2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Deep Learning Models and Fusion Classification Technique for Accurate Diagnosis of Retinopathy of Prematurity in Preterm Newborn

Nazar Salih [1,2]*, ***Mohamed Ksantini*** [2], ***Nebras Hussein*** [3], ***Donia Ben Halima*** [2], ***Ali Abdul Razzaq*** [4], ***Sohaib Ahmed*** [4]

[1] National School of Electronic and Telecommunications, University of Sfax, Tunisia.
[2] Control and Energies Management Laboratory (CEM-Lab), National Engineering School of Sfax, University of Sfax, Sfax, Tunisia.
[3] Biomedical Engineering Department, Al-Khwarizmi College of Engineering University of Baghdad, Iraq.
[4] Ibn AL Haitham Teaching Eye Hospital, Baghdad, Iraq.
*Corresponding Author.

Received 14/03/2023, Revised 14/07/2023, Accepted 16/07/2023, Published Online First 20/10/2023, Published 01/05/2024

Abstract

Retinopathy of prematurity (ROP) is the most common cause of irreversible childhood blindness, and its diagnosis and treatment rely on subjective grading based on retinal vascular features. However, this method is laborious and error-prone, so automated approaches are desirable for greater precision and productivity. This study aims to develop a deep learning-based strategy to accurately diagnose the plus disease of ROP in preterm newborns using transfer learning models and a fusion classification technique. The Private Clinic Al-Amal Eye Center in Baghdad, Iraq, provided us with 2776 ROP screening fundus images between 2015 and 2020, and the images were used to train three deep convolutional neural network models (ResNet50, Densenet161, and EfficientNetB5). A fusion classifier approach was used to merge the three models for a thorough and precise diagnosis. The three models have relative accuracy rates of 69.78%, 80.57 %, and 81.29 % in their respective classifications. The overall accuracy, however, increased to 90.28 percent when the fusion classifier was employed. This shows that the proposed method helps identify ROP in premature infants. The study's findings imply the proposed method has the potential to significantly enhance the precision and speed with which ROP is diagnosed, which in turn could lead to earlier detection and treatment of the illness and a decreased likelihood of childhood blindness.

Keywords: Artificial intelligence, Deep learning, Fusion classifier, Fundus images, Retinopathy of prematurity.

Introduction

Artificial intelligence (AI) has emerged as a crucial tool in several fields, such as image identification, natural language processing, and autonomous driving. However, creating algorithms that can learn from little quantities of data is one of the most challenging problems in AI. Traditional machine

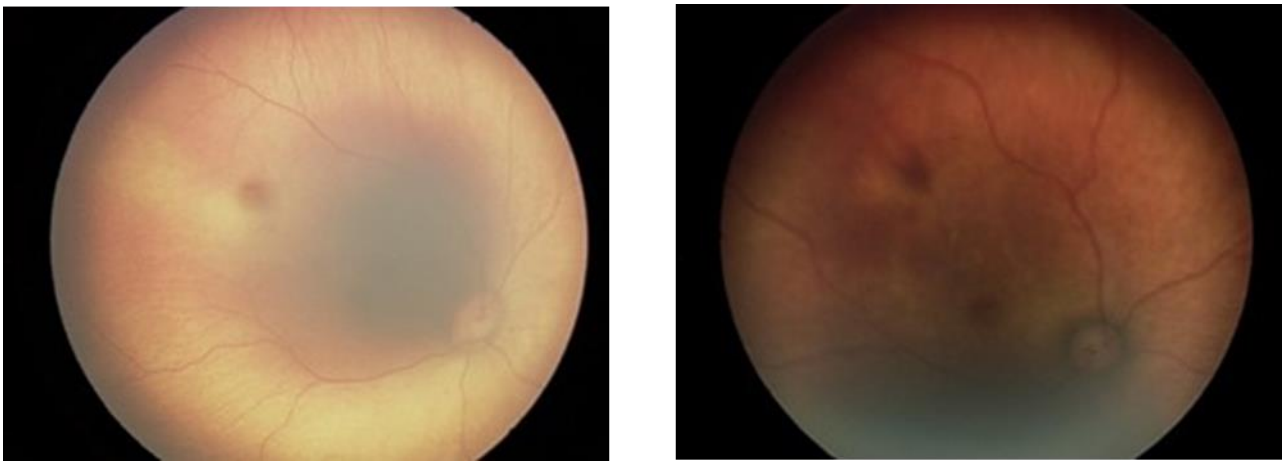
2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

