and birth weight, were integrated with image features like vascular tortuosity for ROP staging. The SVM model achieved a high accuracy of 95%, while the i-ROP DL system assigned a Vascular Severity Score (VSS) on a scale of 1 to 9. This score effectively predicted treatment-requiring ROP with an AUC of 0.95. The study concluded that combining SVM with deep learning improved detection rates and could be integrated into telemedicine programs for remote diagnosis.

In summary, these studies demonstrate the potential of deep learning models in ROP prediction and classification. The VGG19 model emerged as the best-performing architecture, achieving an accuracy of 98.8%. Most models prioritized sensitivity to ensure early detection of ROP, although this often came at the expense of specificity, resulting in more false positives. Incorporating clinical data alongside image features was a common approach to enhance predictive accuracy. These findings underscore the promise of deep learning in early screening and intervention for ROP.

** Internal validation tests the model on unseen data from the same dataset, while external validation evaluates its performance on independent datasets from different sources.*

TABLE I
SUMMARY OF STUDIES

Study	Model(s) Used	Dataset	Results/Accuracies
Study 1: Deep Learning for ROP Prediction	ResNet-50 (feature extraction), OC-Net (ROP occurrence), SE-Net (ROP severity)	7,033 retinal images (training), 763 images (validation)	OC-Net: AUC 0.90, accuracy 52.8%, sensitivity 100%, specificity 37.8%. SE-Net: AUC 0.87, accuracy 68.0%, sensitivity 100%, specificity 46.6%.
Study 2: ROP.AI for Plus Disease Diagnosis	Inception-v3 CNN	6,974 fundal images (training), 90 images (external validation)	Internal Validation: Sensitivity 96.6%, specificity 98.0%, accuracy 97.3%. External Validation: Sensitivity 93.9%, specificity 80.7%, AUROC 0.977.
Study 3: VGG19 for ROP Detection	VGG19, VGG16	6,500 images (augmented to 18,808 images)	VGG19: Accuracy 98.8%, sensitivity 100%, specificity 98.41%. VGG16: Accuracy 96.5%.
Study 4: AI Models for ROP Detection	ResNet-50, CNNs, DNNs (OC-Net and SE-Net)	Various retinal images	OC-Net and SE-Net for ROP occurrence and severity. ROC curves used for evaluating model performance in medical imaging.
Study 5: SVM and Deep Learning in ROP Detection	SVM, DeepROP System, i-ROP DL system	Retinal images, clinical data (gestational age, birth weight)	SVM: 95% accuracy. DeepROP: Sensitivity 96.62%, specificity 99.32%. i-ROP DL: AUC 0.95 (Vascular Severity Score).

IV. MATERIALS AND METHODOLOGY

A. Dataset

The dataset used in this study for training and testing the deep learning model for Retinopathy of Prematurity (ROP) detection was sourced from Kaggle [6]. This dataset contains retinal images of premature infants, accompanied by anonymized patient metadata. It is specifically curated to aid in the development of algorithms for diagnosing retinal diseases in premature infants. Retinal images were evaluated by trained ophthalmologists to ensure quality, and cases were classified based on clinical assessment following standard guidelines.

1) Summary: The dataset comprises 6,004 retinal images collected from 188 newborns, primarily premature infants, during ROP screenings at University Hospital Ostrava, Czech Republic. These images were captured using three imaging systems: Clarity RetCam 3, Natus RetCam Envision, and Phoenix ICON. Along with the images, anonymized metadata crucial for medical analysis is included, such as patient ID, sex, gestational age, birth weight, postconceptual age, diagnosis code, plus form, device type, and series number. To support diverse machine learning workflows, the images are organized in three formats: (1) patient-specific folders grouped by unique IDs and imaging sessions, (2) a bulk folder containing all images for batch processing, and (3) a format without captions for compatibility with untagged inputs.

2) Characterization: Each image in the dataset includes detailed metadata fields, encoded in filenames, to provide essential contextual information for tagging and training the model. These fields are described as follows:

1) *Patient ID:* A unique anonymized identifier grouping images belonging to the same patient across multiple sessions.
2) *Sex (SEX):* The gender of the patient, either male (M) or female (F). The dataset is balanced with 94 male and 94 female patients, containing 3,081 and 2,923 images, respectively.
3) *Gestational Age (GA):* The gestational age at birth, recorded in weeks. Premature infants, particularly those born before 32 weeks, are at increased risk of severe ROP.
4) *Birth Weight (BW):* The birth weight, measured in grams, aids in understanding the relationship between low birth weight and disease progression.
5) *Postconceptual Age (PA):* The sum of gestational and chronological ages in weeks, providing consistency when comparing images captured during the same examination series.
6) *Diagnosis Code (DG):* A code categorizing ROP severity based on expert assessment, with values ranging from Stage 0 (no ROP) to Stage 5 (complete retinal detachment). Additional codes identify specific conditions such as AP-ROP or post-treatment stages.

TABLE II
DIAGNOSIS CODES AND CORRESPONDING STAGES

Diagnosis Code (DG)	ROP Stage
0	No ROP
1	ROP 0
2	ROP 1
3	ROP 2
4	ROP 3
5	ROP 4A
6	ROP 4B
7 till 12	ROP 5 (+ Additional Elements)

7) *Plus Form (PF):* Indicates retinal blood vessel abnormalities, with values for normal vessels (0), pre-plus disease (1), and plus disease (2).
8) *Imaging Device (D):* The specific retinal imaging system used for capturing images, ensuring consistency and robustness in testing.
9) *Series Number (S):* The examination series for a patient, enabling tracking of disease progression over multiple visits.

The dataset provides a structured metadata framework and comprehensive diagnostic information, supporting the development of machine learning models capable of diagnosing Retinopathy of Prematurity (ROP) and assessing disease severity. The dataset is both comprehensive and diverse; however, for our model, we will primarily be using the diagnosis codes.

