



Article

Deep Learning Models for Automated Diagnosis of Retinopathy of Prematurity in Preterm Infants

Yo-Ping Huang ^{1,2,*} , Spandana Vadloori ¹, Hung-Chi Chu ² , Eugene Yu-Chuan Kang ^{3,4}, Wei-Chi Wu ^{3,4,*}, Shunji Kusaka ⁵ and Yoko Fukushima ⁶

¹ Department of Electrical Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; spandanamanoj@gmail.com

² Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung 41349, Taiwan; hcchu@cyut.edu.tw

³ Department of Ophthalmology, Chang Gung Memorial Hospital, Linkou 33305, Taiwan; yckang0321@gmail.com

⁴ College of Medicine, Chang Gung University, Taoyuan 33305, Taiwan

⁵ Department of Ophthalmology, Faculty of Medicine, Kindai University, Osaka 577-8502, Japan; kusaka-ns@umin.net

⁶ Department of Ophthalmology, Osaka University, Osaka 565-0871, Japan; yokofukushima@icloud.com

* Correspondence: yphuang@ntut.edu.tw (Y.-P.H.); weichi666@gmail.com (W.-C.W.)

Received: 18 August 2020; Accepted: 2 September 2020; Published: 4 September 2020



Abstract: Retinopathy of prematurity (ROP) is a disease that can cause blindness in premature infants. It is characterized by immature vascular growth of the retinal blood vessels. However, early detection and treatment of ROP can significantly improve the visual acuity of high-risk patients. Thus, early diagnosis of ROP is crucial in preventing visual impairment. However, several patients refrain from treatment owing to the lack of medical expertise in diagnosing the disease; this is especially problematic considering that the number of ROP cases is on the rise. To this end, we applied transfer learning to five deep neural network architectures for identifying ROP in preterm infants. Our results showed that the VGG19 model outperformed the other models in determining whether a preterm infant has ROP, with 96% accuracy, 96.6% sensitivity, and 95.2% specificity. We also classified the severity of the disease; the VGG19 model showed 98.82% accuracy in predicting the severity of the disease with a sensitivity and specificity of 100% and 98.41%, respectively. We performed 5-fold cross-validation on the datasets to validate the reliability of the VGG19 model and found that the VGG19 model exhibited high accuracy in predicting ROP. These findings could help promote the development of computer-aided diagnosis.

Keywords: deep neural networks; transfer learning; retinopathy of prematurity; retinal fundus images

1. Introduction

Retinopathy of prematurity (ROP) is a disease that can potentially cause blindness in preterm infants. ROP is caused by the pathological neovascularization in the retinal fundus of premature infants [1]. ROP continues to be a major, preventable cause of blindness and visual impairment in children both in developing and developed countries [2]. ROP occurs in babies born prematurely after 32 weeks and with low birth weight (less than 1.5 kg) [3,4]. Globally, 19 million children are estimated to suffer from visual impairment [5]. Over 1.84 million of these children were likely to have developed ROP at any stage, of which approximately 11% would have become totally blind or severely visually impaired and 7% would have developed mild/moderate visual impairment because of ROP [6]. The incidence of ROP in developed and developing countries is estimated to be 9% and 12%, respectively. ROP, like any other disease, can progress from mild to severe stages [7]. Abnormal

growth of the retinal blood vessels is observed in ROP-affected infants. Blindness can also occur because of retinal detachment, unless treated in the initial stages [2]. Laser treatment, anti-VEGF therapy, surgical treatment, or treatment with drugs have proven to be effective in treating ROP [8–10]. ROP is categorized from mild to severe (Stage 1 to Stage 5), depending on the severity [4,11,12]. In brief, Stage 1 is the initial stage, where abnormal growth of the blood vessels occurs due to the occurrence of a thin flat whitish line known as the demarcation line, which separates the retinal regions in the eye. This demarcation line prevents the supply of blood to the outer edges of the retina. In Stage 2, this thin demarcation line transforms into a ridgeline, which means that the thin whitish line becomes broader and is raised and changes in color from white to pinkish. In Stage 3, the ridge demarcation line increases in dimension, and new abnormal blood vessels grow internally (Figure 1). In Stage 4, partial retinal detachment occurs, which may result in complete retinal detachment. Finally, in Stage 5, the person may become blind or suffer from permanent loss of vision [4,12].

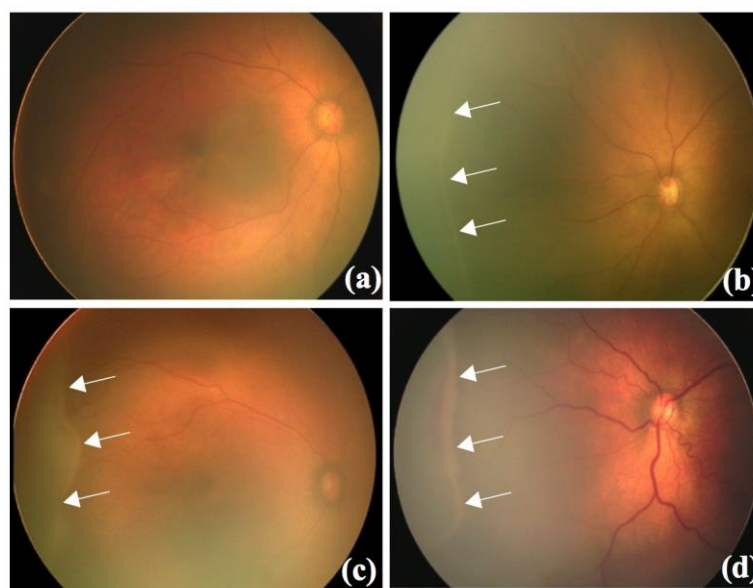


Figure 1. Illustration of retinal fundus images at different stages of retinopathy of prematurity (ROP; indicated by arrows). (a) Normal/NOROP; (b) ROP-Stage 1; (c) ROP-Stage 2; and (d) ROP-Stage 3.

Studies have shown that the condition of an infant with Stage 2 ROP may improve without treatment. However, if the disease has progressed to Stage 3, diagnosis and treatment are crucial to prevent the disease from progressing to later stages. Various strategies for treating ROP are available [13,14]. Regular screening of preterm infants is crucial because distinctive features of ROP could be associated with sequential syndromes such as astigmatism, myopia, glaucoma, cataracts, anisometropia, amblyopia, strabismus, and retinal detachment. ROP can be detected by either pediatric ophthalmologists or retinopathy specialists. However, while the number of cases of ROP is on the rise, the number of ophthalmologists capable of ROP screening is on the decline [15,16]. In rural areas, in particular, the detection of ROP is not easy owing to a lack of ROP specialists. Approximately 36% of neonatologists in the USA were unable to transfer children with ROP to a neonatal intensive care unit for screening owing to a lack of specialists at the care unit [17]. Alternate strategies such as telemedicine computer-aided diagnosis (CAD) of diseases must be adopted to diagnose ROP in patients. Telemedicine has been found to be effective in the diagnosis of ROP [18], and the CAD of ocular diseases has made considerable progress; data reveal its high potential for future breakthroughs [19,20].