learning algorithms frequently require high accuracy, which is impractical or unachievable in many real-world situations. Few-shot learning, which describes a machine learning algorithm's ability to learn from a few samples of a new class rather than requiring enormous amounts of data, has attracted increasing attention in recent years. Few-shot learning may make AI systems more adept at learning new skills and situations more rapidly and effectively. Few-shot learning continues to be a complex subject despite recent advancements. One of the primary problems is the requirement for more robust and adaptable models that can capture intricate patterns and connections between data points and generalize to new samples. Evaluating various algorithms and track development is also challenging because there aren't enough high-quality few-shot learning benchmarks and evaluation measures. This article provides a new method for few-shot learning that considers some of these issues. Our approach is based on a combination of deep neural networks and metric learning, which enables us to learn extremely graphic representations of data that can be used to compare and categorize fresh samples. Furthermore, a brand-new benchmark for few-shot learning is presented, utilizing a well-known image classification dataset containing many complex few-shot learning tasks. Finally, our strategy is compared to other cutting-edge few-shot learning algorithms using this benchmark and find that it performs better. Retinopathy of prematurity (ROP), which causes abnormal blood vessel growth in the eye and vision impairment, is a significant cause of childhood blindness[1-4]. ROP is one of the deadliest and most serious conditions affecting a preterm baby's eyes[1]. Newborns with low birth weight and gestational age of birth of fewer than 32 weeks are premature retinopathy (less than 1.5 kg)[2]. In Fig 1, the severity of ROP is defined according to the International Classification of Retinopathy of Prematurity (ICROP) guidelines published in 1984[3], 2005[4], and 2021[5]. ROP can also be divided into stages (1-5) and zones (1-3) depending on where the disease is located in the body. Even now, ROP remains the world's leading contributor to childhood blindness today. The diagnosis and treatment of ROP are fraught with difficulties: 1- clinical diagnosis is very different, and even among ROP specialists, there is a lot of disagreement about how to diagnose plus disease; 2- Due to issues with logistics, a lengthy training procedure, a time-consuming exam, and a high risk of misbehavior, not enough ophthalmologists and neonatologists are ready and competent to treat ROP.; and 3- the number of babies with ROP is rising. Because of these problems, Scientists have been attempting to use quantitative and objective methods of diagnosing ROP with computerized image analysis[6]. Although numerous institutions have developed Computer-Based Instructional Aid (CBIA) tools for identifying other diseases in ROP, no automated system has shown that it can diagnose as well as a practicing doctor. Care would be ensured using a fully automated, validated CBIA system to aid professionals in diagnosing. It could also make care more accessible by using automated screening systems on a large scale[7]. Medical image analysis is one area where artificial intelligence (AI) based learning models have found widespread use recently[8, 9]. Intelligent diagnostics methods are now widely used to diagnose various ailments[10, 11]. Deep learning (DL) is the cutting-edge answer for various CBIA issues[12]. Deep Convolutional Neural Networks (DCNN's) have been used to successfully diagnose ROP stages[13], Breast Cancer[14], and diabetic retinopathy[15]. Additionally, Convolutional Neural Networks (CNN's) have been utilized to forecast various prior unquantifiable cardiovascular risk variables from fundus images. They have also shown encouraging results for ROP disease identification at 2 levels[16]. Similarly, transfer learning is a valuable paradigm that applies prior learned information and abilities to a new but associated task. Pretrained models such as ResNet50, Densenet161, and EfficientNetB5 were trained using significant data sources like ImageNet, which has over a thousand categories and 1.2 million natural pictures[17]. These models are constructed from the ground up utilizing considerable computer resources. These models have acquired knowledge of characteristics such as edges, forms, illumination, rotation, and geographical data. This data might be used to extract features from photos in various fields. As a result, the amount of available training datasets is critical for a model to attain excellent efficiency; Low performance or overfitting from training the model on sparse data can be improved by transfer learning. Therefore, transfer learning is helpful for categorization; it boosts a model's generalization ability when the training dataset is restricted (less than a thousand examples)[18]. This method is excellent for detecting images and predicting disease in small datasets like the one used in this study. As classifiers themselves, both sets of experts can be joined to form a new classifier, called a fusion

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

classifier. Classifier results can be expressed as numerical vectors whose size is the same as the number of classes. Therefore, the combination problem may be stated as finding a function that accepts N-dimensional score vectors from M classifiers and outputs N final classification scores, such that misclassification costs are minimized. Increased classification rates in challenging information processing issues can be achieved by combining the decisions of multiple classifiers[19]. It has been discovered that fusing several very simple classifiers rather than creating a single complex classifier is preferable in many situations to get higher recognition rates. The primary goal of our research is to create a deep learning-based system for reliable diagnosis of ROP in premature infants, improving patient outcomes by increasing diagnostic accuracy. Among our primary contributions are:

- Automatic identification of ROP in retinal images by developing a deep learning model (DCNNs).
- A fusion classification method was implemented to enhance the accuracy of ROP plus disease diagnosis. This method integrates the results of many CNN models.
- The method, combining deep learning with a fusion classification algorithm, has the potential to advance significantly the detection and treatment of ROP plus disease in premature infants. This work is significant because it can potentially increase the precision and efficacy of ROP diagnosis, ultimately resulting in better outcomes for premature infants at risk of visual loss. This study may offer a new avenue for creating more efficient diagnostic tools for ROP and other ophthalmologic disorders by utilizing the strength of deep learning and fusion classification algorithms.