It is worth noting that ROP is categorized into 13 classes, but our dataset contains only 11 classes, with each diagnosis code signifying a specific form of ROP. We will use these 11 classes for model training and evaluation.

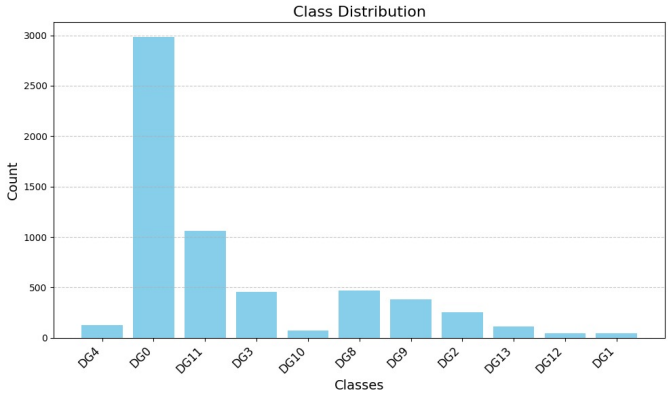

Fig. 1. Distribution of Diagnosis Codes (DG) in the Dataset.

As shown in Fig. 1, The diagnosis codes offer crucial insights into the representation of different stages of ROP. A significant portion of the dataset is classified as diagnosis code 0 (No ROP), indicating a class imbalance, where cases with no ROP are more prevalent. In contrast, higher diagnosis codes, corresponding to severe stages of ROP or other retinal abnormalities, are comparatively less represented, highlighting the rarity of severe conditions in the dataset.

The statistical analysis indicates that the mean diagnosis code is approximately 3.59, suggesting a skew toward milder stages of ROP. Additionally, the standard deviation of 4.8 reflects a wide spread of data, ranging from "No ROP" to severe retinal conditions. Fig. 2 shows some stages of ROP for better visualization and understanding.

B. Data Augmentation

To enhance the diversity of the training dataset and improve the robustness of our machine learning model, we employed a systematic data augmentation approach. The original dataset, consisting of 6004 retinal images spanning 11 diagnostic classes, was augmented using the TensorFlow ImageDataGenerator library. The augmentation pipeline involved applying transformations such as small random rotations (up to 2°), shifts (up to 5% of the image dimensions), scaling variations (0.85 to 1.15), horizontal flips, and nearest-mode filling for empty areas. Images were categorized into diagnostic classes (DG0, DG1, DG10, DG11, DG12, DG13, DG2, DG3, DG8, DG9) based on filenames, and a maximum of 2 augmented images were generated per original image. Although the original dataset contained 11 diagnostic classes, one class with very few and noisy images was excluded during augmentation, resulting in a total of 12,342 images across 10 classes. This process enhanced class representation, addressed imbalance, and improved variability for better model generalization.

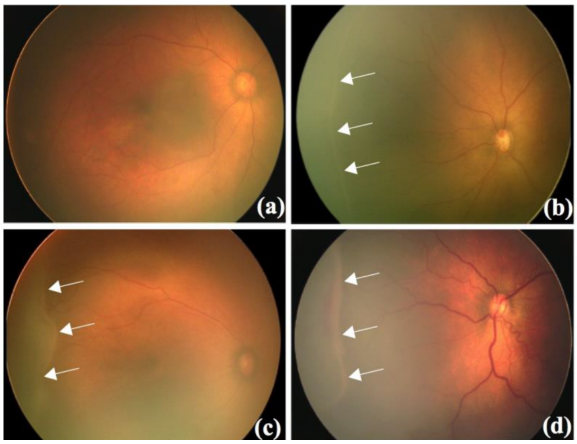

Fig. 2. Illustration of retinal fundus images at different stages of retinopathy of prematurity (ROP; indicated by arrows). (a) Normal/NOROP; (b) ROP-Stage 1; (c) ROP-Stage 2; and (d) ROP-Stage 3.

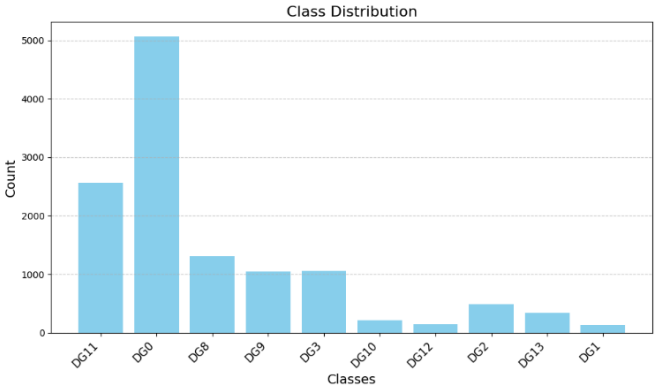

Fig. 3. Class distribution of the augmented dataset.

The class distribution of the augmented dataset is shown in Fig. 3. DG0 (No ROP) dominates the dataset, followed by DG11, while other diagnostic classes have fewer samples. The augmentation process significantly improved representation for underrepresented classes.

C. Loading & Splitting Dataset

To prepare the augmented dataset for model training, we employed TensorFlow's dataset and preprocessing utilities. The dataset, consisting of augmented retinal fundus images, was loaded using the image_dataset_from_directory methord. Then Random shuffling was done and the dataset was partitioned into training, validation, and testing subsets using a custom function. The partitioning strategy split the dataset

into 80% for training, 10% for validation, and 10% for testing.To standardize the images, a preprocessing pipeline was implemented, which resized all images to 256×256 and rescaled pixel values to the range [0, 1]. This comprehensive preprocessing approach ensured the dataset was adequately prepared for training our machine learning models.

D. General Training Details

All models were implemented using TensorFlow and Keras. Initially tested with a batch size of 32, the models were fine-tuned to a batch size of 16 for optimal performance. Each model processed images with an input size of 256×256 pixels, consisting of 3 channels for RGB images.