The use of artificial intelligence (AI) in the field of medicine has increased in recent years owing to advancements in AI technologies. Deep learning models have made incredible progress in the field of medical diagnosis and have been employed practically in various computer vision tasks, including image classification, object detection, image segmentation, and disease diagnosis. Owing to

the advancement in deep network architectures and access to big data, the use of AI has been proposed to reduce the burden on medical experts. Traditional machine learning algorithms such as logistic regression, support vector machines, and fuzzy decision trees have been used in the field of image recognition and classification. However, other techniques such as feature extraction and dimensionality reduction are required to accomplish the task, which is time-consuming. Moreover, the conversion of the image matrix to a one-dimensional vector leads to the loss of some critical information from the image, which could lower the performance of the models. In the case of a convolutional neural network (CNN), classification is accomplished by extracting features from raw input images by the convolutional layers followed by dimensionality reduction by the pooling layer.

Transfer learning is a useful concept in CNNs, which use previously acquired knowledge and skills and apply them to a different but related problem. Pretrained models such as VGG16 and InceptionV3 have been trained using rich data sources such as ImageNet, which contains 1.2 million natural images with more than 1000 categories [21]. These models are built from scratch using substantial computational resources. These models have learned features such as edges, shapes, lighting, rotation, and spatial information. This knowledge is useful for extracting features from images in a different domain. Thus, the availability of vast training datasets is essential for a model to achieve high performance; training the data with a small dataset may lead to underperformance or overfitting, which can be overcome by transfer learning. Thus, transfer learning is particularly useful in classification tasks; it improves the generalization ability of a model when the training dataset is small (not even in the thousands) [22]. This strategy is useful for classifying images and predicting disease where the dataset is small, such as the dataset used in the present study.

CNNs have been used in image classification, and since 2012, they have exhibited high performance in the diagnosis of diseases [23]. CNNs have been successfully used in the diagnosis of lung cancer [9], glioma [24], pneumonia [25], skin cancer [26], brain tumor [27], and other medical conditions [28]. Recently, deep learning was also used for the accurate diagnosis of the COVID-19 symptoms by using CT images [29]. Deep learning has also been used for the diagnosis of eye diseases such as diabetic retinopathy [30,31] and glaucoma [32], which are eye diseases associated with ROP. Studies have developed a deep learning algorithm for the automated diagnosis of plus disease by using fundus images [33]. Transfer learning has been used to pretrain models for classifying ROP images [34]. Studies have also employed a CAD system for plus disease and the measurement of tortuosity from retinal fundus images [35]. Owing to the excellent results achieved with CNNs in the medical image processing field, researchers proposed a novel CNN architecture for diagnosing plus disease in ROP by using a pretrained GoogLeNet to visualize feature maps of pathologies learned directly from the data [36]. The field has advanced with the use of two CNN methods to diagnose plus disease in ROP [37]. Recently, ROP was screened using deep neural networks (DNNs) [38–40]. In these studies, retinal fundus images were used to train and test fundus images for detecting ROP.

A robust and reliable automated ROP detection system is currently required to diagnose ROP in the initial stages of development. To this end, the present study aimed to achieve high accuracy in the diagnosis of ROP by using RetCam fundus images captured from preterm infants. The system was trained with a dataset, and it tested eye-based diseases to predict the classification performance. We also applied transfer learning to the deep CNNs. The first step was identifying whether the eye condition was normal (NOROP) or abnormal (ROP). Furthermore, based on the severity of the abnormal condition, we classified it as either mild-ROP or severe-ROP. Blindness due to ROP in infants can be prevented through early diagnosis. Therefore, early identification of the disease is crucial for administering proper treatment to prematurely born infants to prevent blindness. Detection in the initial stages of ROP development is essential for understanding the progression of the disease. This study presents an automated diagnosis of ROP by using various classification models. Our findings have the potential to assist ophthalmologists in diagnosing the disease at an early stage. The purpose of the present study was to provide a CAD system in a clinical setting for diagnosing ROP. We applied transfer learning to the deep CNN models and achieved high accuracy in the prediction of

eye-based cases. Moreover, the different stages of ROP were accurately classified based on the severity of the disease.

In this study, we applied transfer learning to deep CNN models and compared their capabilities in the detection of ROP by using retinal fundus images. We aimed to determine the absence or presence of ROP (NOROP or ROP) in a preterm infant as well as the severity of the disease (mild-ROP or severe-ROP). We used five pretrained models with different architectures, namely VGG19, VGG16, InceptionV3, DenseNet, and MobileNet. The major contributions of the study are as follows:

1. We investigated a large variety of backbone models of different architectures; these models differed in the number of convolutional layers they had. The pretrained models and their number of convolutional layers are listed in Table 1.
2. We comprehensively explored different backbone architectures in terms of performance. We demonstrated significant variation in performance across backbone models.
3. Owing to the variation in performance across the different backbones in this domain, our work becomes significant as it indicates the necessity to improve on backbone models selection and provides clear benchmarks to assist it.
4. We achieved the optimal results with the VGG19 model in terms of classifying ROP and NOROP and identifying the severity of ROP with high sensitivity and specificity.
5. We performed 5-fold cross-validation on the datasets to evaluate the performance of the VGG19 model.

Table 1. Pretrained models and their number of convolutional layers.

Classification Model	No. of Convolutional Layers
VGG16	13
VGG19	16
MobileNet	28
InceptionV3	48
DenseNet	103

The rest of the paper is organized as follows: In Section 2, we provide a brief description of the dataset, an overview of the training of classification models, and the evaluation method. In Section 3, we present our approach and its results on the performance of the classification models in the diagnosis of ROP, along with a discussion. In Section 4, we summarize our findings, draw some conclusions, and state directions for future work.

2. Materials and Methods

In this study, we aimed to predict the occurrence of ROP in a preterm infant's eyes. We examined the retinal fundus images from patients' eyes, which indicated the absence or presence of ROP.

2.1. Dataset

All the fundus images were captured by expert technicians using the RetCam imaging system (Clarity Medical System, Pleasanton, CA, USA). The datasets were procured from the neonatal intensive care units of (1) Chang Gung Memorial Hospital, Linkou, Taiwan, and (2) Osaka Women's and Children's Hospital, Japan. They are specialized hospitals and have been providing ROP screening services for several years. A total of 5–22 images were collected during each ROP screening session, and the dataset from each patient was split into two eye cases such as NOROP or ROP. The patients' demographic datasets were captured before July 2019. The patients had to satisfy at least one of the following criteria in order to be selected in this study: the babies had to be born within 37 weeks of gestation and/or had to weigh ≤ 1500 g at birth.

2.2. Image Labeling

Three senior ophthalmologists who had over 10 years of experience working with patients with ROP were involved in the study. These experts labeled the fundus images as NOROP (normal/without disease) or ROP (with the disease) according to the guidelines set by the International Classification of Retinopathy of Prematurity. Furthermore, the different stages of ROP were classified as Stage 1, Stage 2, and Stage 3. The three ophthalmologists first labeled the images independently; the images were then compared to identify any inconsistency in the labeling process (i.e., to identify whether a particular image was assigned different labels by the experts). Subsequently, the labels were sorted collectively after a discussion among the experts and a label was assigned to such images. The ophthalmologists defined the severity of the disease as mild-ROP, if the eye cases belonged to Stage 1 and Stage 2 ROP, or severe-ROP, if the eye cases belonged to Stage 3 ROP. A description of the different ROP stages can be found in the literature [11,41].

First, the present study aimed to identify from fundus images whether an infant had ROP. Then, the images that indicated the presence of ROP were further classified as mild-ROP or severe-ROP. All the different test cases from the patients were manually labeled by these experts and compared with the DNN model predictions.