The models were compiled using the Adam optimizer and a sparse categorical cross-entropy loss function, suitable for multi-class classification tasks. A learning rate of 1×10^{-4} was employed to fine-tune the training process. Early stopping was used to monitor validation loss, restoring the best weights after 5 epochs of no improvement to prevent overfitting.

At the end of training, the models were saved using the model.save() method and reloaded to evaluate their performance on the test dataset. Metrics such as test loss and test accuracy were computed to assess the models' ability to generalize to unseen data.

E. Model Architectures

1) Convolutional Neural Network (CNN): The CNN model was designed to classify retinal images based on diagnostic categories. It featured four convolutional layers, each followed by max-pooling operations to reduce spatial dimensions. The convolutional layers used 32 and 64 filters with a 3×3 kernel and ReLU activation. After flattening the output, the model included two dense layers, with the final layer having n neurons (corresponding to the number of diagnostic classes) and a softmax activation function for multi-class classification.

2) ResNet (Residual Neural Network): The ResNet model followed a bottleneck design using residual blocks to enable deep feature extraction while maintaining efficient gradient flow. It started with an initial convolutional layer followed by residual blocks composed of 1×1, 3×3, and 1×1 convolutions. Each block incorporated batch normalization and ReLU activation. The network included three stages, with increasing complexity as it progressed, using 32, 64, and 128 filters. After the convolutional layers, global average pooling was applied, followed by a fully connected layer with a softmax output for multi-class classification.

3) EfficientNet-like Model: The EfficientNet-like model implemented principles of the EfficientNet architecture, including Mobile Inverted Bottleneck Convolution (MBConv) blocks and Squeeze-and-Excitation (SE) modules. The architecture began with an initial convolutional block, followed by multiple MBConv blocks with progressively larger filters and kernel sizes. Depthwise separable convolutions were employed for computational efficiency. SE modules recalibrated channel-wise feature responses. The final layers included global average pooling, dropout for regularization, and a fully connected layer with a softmax activation for multi-class classification.

4) Inception-like Model: Inspired by the GoogleNet Inception architecture, the Inception-like model utilized parallel feature extraction pathways. Each Inception block contained convolutional layers with 1×1, 3×3, and 5×5 filters, alongside a max-pooling branch. Outputs from these parallel paths were concatenated along the channel dimension. The architecture consisted of multiple Inception blocks with increasing filter sizes, interspersed with max-pooling layers to downsample feature maps. The final layers included global average pooling, dropout for regularization, and a fully connected layer with a softmax activation function for multi-class classification.

5) VGG19 (Reduced VGG19-like Architecture): The reduced VGG19 architecture retained the essence of the original model while being optimized to approximately 1 million parameters for computational efficiency. It consisted of a series of convolutional and max-pooling blocks. The first two blocks had one convolutional layer each with 32 and 64 filters, respectively. The next two blocks each included two convolutional layers with 128 and 256 filters. All convolutional layers used a 3×3 kernel with ReLU activation and padding to preserve spatial dimensions. After the convolutional layers, global average pooling reduced the feature map size. This was followed by a dense layer with 128 neurons and ReLU activation, and a dropout rate of 0.4 was applied to mitigate overfitting. The final dense layer employed a softmax activation function for multi-class classification.

V. Results

A. CNN

The CNN model achieved a test accuracy of 98.48% with a test loss of 0.06, demonstrating superior performance compared to other models. This reflects the model's ability to generalize well, with minimal misclassifications, particularly for DG1 and DG2. The training and validation losses decreased steadily, indicating stable training throughout the process. The following figures illustrate the performance of the Convolutional Neural Network (CNN) model. Figure 4 presents the training and validation loss curves over time, showing that both losses decrease steadily, indicating model convergence. Figure 7 shows the confusion matrix, with the highest accuracy achieved for class DG0 (489 correct predictions). Misclassifications are notably frequent for DG11 (6 errors) and DG9 (5 errors), while DG1 (15 misclassifications) and DG2 (44 misclassifications) show relatively lower accuracy. Figures 5 and 6 display the test loss and test accuracy, respectively, on a batch-wise basis.

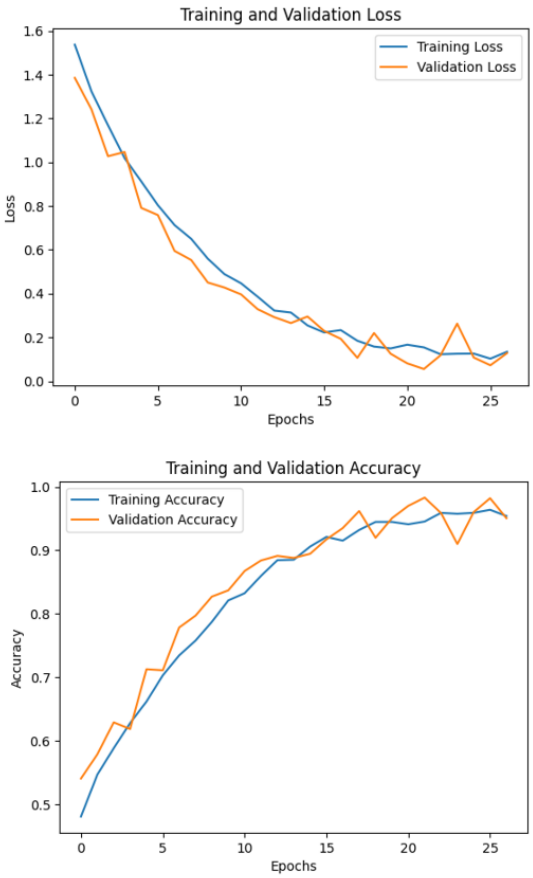

Fig. 4. Training and Validation Loss and accuracy for CNN

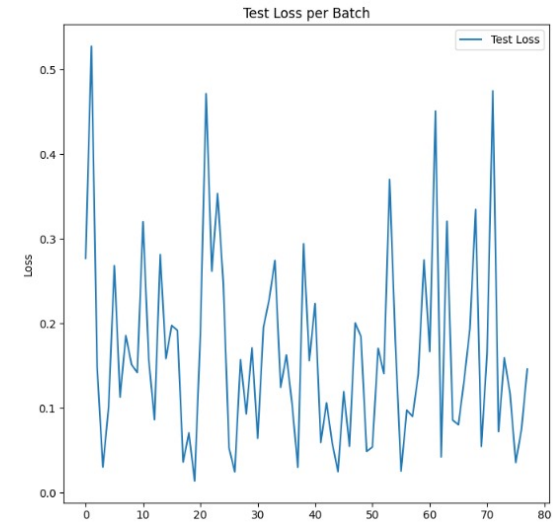

Fig. 5. Test Loss for CNN

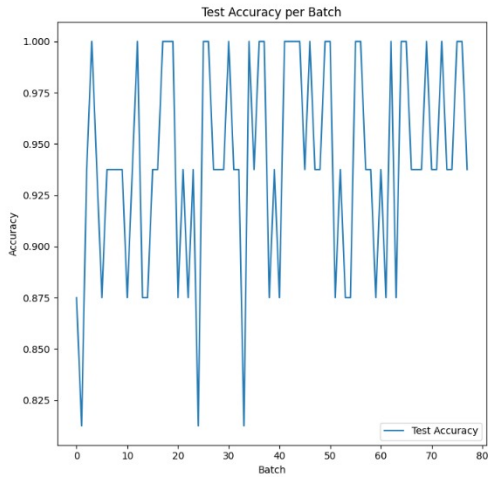

Fig. 6. Test Accuracy for CNN

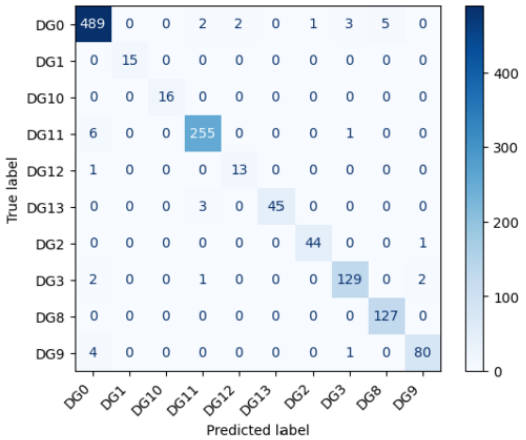

Fig. 7. Confusion Matrix for CNN

B. ResNet

The ResNet model achieved a test accuracy of 95.35% with a test loss of 0.14. The model performed well, except for DG10 and DG11, where minimal misclassifications occured. The accuracy curve showed steady improvement over epochs, and both training and validation losses decreased consistently, confirming the model's stability during training. The following figures illustrate the performance of the ResNet model. Figure 8 presents the training and validation loss curves over time. As observed, both the training and validation losses decrease steadily, indicating that the ResNet model converges effectively during training. Figure 11 shows the confusion matrix of the ResNet model. This model shows improved accuracy for DG0, with 514 correct predictions, and for DG9, with 109 correct predictions, compared to the first model. However, it still struggles with DG11, where 25 misclassifications are observed. Figures 9 and 10 display test loss and test accuracy, respectively, batch-wise.