2.3. Dataset Description and Preprocessing

The resolutions of the multiple fundus images of infants' eyes were 1600×1200 for the Taiwanese dataset and 640×480 for the Japanese dataset. A total of 6500 images of left and right eyes of 210 infants were collected. We used data of 106 patients for training the ROP/NOROP model. The unclear images, blurred images, dark images, etc., were omitted from the analysis. We considered the fundus images showing the different stages of ROP in the same infant for analysis, ensuring no overlaps between the patients from the training dataset and test dataset.

2.3.1. Image Normalization

All the data were first subjected to preprocessing to run the classification models. The preprocessing step included resizing the images to $224 \text{ pixels} \times 224 \text{ pixels} \times 3 \text{ pixels}$ for the MobileNet, DenseNet, VGG16, and VGG19 models and $299 \text{ pixels} \times 299 \text{ pixels} \times 3 \text{ pixels}$ for the InceptionV3 model. These images were then loaded using "OpenCV," resized, and converted to a NumPy array. Normalization was further carried out on the input images, where they were rescaled to have pixel values between 0 and 1 by dividing all the pixel values with the highest pixel value of 255.

Using preprocessing tools with the Keras API of the ImageDataGenerator, we performed data augmentation and loaded the model with weights on convolutional layers.

2.3.2. Data Augmentation

Training the model with a small amount of data can lead to overfitting during training. To overcome this issue, we employed data augmentation to create new retinal fundus images from the existing training dataset. Data augmentation was used to generate more datasets. In this study, we used various augmentation techniques that included `rotation_range` $[-3, 3]$, `width_shift_range` $[-0.1, 0.1]$, `height_shift_range` $[-0.1, 0.1]$, `zoom_range` $[0.85, 1.15]$, and `horizontal_flip`. The training dataset was augmented seven times, resulting in a total of 18,808 images for training. From our initial tests, as expected, we observed that data augmentation was useful in increasing the prediction accuracies of the ROP and NOROP datasets in the case of the VGG19 and VGG16 models. Hence, we applied the augmentation techniques to all the classification models used in the present study.

Blurry or bright images or images that were not clear were filtered out from the image datasets of all the patients. For the NOROP training dataset, we selected 108 eye cases from 54 patients and obtained a total of 1222 images. For training the ROP dataset, which included Stage 1, Stage 2, and Stage 3 ROP cases, we selected a similar number of patients and images randomly to balance with

the NOROP dataset. Overall, for training the ROP dataset, we selected 159 cases from 52 patients and obtained 1129 images. For testing the accuracy of the classification models, data from 25 patients were used. Details on the number of patients and eye cases used in the training set and test set are presented in Table 2. The training set and test set for identifying the severity of the disease as mild-ROP or severe-ROP are presented in Table 3.

Table 2. NOROP (absence of ROP) and ROP (presence of ROP) dataset used for training and testing.

Dataset	Training Set			Test Set	
	Patients	No. of Cases	Images	Patients	No. of Cases
NOROP	54	108	1222	21	42
ROP	52	159	1129	26	59
Total	106	267	2351	47	101

Table 3. Mild-ROP and severe-ROP dataset used for training and testing.

Dataset	Training Set			Test Set	
	Patients	No. of Cases	Images	Patients	No. of Cases
Mild-ROP	45	146	1189	25	63
Severe-ROP	54	108	1174	11	22
Total	99	254	2363	36	85

2.4. Classification Model Training

We applied transfer learning to the models for ROP identification and ROP severity classification. Transfer learning was performed by freezing the initial layers of the pretrained model and replacing the three fully connected (FC) layers with the final layer as a classification layer. The weights from the convolution layers were copied instead of weights of the entire network with FC layers. An illustration of the model is shown in Figure 2. In the present study, we confirmed the relevant parameters of the FC layers through testing with different layer sizes of 100–600 to obtain optimal results (i.e., high accuracy on the validation set, low error rate, and no overfitting). The results of a comparison of the models are shown as a confusion matrix.

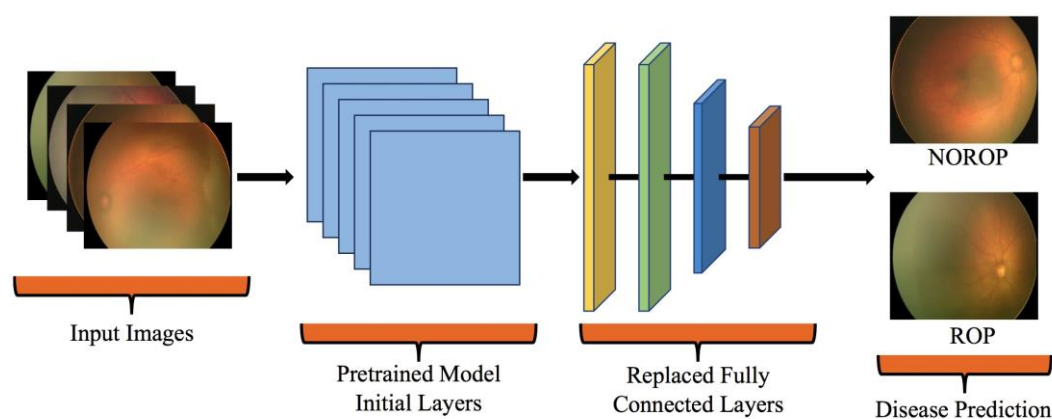


Figure 2. Graphical representation of transfer learning with pretrained models for disease prediction.

In the present study, we covered a large variety of backbone models by selecting them from different architecture types such as the VGG family (VGG11, VGG13, VGG16, and VGG19), MobileNet group (MobileNet, MobileNetv2, ShuffleNet, and FD-MobileNet), Inception (Inception, InceptionV1, InceptionV2, and InceptionV3), and DenseNet group (DenseNet, HarDNet, and S-Net). In this study, we selected two models from the VGG group and one each from the remaining groups. In total,

we selected five different classification models, with each having a different number of layers (ranging from 13 to 103). The models VGG19 and VGG16 [42], which belong to the same family, have 16 and 13 convolutional layers, respectively. The InceptionV3 [43], DenseNet [44], and MobileNet [45] models have 48, 28, and 103 convolutional layers, respectively. All these five DNN models were selected to achieve our primary aim of identifying ROP. From the results, the performance of the models was then evaluated, and two models that exhibited the best performance were chosen for identifying the severity of the disease. Our method included the loading of weights of the pretrained model provided by Keras. We added our classifiers by replacing the FC layers of the model with four dense layers and fine-tuned them. In the VGG16 and VGG19 models, the first and second FC layers had a size of 200 each, and the dropout layers had a 50% drop rate. The third FC layer had a size of 64, and the third dropout layer had a 50% drop rate. The final layer was a softmax layer, which was stacked at the end for classifying the fundus images, followed by the FC layer output to determine whether the image should be classified as ROP or NOROP. In the Inception V3, MobileNet, and DenseNet models, the first and second FC layers had a size of 100 and 64, respectively. The first, second, and third dropout layers had a drop rate of 50%, and the final layer was a softmax layer. In the identification of the severity of ROP (mild-ROP or severe-ROP), the optimal results were obtained with two FC layers, with the third layer as the classification layer. Here, the first and second FC layers had a size of 512 and 200, respectively, and the dropout layers had a 50% dropout rate. The Adam optimizer was used at a learning rate of 2×10^{-5} , categorical cross-entropy was used as the loss function, and the batch size was set to 10.

2.5. Model Evaluation

The findings of the classification models are represented as a confusion matrix. In binary classification, a confusion matrix represents information of the classes with a number of instances/cases in true positives (TPs—instances correctly predicted to the class of interest), true negatives (TNs—instances correctly predicted that belong to the other class of interest), false positives (FPs—instances assigned to the class of interest but do not belong to it), and false negatives (FNs—instances assigned to the class of interest but belong to the complementary class). A conventional illustration of the confusion matrix is given in Figure 3.

		Predicted labels	
		0	1
True labels	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Figure 3. Illustration of the confusion matrix.

We evaluated and compared the performance of the five models by calculating the sensitivity, specificity, precision, accuracy, true positive rate, and false positive rate using the equations given below. In brief, sensitivity refers to the percentage of TP that are correctly predicted by the classification model that performs the testing of the test cases, whereas specificity refers to the percentage of TN that are correctly identified by the model [46]. Precision is a measure of the percentage of instances where a classifier is labeled as positive to the total predictive positive cases [47].

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \times 100\%, \quad (1)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \times 100\%, \quad (2)$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \times 100\%, \quad (3)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \times 100\%. \quad (4)$$

Figure 4 shows the schematic of the entire workflow of the classification process. The entire dataset was first divided into training and test datasets. These datasets then underwent preprocessing and normalization. The preprocessed training data were then subjected to augmentation. After model testing and hyperparameter tuning to obtain the optimal results on the validation dataset, the model was deployed to the test dataset for binary classification. The model performance was then evaluated in terms of prediction accuracy, sensitivity, and specificity. Additionally, we calculated the area under the curve (AUC) for evaluating the performance of the models and visualized the problems presented by different models in classifying the stages of ROP.

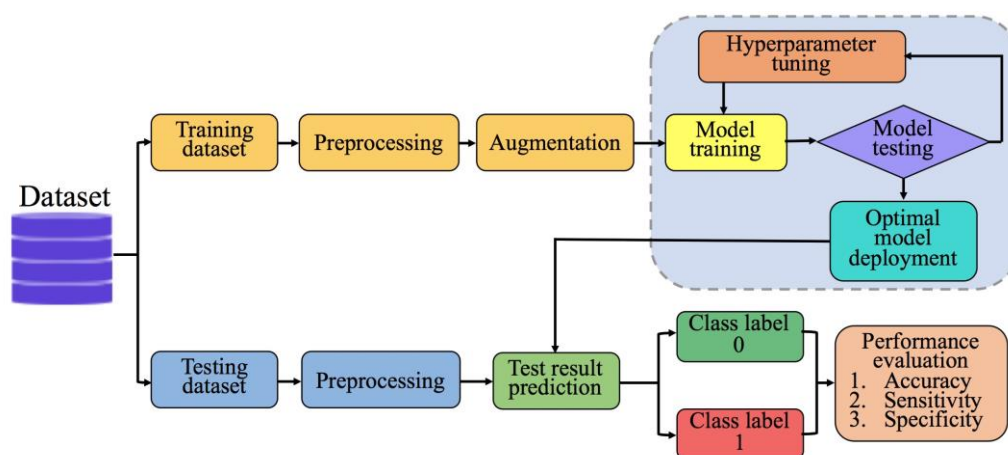


Figure 4. Schematic of the classification process.

2.6. 5-Fold Cross-Validation

Cross-validation was used to improve the accuracy and reliability of the model using the training and test samples multiple times. We evaluated the performance of the VGG19 model through 5-fold cross-validation on ROP/NOROP data. All the patients' datasets were combined and divided into 5 folds: 80% (4 folds) as a training dataset and 20% (1 fold) as a test dataset. We validated the performance of the model on each of the 5 folds. To train the ROP and NOROP data, 75 and 78 patients were used, respectively; the number of eye cases was 150 and 218, respectively. Each fold contained the data of at least 15 patients. Similarly, we performed 5-fold cross-validation on mild-ROP and severe-ROP patient data. 70 and 65 patients were used to train the mild-ROP and severe-ROP data, respectively; the number of eye cases was 209 and 130, respectively.

3. Results and Discussion

In this study, we first identified infants with and without ROP and classified the eyes of the patients as ROP or NOROP. Since ROP is a progressive disease, we then classified the different stages of disease development. Stage 1 and Stage 2 are considered the preliminary stages of ROP, and Stage 3 is considered the critical stage. In Stage 3, treatment must be initiated to prevent an infant from losing vision. Thus, we classified the disease stages as mild-ROP or severe-ROP. Here, mild-ROP was defined as the group that comprised Stage 1 and Stage 2 ROP patients, whereas severe-ROP comprised of Stage 3 ROP patients.

3.1. Experimental Setup

The proposed ROP classification was performed in Python on a Windows operating system with a configuration of Intel Core i5-CPU @ 2.7 GHz with 24 GB RAM on NVIDIA GEFORCE GTX 1050TI (Santa Clara, CA, USA). The classification was performed on training, validation, and test datasets.

To accomplish our primary objective, we evaluated the accuracy of the classification models in identifying whether infants' eyes were indicative of NOROP or ROP. We first performed a classification study using five pretrained deep learning models on the datasets consisting of ROP and NOROP to select the model that most accurately predicts the presence or absence of the disease from the patients' eyes. Our strategy was identifying the eyes of the patients. The models were initially trained to achieve high accuracy using the training model. Then, the test cases were predicted using the model. Several images belonging to one particular eye from a patient was considered as an eye case. These images of an eye case were fed into the classification model to record the performance of the model. The results were compared with the results of the labeling performed by the ophthalmologists.

3.2. Diagnosis of ROP by DNN Models

A total of 42 test cases labeled as having no symptoms of the disease by the ophthalmologists were assigned as NOROP, whereas 59 test cases labeled as having any of the stages of ROP were assigned as ROP. The models were trained using the training dataset and validated with the test dataset. After a model was sufficiently trained, the test cases were predicted. A set of images belonging to a particular eye from a patient was given as input, and the output was obtained as an array. It included the prediction of the model; the output predicted by the model was labeled as either 0 or 1 for NOROP and ROP, respectively. The test cases were predicted based on the label given by the model for each image. For example, when the classifier labeled all the images as 0 from a test case of NOROP, which was previously labeled by ophthalmologists as 0, then we considered the test case prediction as NOROP. Otherwise, the eye was labeled as 1 (i.e., ROP). This test case was considered a misclassification. All the test cases were predicted in the same manner; the results are listed in Table 4. In the case of the VGG19 model, for the 42 NOROP test cases, 40 cases were correctly classified as NOROP (TN), and 2 cases were misclassified as ROP (FP). Likewise, for the 59 ROP test cases, 57 were correctly classified as ROP (TP), and 2 were misclassified as NOROP (FN).

Table 4. Confusion matrix of the prediction of the NOROP and ROP derived from the classification models.

		0	1
VGG19	0	40	2
	1	2	57
VGG16	0	32	10
	1	2	57
InceptionV3	0	17	25
	1	3	56
DenseNet	0	37	5
	1	19	40
MobileNet	0	36	6
	1	8	51

Similarly, in the VGG16 model, for the 42 NOROP test cases, 32 cases were correctly classified as NOROP, and 10 were misclassified as ROP. For the 59 test cases of ROP, the VGG16 correctly classified 57 of the test cases as ROP and misclassified 2 test cases as NOROP. The details of the other classifiers—InceptionV3, DenseNet, and MobileNet—are listed in Table 4 in the form of a confusion matrix, where 0 and 1 indicate NOROP and ROP, respectively. Some of the misclassified images from the test cases are shown in Figure 5. Here, the prediction was performed using the test dataset to obtain the correct prediction and incorrect prediction. For example, in a situation where a NOROP test case containing eight images was sent as input for prediction, and seven images were correctly classified in the array and one image was incorrectly classified, we considered the test case as ROP, even if the classifier mislabeled a single image.