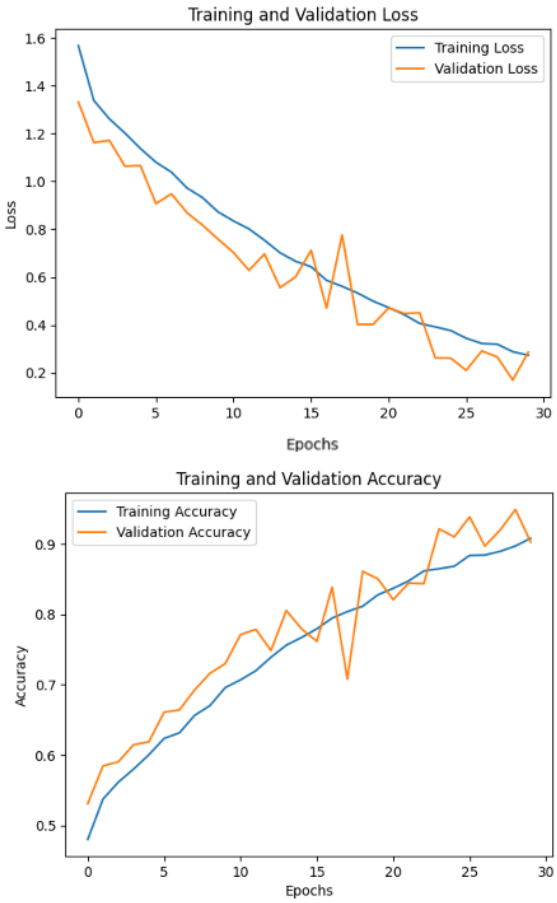

Fig. 8. Training and Validation Loss and accuracy for ResNet

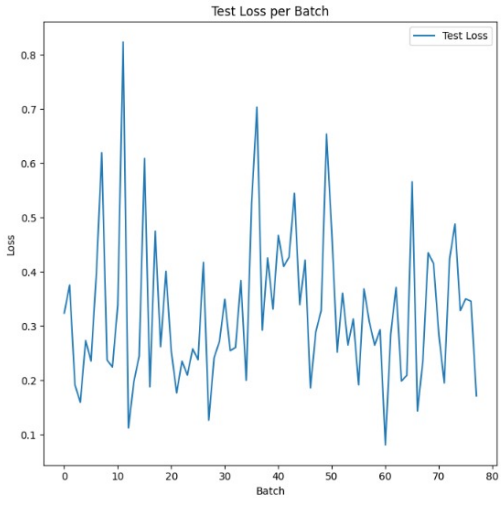

Fig. 9. Test Loss for ResNet

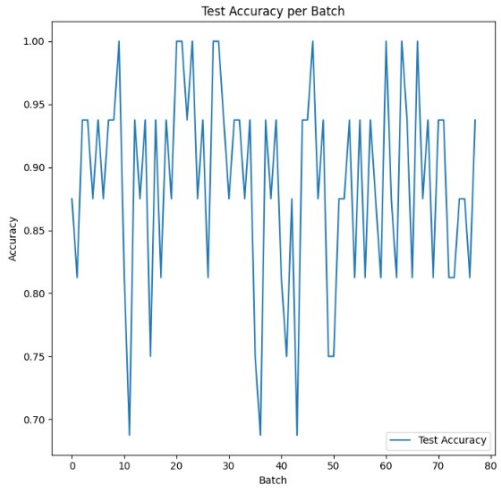

Fig. 10. Test Accuracy for ResNet

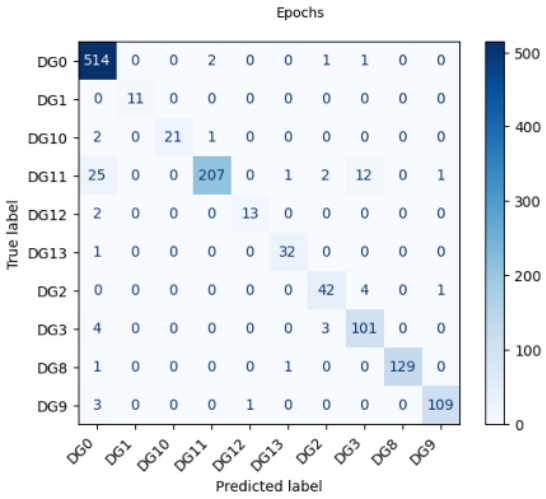

Fig. 11. Confusion Matrix for ResNet

C. *EfficientNet*

EfficientNet reached a test accuracy of 96.07% with a test loss of 0.13. The model demonstrated strong performance, particularly for DG0 and DG11, with few misclassifications. The accuracy curve showed consistent progress, and the stable loss trends indicate effective learning with a steady reduction in both training and validation losses. The following figures illustrate the performance of the EfficientNet model. Figure 12 presents the training and validation loss curves over time. As seen, both the training and validation losses decrease steadily, indicating that the EfficientNet model also converges effectively during training. Figure 15 shows the confusion matrix of the EfficientNet model. For this model, the performance trends are similar to the first two, with DG0 (481 correct predictions) and DG11 (257 correct predictions) remaining the strongest classes. The main areas of misclassification are DG9 (92 errors) and DG3 (113 errors). While DG0 shows

a slight decrease in performance, marginal gains are seen in other classes like DG2, which improves to 52 correct predictions. Figures 13 and 14 display test loss and test accuracy, respectively, batch-wise.