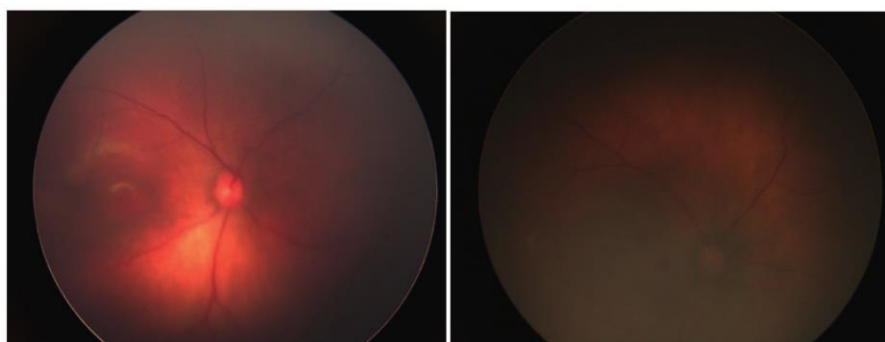


Figure 5. Examples of some of the misclassified retinal fundus images. NOROP test cases were predicted to be ROP test cases.

Our results revealed that VGG19 outperformed all other models in identifying ROP and NOROP with a sensitivity and specificity of 96.6% and 95.2%, respectively. Without data augmentation, the accuracies of the VGG19 and VGG16 models were 82.6% and 74.5%, respectively. Transfer learning helps develop robust models. We trained the VGG19 and VGG16 models from scratch. However, without transfer learning, we obtained poor results in the identification of the disease. After augmentation and transfer learning, the VGG19 model exhibited the highest accuracy among the five DNN classifiers; it had a prediction accuracy of 96.0% with the NOROP and ROP test cases and an AUC value of 0.97. The second-best classification model was VGG16, with an accuracy of 88.1%. The sensitivity of this model was equal to that of the VGG19 model. However, the other performance metrics of the VGG16 model (specificity, precision, and AUC) were inferior to those of the VGG19 model. For the remaining three classifiers, the accuracies in descending order were 86.1%, 76.2%, and 72.3% for MobileNet, DenseNet, and InceptionV3, respectively (Table 5). The performance of all the five models is shown in the form of receiver operating characteristics (ROC) curves in Figure 6. A comparative analysis of these five models with the test dataset suggested that the VGG19 model was the best in identifying the presence or absence of ROP.

Table 5. Performance evaluation of the five deep neural network (DNN) models.

Classification Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
VGG19	96.0	96.6	95.2	95.2	0.97
VGG16	88.1	96.6	76.2	94.1	0.96
InceptionV3	72.3	94.9	40.5	85.0	0.76
DenseNet	76.2	67.8	88.1	66.1	0.77
MobileNet	86.1	86.4	85.7	81.8	0.87

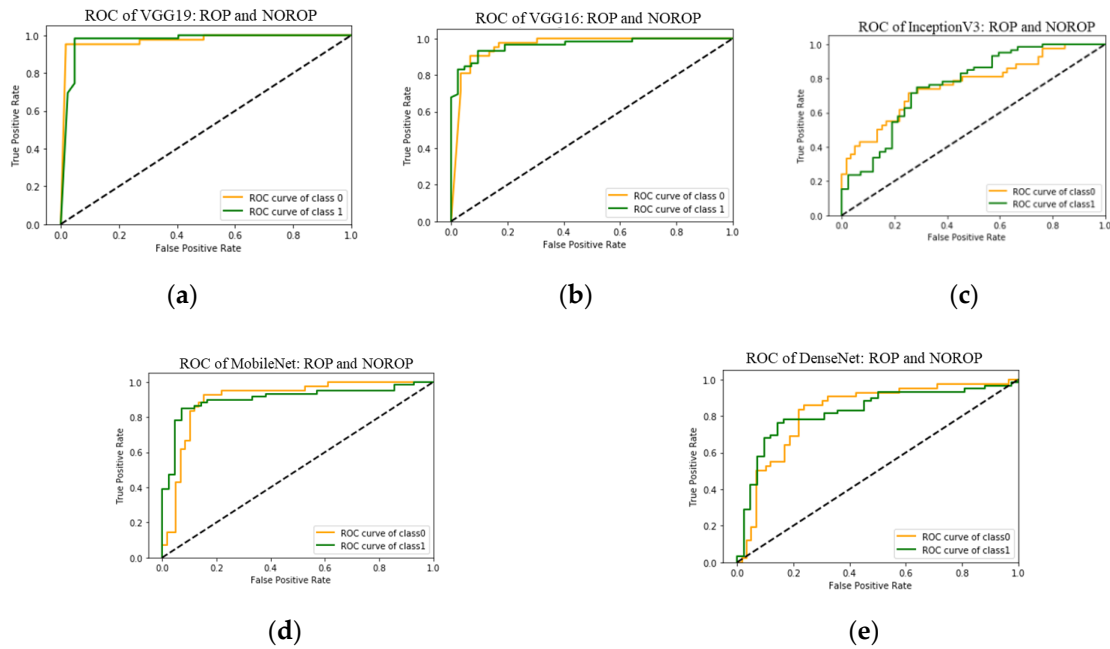


Figure 6. Receiver operating characteristics (ROC) curves of the DNN classification models. (a) VGG19, (b) VGG16, (c) InceptionV3, (d) MobileNet and (e) DenseNet. The orange and green ROC curves in the individual plots represent Class 0 (NOROP) and Class 1 (ROP), respectively.

To evaluate the performance of the VGG19 model, we performed 5-fold cross-validation. We divided the data into five folds and tested the accuracy of each fold. We observed good accuracy in most of the folds. The highest accuracy achieved was 94.6% with fold 4 (Table 6), which exhibited a sensitivity, specificity, and precision of 91.1%, 99.2%, and 99.3%, respectively. The results of the 5-fold cross-validation are listed in Table 6. The ROC curves with the AUC are shown in Figure 7. The results indicate the performance of the VGG19 model.

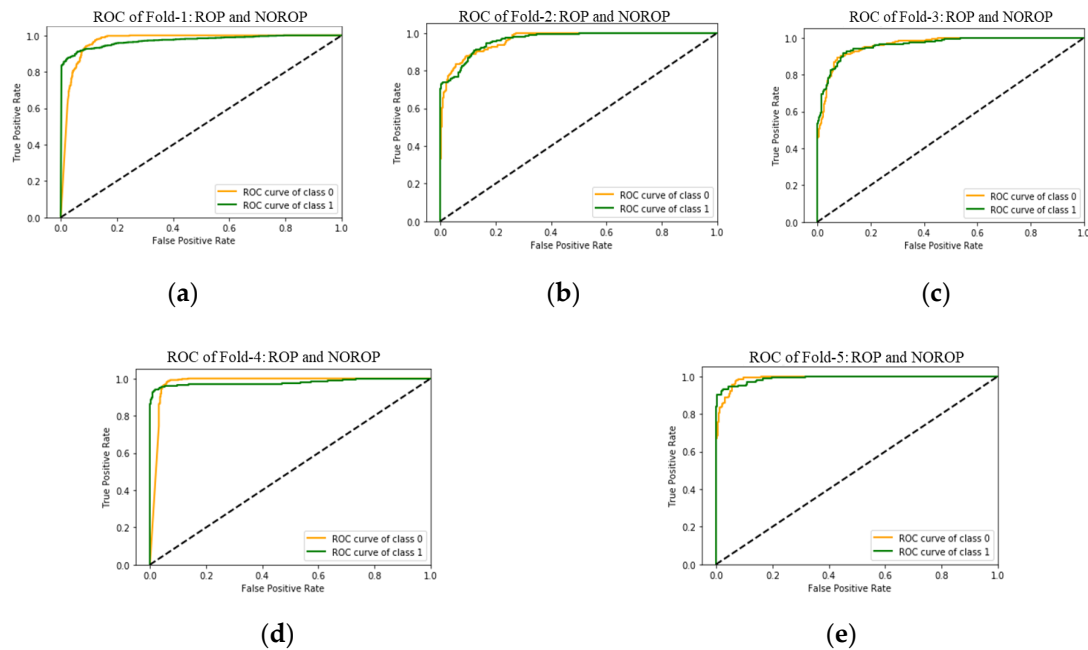


Figure 7. ROC curves for each of the 5 folds of data (a–e) obtained using the VGG19 classification model. The orange and green ROC curves in individual plots represent class 0 (NOROP) and class 1 (ROP), respectively.

Table 6. Performance evaluation of the 5-fold cross-validation of the VGG19 model.

Classification Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
Fold 1	92.1	92.5	91.2	95.6	0.97
Fold 2	91.0	92.8	87.4	93.6	0.97
Fold 3	90.6	95.0	87.0	86.0	0.97
Fold 4	94.6	91.1	99.2	99.3	0.98
Fold 5	94.1	94.4	93.6	96.0	0.99

In our first step, we obtained high accuracies with VGG19 and VGG16 in the prediction of the disease (Table 4). The detection of the incidence and severity of ROP is crucial for treatment. Thus, we performed binary classification to determine whether the severity of the disease was low (mild-ROP) or high (severe-ROP). Such a diagnosis would allow proper treatment to be administered on time. In the second step, which involved predicting whether the disease was in the mild or severe stage, we trained the VGG19 and VGG16 models, which showed better performance in the first step of ROP identification. We provided the test case images as input to the model, and based on the prediction result, the accuracy was determined. Our results show that the VGG19 model had 98.8% accuracy in predicting the severity of the disease.

Our results with the VGG19 model show that out of 63 test cases of mild-ROP, the model predicted 62 cases correctly, and only one test case was missed. In terms of severe-ROP, the model did not mispredict any test case (Table 7). The accuracy of the VGG19 model in predicting mild-ROP and severe-ROP was 98.8%. Similar predictions were made with the VGG16 classifier; two of the mild-ROP cases were misclassified, and only one of the severe-ROP cases was misclassified, giving the model an overall prediction accuracy of 96.5%. Details of the performance metrics are presented in Table 8, and the ROC curves are shown in Figure 8. The identification of the severity of the disease is important for ophthalmologists because proper treatment on time could potentially prevent an infant from becoming blind.

Table 7. Confusion matrix of the test cases predicted for mild-ROP, denoted as 0, and severe-ROP, denoted as 1, by the VGG19 and VGG16 models.

		0	1
VGG19			
	0	62	1
	1	0	22
VGG16			
	0	61	2
	1	1	21

Table 8. Performance evaluation of the VGG19 and VGG16 models for predicting mild-ROP and severe-ROP.

Classification Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
VGG19	98.8	100.0	98.4	95.7	0.99
VGG16	96.5	95.5	96.8	91.3	0.96

We also performed 5-fold cross-validation to evaluate the performance of the VGG19 model with the mild-ROP and severe-ROP data. Our results show 100% accuracy in one of the folds, with 100% sensitivity, specificity, and precision. The lowest accuracy observed was 97.8% (Table 9). The performance was also evaluated using ROC curves, as shown in Figure 9, which indicates the consistency of the VGG19 model.

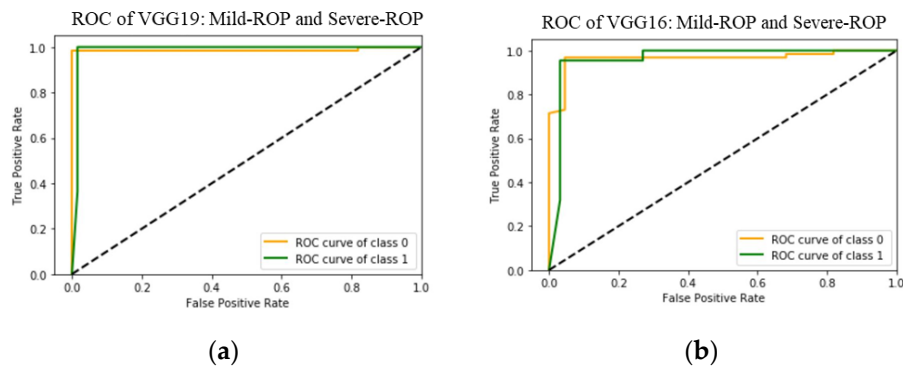


Figure 8. ROC curves of (a) VGG19 and (b) VGG16 models for classifying mild-ROP and severe-ROP.

Table 9. Performance evaluation of the 5-fold cross-validation of the VGG19 model.

Classification Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
Fold 1	99.0	98.4	99.2	98.4	0.999
Fold 2	100.0	100.0	100.0	100.0	1.000
Fold 3	97.8	98.6	96.6	97.6	0.998
Fold 4	97.8	98.0	97.7	98.0	0.992
Fold 5	98.8	98.8	98.8	99.2	0.999

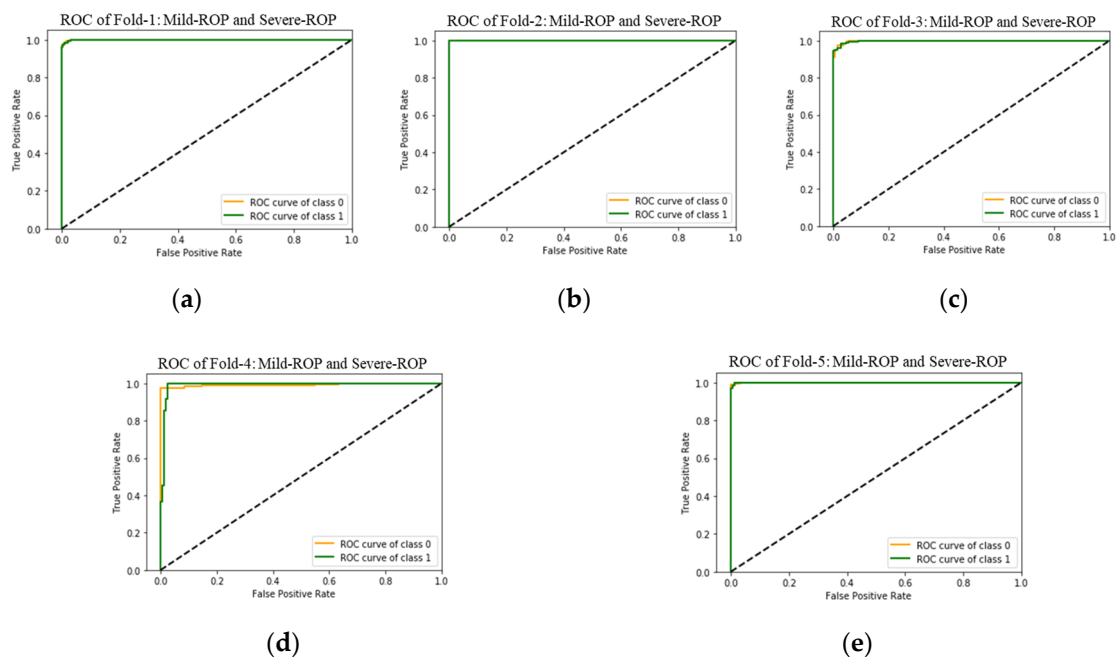


Figure 9. ROC curves for each of the 5 folds of data (a–e) obtained using the VGG19 classification model. The orange and green ROC curves in individual plots represent class 0 (mild-ROP) and class 1 (severe-ROP), respectively.