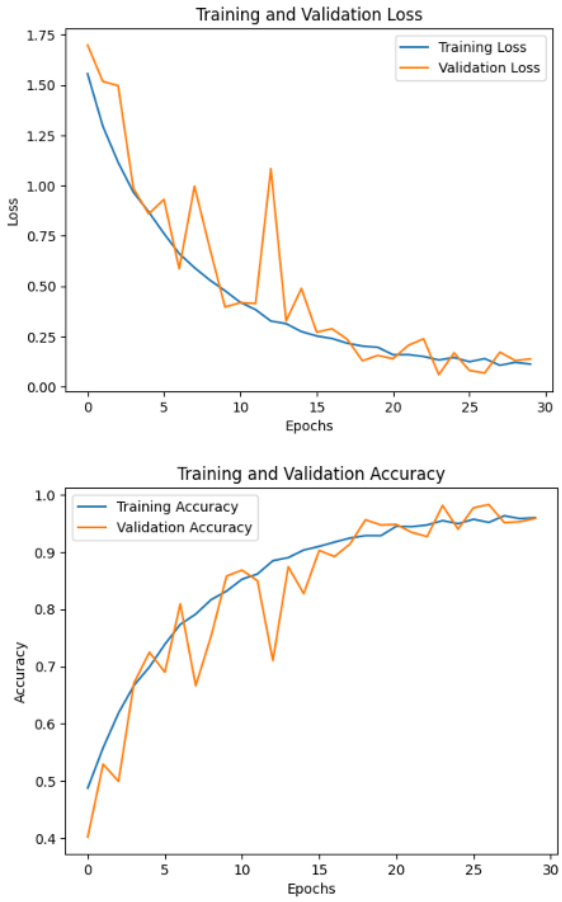

Fig. 12. Training and Validation Loss and accuracy for EfficientNet

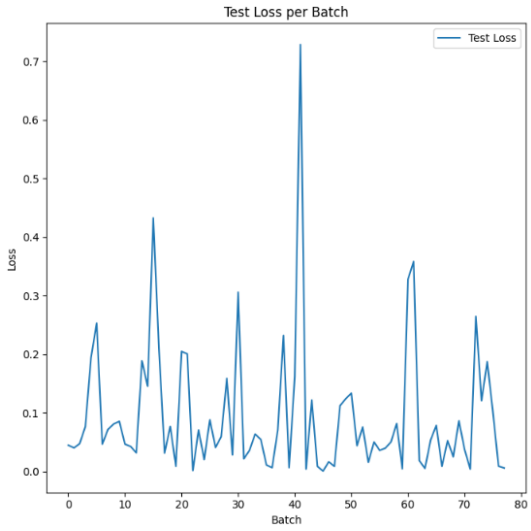

Fig. 13. Test Loss for EfficientNet

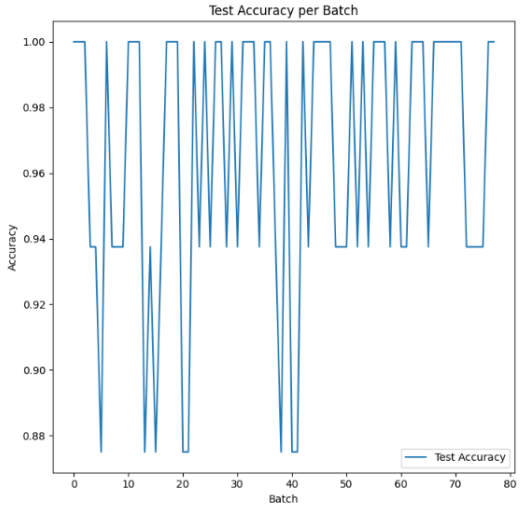

Fig. 14. Test Accuracy for EfficientNet

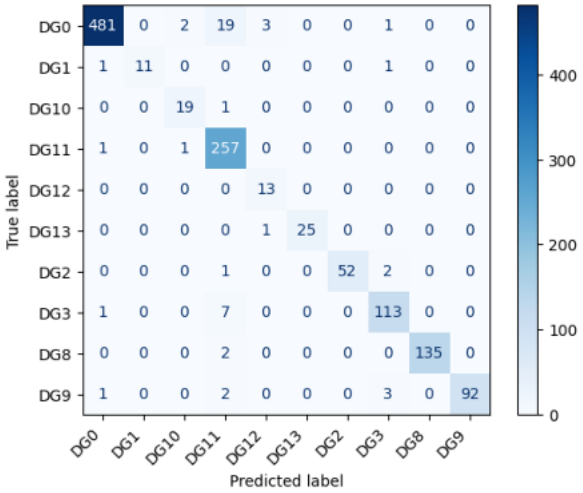

Fig. 15. Confusion Matrix for EfficientNet

D. Inception

The Inception model achieved a test accuracy of 91.83% with a test loss of 0.25. While the model showed reasonable classification performance, there were more misclassifications compared to other models, especially for DG0 and DG11. The accuracy curve demonstrated slower convergence, and the loss trends indicate moderate stability during training. The following figures illustrate the performance of the Inception model. Figure 16 presents the training and validation loss curves over time. As shown, both the training and validation losses decrease steadily, demonstrating the effective convergence of the Inception model during the training process. Figure 19 shows the confusion matrix of the Inception model. The matrix shows strong performance for classes DG11, DG0 and DG8, with most instances correctly classified. However, the model struggles with distinguishing between DG0 and DG1, DG10

and DG11, and DG2 and DG3. Figures 17 and 18 display the test loss and test accuracy, respectively, batch-wise.

Fig. 16. Training and Validation Loss and accuracy for Inception

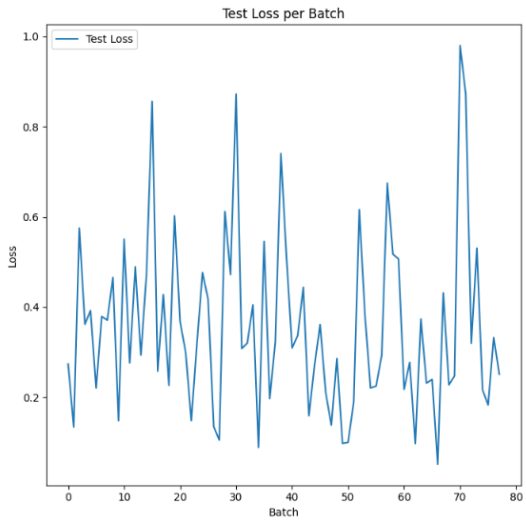

Fig. 17. Test Loss for Inception

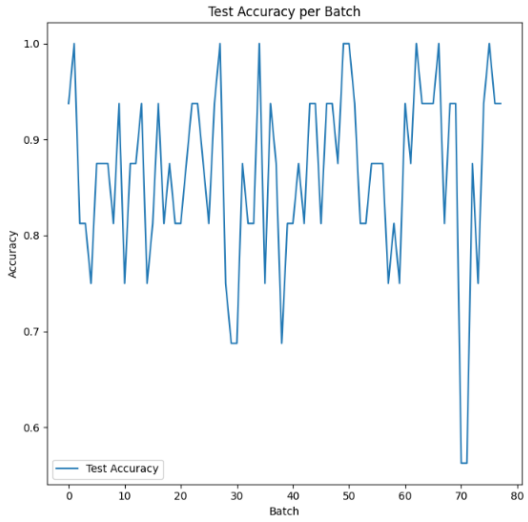

Fig. 18. Test Accuracy for Inception

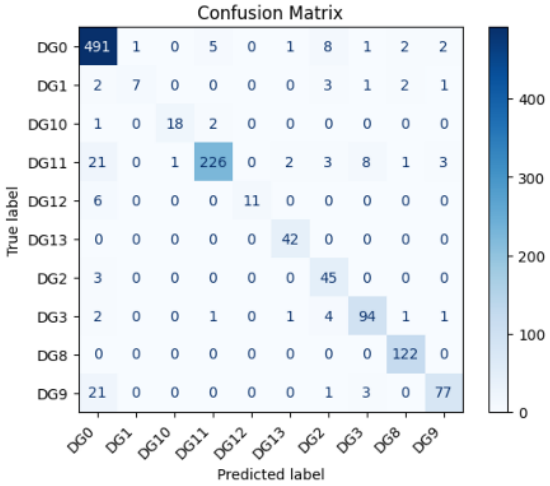

Fig. 19. Confusion Matrix for Inception

E. VGG

The VGG model achieved the lowest performance, with a test accuracy of 88.86% and a test loss of 0.32. The model exhibited higher misclassification rates, particularly for DG0 and DG11. The accuracy curve showed slow improvement, and the loss trends reflected less consistent reductions in both training and validation losses. The following figures illustrate the performance of the VGG model. Figure 20 presents the training and validation loss curves over time, showing that both losses decrease consistently, indicating that the VGG model converges effectively during the training process. Figure 23 shows the confusion matrix, where the VGG model demonstrates strong overall performance with high accuracy for most classes, particularly DG11. However, more misclassifications are observed in DG0 and DG11, even more so than the other models tested, hence making it the worst performer. Figures

21 and 22 display the test loss and test accuracy, respectively, on a batch-wise basis.