Wang et al. [38] used Id-Net and Gr-Net to diagnose ROP and identify its severity. They reported a sensitivity and specificity of 88.5% and 92.3%, respectively, in identifying the severity of the disease. In our study, with 5-fold cross-validation by using the VGG19 model for identifying ROP severity, we obtained average sensitivity and specificity values of 98.7% and 98.5%, respectively. Hu et al. [38] used the DNN models InceptionV2, VGG16, and ResNet-50 to identify ROP severity; they achieved an accuracy of 84.0%. In our study with the VGG19 model, we achieved an average accuracy 98.7% after 5-fold cross-validation. This indicates that our results were a significant improvement over the results of previous studies on the identification of ROP severity.

4. Conclusions

The present study describes the application of transfer learning to a deep convolutional neural network for the automated detection of ROP disease in infants. Developing a system with high prediction accuracy is essential, especially for those in rural areas where there is a lack of ophthalmology specialists and a high number of preterm infants. We used pretrained models with transfer learning to improve the accuracy in predicting ROP. The results showed that our approach could improve the accuracy in ROP prediction, even if the dataset is small, as was the case in the current study. Five different DNN classification models with transfer learning were studied for identifying the disease. Our results showed that VGG19 was the most efficient classification model for predicting the disease. The model was also efficient in detecting the severity of the disease. Since the early detection of ROP in preterm infants was considered essential for reducing the number of cases of ROP-related blindness, the proposed system was proven to be an efficient ROP diagnosis method. In the future, we aim to use the present approach to develop a mobile application that could conveniently be used for preliminary screening of high-risk patients in rural areas to detect the disease even in the absence of medical experts.

Author Contributions: Conceptualization, Y.-P.H., S.V., H.-C.C., E.Y.-C.K., and W.-C.W.; methodology, Y.-P.H., S.V., and H.-C.C.; software, S.V.; validation, Y.-P.H., E.Y.-C.K., and W.-C.W.; formal analysis, Y.-P.H. and S.V.; investigation, Y.-P.H., S.V., E.Y.-C.K., and W.W.; resources, Y.-P.H. and W.-C.W.; data curation, E.Y.-C.K., W.-C.W., S.K., and Y.F.; writing—original draft preparation, Y.-P.H., S.V., and H.-C.C.; writing—review and editing, Y.-P.H., S.V., E.Y.-C.K., and W.-C.W.; visualization, E.Y.-C.K. and W.-C.W.; supervision, Y.-P.H., E.Y.-C.K., W.W., S.K., and Y.F.; project administration, Y.-P.H. and W.-C.W.; funding acquisition, Y.-P.H. and W.-C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded in part by the Ministry of Science and Technology, Taiwan, under Grants MOST108-2221-E-027-111-MY3 and MOST108-2321-B-027-001-, and by a joint project between the National Taipei University of Technology and the Chang Gung Memorial Hospital under Grant NTUT-CGMH-109-01. This study was also supported by Chang Gung Memorial Hospital Research Grants (CMRPG310071~3 and CMRPG3G30581~3) and the Ministry of Science and Technology, Taiwan research Grants (MOST 106-2314-B-182A-040-MY3).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data Availability: The datasets used in the current study are available from the corresponding author upon request.

References

1. Nguyen, Q.D.; Tawansy, K.; Hirose, T. Recent Advances in Retinopathy of Prematurity. *Int. Ophthalmol. Clin.* **2001**, *41*, 129–151. [[CrossRef](#)] [[PubMed](#)]
2. Hansen, E.D.; Hartnett, M.E. A review of treatment for retinopathy of prematurity. *Expert Rev. Ophthalmol.* **2019**, *14*, 73–87. [[CrossRef](#)] [[PubMed](#)]
3. Palmer, E.A.; Flynn, J.T.; Hardy, R.J.; Phelps, D.L.; Phillips, C.L.; Schaffer, D.B.; Tung, B. Cryotherapy For Retinopathy of Prematurity Cooperative Group Incidence and Early Course of Retinopathy of Prematurity. *Ophthalmology* **2020**, *127*, S84–S96. [[CrossRef](#)] [[PubMed](#)]
4. Shah, P.K.; Prabhu, V.; Karandikar, S.S.; Ranjan, R.; Narendran, V.; Kalpana, N. Retinopathy of prematurity: Past, present and future. *World J. Clin. Pediatr.* **2016**, *5*, 35–46. [[CrossRef](#)] [[PubMed](#)]
5. Pascolini, D.; Mariotti, S.P. Global estimates of visual impairment: 2010. *Br. J. Ophthalmol.* **2011**, *96*, 614–618. [[CrossRef](#)]
6. Blencowe, H.; E Lawn, J.; Vazquez, T.; Fielder, A.; Gilbert, C. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatr. Res.* **2013**, *74*, 35–49. [[CrossRef](#)]
7. Chang, J.W. Risk factor analysis for the development and progression of retinopathy of prematurity. *PLoS ONE* **2019**, *14*, e0219934. [[CrossRef](#)]
8. Rajan, R.P.; Kohli, P.; Babu, N.; Dakshayini, C.; Tandon, M.; Ramasamy, K. Treatment of retinopathy of prematurity (ROP) outside International Classification of ROP (ICROP) guidelines. *Graefes Arch. Clin. Exp. Ophthalmol.* **2020**, *258*, 1205–1210. [[CrossRef](#)]

9. Xu, X.X.; Wu, Y.J.; Wu, H.Y.; Hu, Y.X.; Cheng, Y.; Yan, L.; Rao, J.; Wu, N.; Wu, X.R. Advances in Retinopathy of Prematurity. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* **2019**, *41*, 261–266.
10. Eldweik, L.; Mantagos, I.S. Role of VEGF Inhibition in the Treatment of Retinopathy of Prematurity. *Semin. Ophthalmol.* **2016**, *31*, 163–168. [[CrossRef](#)]
11. E Quinn, G. The International Classification of Retinopathy of Prematurity Revisited. *Arch. Ophthalmol.* **2005**, *123*, 991. [[CrossRef](#)]
12. Wheatley, C.M.; Dickinson, J.; Mackey, D.; Craig, J.; Sale, M.M. Retinopathy of prematurity: Recent advances in our understanding. *Arch. Dis. Child. Fetal Neonatal Ed.* **2002**, *87*, F78–F82. [[CrossRef](#)] [[PubMed](#)]
13. Hartnett, M.E.; Capone, A. Advances in diagnosis, clinical care, research, and treatment in retinopathy of prematurity. *Eye Brain* **2016**, *8*, 27–29. [[CrossRef](#)] [[PubMed](#)]
14. Mutlu, F.M.; Sarici, S.U. Treatment of retinopathy of prematurity: A review of conventional and promising new therapeutic options. *Int. J. Ophthalmol.* **2013**, *6*, 228–236. [[PubMed](#)]
15. Vartanian, R.; Besirli, C.G.; Barks, J.D.; Andrews, C.A.; Musch, D.C. Trends in the Screening and Treatment of Retinopathy of Prematurity. *Pediatrics* **2016**, *139*, 139. [[CrossRef](#)] [[PubMed](#)]
16. Kemper, A.R.; Freedman, S.F.; Wallace, D.K. Retinopathy of prematurity care: Patterns of care and workforce analysis. *J. Am. Assoc. Pediatr. Ophthalmol. Strabismus* **2008**, *12*, 344–348. [[CrossRef](#)]
17. Kemper, A.R.; Wallace, D.K. Neonatologists' practices and experiences in arranging retinopathy of prematurity screening services. *Pediatrics* **2007**, *120*, 527–531. [[CrossRef](#)]
18. Richter, G.M.; Sun, G.; Lee, T.C.; Chan, R.P.; Flynn, J.T.; Starren, J.; Chiang, M.F. Speed of Telemedicine vs. Ophthalmoscopy for Retinopathy of Prematurity Diagnosis. *Am. J. Ophthalmol.* **2009**, *148*, 136–142.e2. [[CrossRef](#)]
19. Zhang, Z.; Srivastava, R.N.; Liu, H.; Chen, X.; Duan, L.; Wong, D.W.K.; Kwok, C.K.; Wong, T.Y.; Liu, J. A survey on computer aided diagnosis for ocular diseases. *BMC Med. Inform. Decis. Mak.* **2014**, *14*, 80. [[CrossRef](#)]
20. Mookiah, M.R.K.; Acharya, U.; Chua, C.K.; Lim, C.M.; Ng, E.; Laude, A. Computer-aided diagnosis of diabetic retinopathy: A review. *Comput. Biol. Med.* **2013**, *43*, 2136–2155. [[CrossRef](#)]
21. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
22. Byra, M.; Styczyński, G.; Szmigielski, C.; Kalinowski, P.; Michałowski, Ł.; Paluszkiewicz, R.; Ziarkiewicz-Wróblewska, B.; Zieniewicz, K.; Sobieraj, P.; Nowicki, A. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1895–1903. [[CrossRef](#)] [[PubMed](#)]
23. Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **2017**, *29*, 2352–2449. [[CrossRef](#)] [[PubMed](#)]
24. Ertosun, M.G.; Rubin, D.L. Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks. In *AMIA Annual Symposium Proceedings, AMIA Symposium*; American Medical Informatics Association: Bethesda, MD, USA, 2015; Volume 2015, pp. 1899–1908.
25. World Health Organization. Pneumonia Vaccine Trial Investigators' Group & World Health Organization. 2001. Available online: <https://apps.who.int/iris/handle/10665/66956> (accessed on 15 May 2020).
26. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [[CrossRef](#)] [[PubMed](#)]
27. Nadeem, M.W.; Al Ghamdi, M.A.; Hussain, M.; Khan, M.A.; Masood, K.; AlMotiri, S.H.; Butt, S.A. Brain Tumor Analysis Empowered with Deep Learning: A Review, Taxonomy, and Future Challenges. *Brain Sci.* **2020**, *10*, 118. [[CrossRef](#)]
28. Choi, O.; Choi, J.; Kim, N.; Lee, M.C. Combustion Instability Monitoring through Deep-Learning-Based Classification of Sequential High-Speed Flame Images. *Electronics* **2020**, *9*, 848. [[CrossRef](#)]
29. Song, Y.; Zheng, S.; Li, L.; Zhang, X.; Zhang, X.; Huang, Z.; Chen, J.; Zhao, H.; Jie, Y.; Wang, R.; et al. Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT Images; Medrxiv: Guangdong, China, 2020. [[CrossRef](#)]

30. Abràmoff, M.D.; Leng, T.; Ting, D.S.; Rhee, K.; Horton, M.B.; Brady, C.J.; Chiang, M.F. Automated and Computer-Assisted Detection, Classification, and Diagnosis of Diabetic Retinopathy. *Telemed. e-Health* **2020**, *26*, 544–550. [\[CrossRef\]](#)
31. Ting, D.S.W.; Pasquale, L.R.; Peng, L.; Campbell, J.P.; Lee, A.Y.; Raman, R.; Tan, G.S.W.; Schmetterer, L.; Keane, P.; Wong, T.Y. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **2018**, *103*, 167–175. [\[CrossRef\]](#)
32. Devalla, S.K.; Liang, Z.; Pham, T.H.; Boote, C.; Strouthidis, N.G.; Thiery, A.H.; Girard, M.J. Glaucoma management in the era of artificial intelligence. *Br. J. Ophthalmol.* **2019**, *104*, 301–311. [\[CrossRef\]](#)
33. Tan, Z.; Simkin, S.; Lai, C.; Dai, S. Deep Learning Algorithm for Automated Diagnosis of Retinopathy of Prematurity Plus Disease. *Transl. Vis. Sci. Technol.* **2019**, *8*, 23. [\[CrossRef\]](#)
34. Zhang, Y.; Wang, L.; Wu, Z.; Zeng, J.; Chen, Y.; Tian, R.; Zhao, J.; Zhang, G. Development of an Automated Screening System for Retinopathy of Prematurity Using a Deep Neural Network for Wide-Angle Retinal Images. *IEEE Access* **2018**, *7*, 10232–10241. [\[CrossRef\]](#)
35. Oloumi, F.; Rangayyan, R.M.; Ells, A.L. Computer-aided diagnosis of plus disease in retinal fundus images of preterm infants via measurement of vessel tortuosity. In *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015*; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2015; Volume 2015, pp. 4338–4342.
36. Carneiro, G.; Mateus, D.; Loic, P.; Bradley, A.; Tavares, J.M.R.S.; Belagiannis, V.; Papa, J.P.; Nascimento, J.C.; Loog, M.; Lu, Z.; et al. *Deep Learning and Data Labeling for Medical Applications*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; Volume 10008, p. 9. [\[CrossRef\]](#)
37. Brown, J.; Campbell, J.P.; Beers, A.; Chang, K.; Ostmo, S.; Chan, R.P.; Dy, J.; Erdogmus, D.; Ioannidis, S.; Kalpathy-Cramer, J.; et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA Ophthalmol.* **2018**, *136*, 803. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Wang, J.; Ju, R.; Chen, Y.; Zhang, L.; Hu, J.; Wu, Y.; Dong, W.; Zhong, J.; Yi, Z. Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine* **2018**, *35*, 361–368. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Hu, J.; Chen, Y.; Zhong, J.; Ju, R.; Yi, Z. Automated Analysis for Retinopathy of Prematurity by Deep Neural Networks. *IEEE Trans. Med. Imaging* **2018**, *38*, 269–279. [\[CrossRef\]](#)
40. Lee, D.-G.; Jang, Y.; Seo, Y.-S. Intelligent Image Synthesis for Accurate Retinal Diagnosis. *Electronics* **2020**, *9*, 767. [\[CrossRef\]](#)
41. Agarwal, K.; Jalali, S. Classification of retinopathy of prematurity: From then till now. *Community Eye Health* **2018**, *31*, S4–S7. [\[PubMed\]](#)
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015*.
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; pp. 2818–2826.
44. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; pp. 2261–2269.
45. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
46. Altman, D.G.; Bland, J.M. Statistics Notes: Diagnostic tests 1: Sensitivity and specificity. *BMJ* **1994**, *308*, 1552. [\[CrossRef\]](#)
47. Metz, C.E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **1978**, *8*, 283–298. [\[CrossRef\]](#)