Fig. 20. Training and Validation Loss and accuracy for VGG

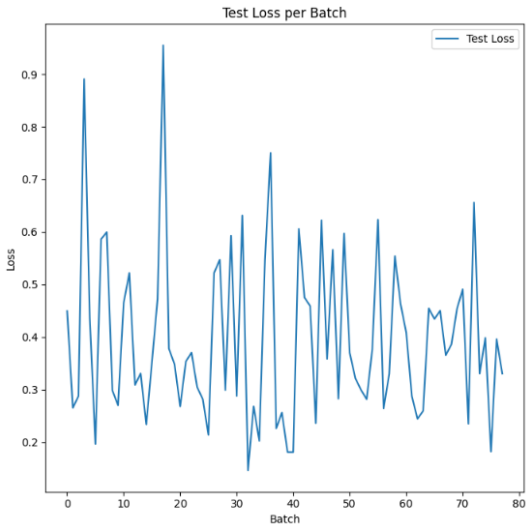

Fig. 21. Test Loss for VGG

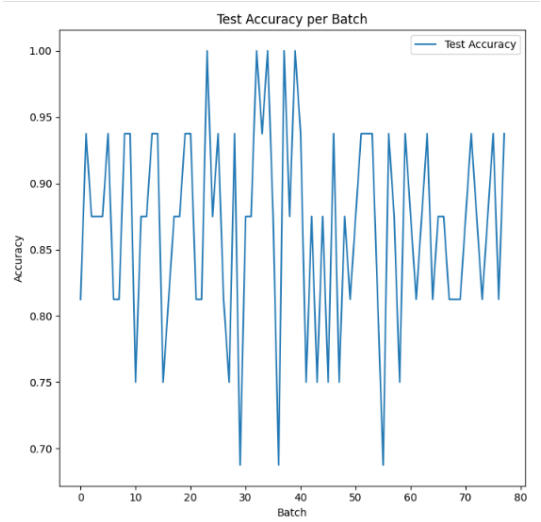

Fig. 22. Test Accuracy for VGG

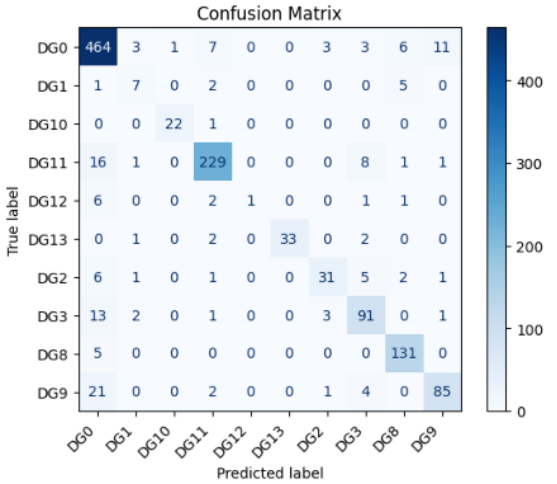

Fig. 23. Confusion Matrix for VGG

F. Summary

TABLE III
SUMMARY OF TEST RESULTS FOR ALL MODELS

Model	Test Loss	Test Accuracy
CNN	0.06	98.48%
ResNet	0.14	95.35%
EfficientNet	0.13	96.07%
Inception	0.25	91.83%
VGG	0.32	88.86%

The performance of the five models is summarized in Table III. Among all models, the CNN model outperformed others with the lowest test loss of 0.06 and the highest test accuracy of 98.48%. EfficientNet followed closely with a test accuracy of 96.07% but had a slightly higher test loss of 0.13. ResNet achieved a test accuracy of 95.35% but showed a higher test loss compared to EfficientNet. Inception and VGG had the

lowest test accuracies of 91.83% and 88.86%, respectively, with higher losses as well.

Based on these results, it can be concluded that the CNN model demonstrates superior performance in terms of both accuracy and stability, making it the most reliable choice for detection and classification of Retinopathy of Prematurity (ROP).

VI. Discussion

The results from this study demonstrate that the CNN model outperforms other architectures in detecting and classifying Retinopathy of Prematurity (ROP), with a test accuracy of 98.48% and a test loss of 0.06. This aligns with findings from the literature, where CNNs have consistently delivered strong performance in medical imaging tasks. Specifically, the study by [1] showed that CNN-based models, such as OC-Net and SE-Net, achieved high sensitivity (100%) in predicting ROP, albeit with lower specificity. This aligns with our findings, where the CNN model in this study also demonstrated strong sensitivity and generalization to various ROP classes.

In comparison, models like ResNet (95.35%) and EfficientNet (96.07%) performed well but were not as accurate as the CNN model. These results are similar to [2], where the Inception-v3 CNN model showed high sensitivity (96.6%) but slightly lower performance on external validation, suggesting that while these models perform well in controlled settings, they may face challenges in more diverse data environments. This is reflected in the misclassifications observed for certain ROP classes, such as DG9 and DG3, in our ResNet and EfficientNet models.

The Inception and VGG models in this study performed the weakest, with accuracies of 91.83% and 88.86%, respectively. This is consistent with findings in [3], where VGG models achieved high accuracy but faced challenges in specific ROP severity classifications. The lower performance of these models in our study highlights their limitations for ROP detection when compared to the CNN and ResNet architectures, suggesting that more advanced architectures may be necessary for improving classification accuracy.

While the CNN model excels in terms of accuracy and stability, similar to the findings in [5], where deep learning combined with SVMs showed high accuracy for ROP detection, the need to balance sensitivity and specificity remains. Most studies, including ours, have prioritized high sensitivity to minimize missed diagnoses, which often leads to increased false positives. Future work could explore methods to enhance specificity while maintaining the high sensitivity demonstrated by models like CNN.

In conclusion, this study confirms that CNN is the best-performing model for ROP detection, outperforming other deep learning architectures in terms of both accuracy and generalization. The results are consistent with the literature, reinforcing CNN's utility in medical image classification tasks, especially for early diagnosis and intervention in ROP.

VII. Conclusion and Future Prospects

This study evaluated five deep learning models—custom CNN, ResNet, EfficientNet, Inception, and VGG—for the detection and classification of Retinopathy of Prematurity (ROP) in retinal images. The custom CNN achieved the highest performance, with a test accuracy of 98.48% and a test loss of 0.06, making it the most reliable model for ROP detection. EfficientNet and ResNet followed with accuracies of 96.07% and 95.35%, respectively, while Inception and VGG showed lower accuracies of 91.83% and 88.86%. These results underscore the potential of deep learning for improving ROP detection, particularly in resource-limited areas, by automating screening and enhancing diagnostic capabilities.

Looking ahead, there are several avenues for improving the model's performance and clinical applicability. Data augmentation techniques, such as geometric, color, and noise augmentations, could enhance generalization, especially with a more diverse dataset. Incorporating explainable AI tools will increase model transparency and trust, which is crucial for clinical adoption. Additionally, integrating clinical data, like patient history and birth weight, alongside retinal images could refine predictions and provide a more comprehensive assessment of ROP severity.

To ensure the model's generalizability, it should be validated on external datasets from diverse populations. Ensemble learning methods could further improve accuracy by combining predictions from multiple models. Finally, integrating the model into a Clinical Decision Support System (CDSS) would streamline diagnosis and assist in prioritizing urgent cases.

In conclusion, while deep learning models show great promise for ROP detection, further advancements in data diversity, model interpretability, and clinical integration are essential for real-world deployment. Addressing these challenges could significantly improve early ROP detection and reduce blindness in premature infants. Thus, we conclude our research overview.

References

[1] S. Shah, E. Slaney, E. VerHage, J. Chen, R. Dias, B. Abdelmalik, A. Weaver, and J. Neu, "Application of artificial intelligence in the early detection of retinopathy of prematurity: Review of the literature," *Neonatology*, vol. 120, no. 5, pp. 558–565, Jul. 2023. doi: 10.1159/000531441. Available: https://pubmed.ncbi.nlm.nih.gov/37490881/

[2] B. A. Scruggs, R. V. P. Chan, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Artificial intelligence in retinopathy of prematurity diagnosis," *Translational Vision Science & Technology*, vol. 9, no. 2, p. 5, Feb. 2020. doi: 10.1167/tvst.9.2.5. Available: https://pubmed.ncbi.nlm.nih.gov/32704411/

[3] Z. Tan, S. Simkin, C. Lai, and S. Dai, "Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease," *Translational Vision Science & Technology*, vol. 8, no. 6, p. 23, Dec. 2019. doi: 10.1167/tvst.8.6.23. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6892443/

[4] Q. Wu, Y. Hu, Z. Mo, R. Wu, X. Zhang, Y. Yang, B. Liu, Y. Xiao, X. Zeng, Z. Lin, Y. Fang, Y. Wang, X. Lu, Y. Song, W. W. Ng, S. Feng, and H. Yu, "Development and validation of a deep learning model to predict the occurrence and severity of retinopathy of prematurity," *JAMA Network Open*, vol. 5, no. 6, p. e2217447, Jun. 2022. doi: 10.1001/jamanetworkopen.2022.17447. Available: https://pubmed.ncbi.nlm.nih.gov/35708686/

[5] Y.-P. Huang, S. Vadloori, H.-C. Chu, E. Y.-C. Kang, W.-C. Wu, S. Kusaka, and Y. Fukushima, "Deep learning models for automated diagnosis of retinopathy of prematurity in preterm infants," *Electronics*, vol. 9, no. 9, p. 1444, Sep. 2020. doi: 10.3390/electronics9091444. Available: https://doi.org/10.3390/electronics9091444

[6] J. Timkovič, J. Nowaková, J. Kubíček, 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. Available: https://doi.org/10.1038/s41597-024-03409-7, Dataset available at: https://www.kaggle.com/datasets/jananowakova/retinal-image-dataset-of-infants-and-rop/data.

[7] M. Hasal, J. Nowaková, D. Hernández-Sosa, and J. Timkovič, "Image enhancement in retinopathy of prematurity," in *Advances in Intelligent Networking and Collaborative Systems*, L. Barolli and H. Miwa, Eds., vol. 527, Springer, Cham, 2022, pp. 385–392. doi: 10.1007/978-3-031-14627-5_43. Available: https://doi.org/10.1007/978-3-031-14627-5_43

[8] M. Hasal, M. Pecha, J. Nowaková, D. Hernández-Sosa, V. Snášel, and J. Timkovič, "Retinal vessel segmentation by U-Net with VGG-16 backbone on patched images with smooth blending," in *Advances in Intelligent Networking and Collaborative Systems*, L. Barolli, Ed., vol. 182, Springer, Cham, 2023, pp. 431–440. doi: 10.1007/978-3-031-40971-4_44. Available: https://doi.org/10.1007/978-3-031-40971-4_44

[9] J. Nowaková, "Image enhancement in retinopathy of prematurity," GitHub Repository, 2023. Available: https://github.com/JanaNowakova/Image_enhancement_retinopathy_of_prematurity