(a) **(b)**

Figure 1. Retina images of ROP's normal and plus disease: (a) Normal; (b) Plus disease.

There are five sections in this paper. The research challenge is introduced in Section 1, which emphasizes the value of retinopathy of prematurity (ROP) in preterm neonates and the possibility of deep learning models to enhance the procedure. The related works in computational ROP diagnosis are discussed in Section 2. The study's materials and methods, including dataset collecting, labeling, preprocessing, data augmentation, and the deep learning models utilized in the suggested methodology, are described in Section 3. This section also presents the evaluation measures used to gauge the categorization models' effectiveness. The study's findings are discussed in Section 4 along with the effectiveness of the fusion classifier and each deep-learning model. Section 5 offers a conclusion summarizing the results and discusses additional research.

Related Work

Researchers worldwide have become increasingly interested in studying ROP and creating computer-aided diagnostic techniques for ROP screening. In recent years, numerous deep learning-based ROP diagnosis tools that utilize Retcam imaging modalities have been proposed in the literature. Such as Brown et al.[20] developed and tested a CNN-based DL method for ROP detection at three levels (normal, pre-plus disease, and plus disease). A total of 5511 retinal images were used to train the DCNN.

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Three independent experts graded each image, and one expert made a clinical diagnosis to establish a Reference Standard Diagnostic (RSD) for each image (i.e., normal, pre-plus disease, or plus disease). Three independent experts graded each image, and one expert made a clinical diagnosis to establish a Reference Standard Diagnostic (RSD) for each image. Using 5-fold cross-validation, the technique was examined and verified on an independent sample of 100 images. The Imaging and Informatics in ROP (i-ROP) cohort research collected images from eight academic institutions. Eight ROP experts were used to test the deep learning algorithm. Also, Reed et al.[21] developed a neural network to rate the severity of retinal vascular anomalies on a scale from 1 to 9, allowing for a more precise diagnosis (i-ROP plus score). The overall category of ROP illness was developed based on a consensus model reference diagnostic that included clinical and image-based diagnosis. Experts then assessed the severity of ROP in a second data collection of 100 back pictures. They looked at 870 children's 4861 exams. A normal reference diagnosis of type 1 ROP was found in 155 investigations (3 percent). Moreover, Tan et al.[22] presented a deep learning system that detects ROP plus disease in fundal pictures. There are 3487 images in a local database used for the scheme. One ROP professional graded all the photos in the training collection, and another expert graded the test set. Data augmentation was used to preprocess the images for training to double the images. CNN's were utilized for feature extraction and picture recognition. The pictures were then sorted and labeled as either normal or illness plus. This method was developed by the RMSProp optimizer to increase output by adjusting the system's operating points. Furthermore, Mao et al.[23] conducted automatic diagnosis and quantitative analysis for more diseases. The machine makes a diagnosis choice and quantitatively analyzes the illness's features, allowing doctors to evaluate and make their best decisions. The deep study system used in this research divided the retinal vessels and the Optical Disk (OD). The vessel segmentation process enabled the system to automatically assess essential features such as plus disease, tortuosity, width, fractal dimensions, and vessel density. Hu et al.[24] suggested a CNN architecture for on-the-spot ROP disease diagnosis and severity assessment. ROP is classified into mild, moderate, and severe categories based on the severity of the condition. The suggested structure is made up of two individual networks that are linked by a third network that aggregates features. The primary network is optimized for recognizing fundus pictures and extracting high-level information. The aggregate operator combines these characteristics from several pictures into a single set, which is then sent into the second subnetwork to provide a classification prediction. The model is trained and tested using a sizable dataset captured by RetCam 3. Additionally, Wang et al.[25] used Deep Neural Networks to create an automated ROP detection system called DeepROP (DNNs). ROP recognition was broken down into ROP recognition and ROP scoring. Two explicit DNN models (Id-Net and Gr-Net) were developed for recognition and classifying tasks. Clinical ophthalmologists labeled photos from ROP examinations to create massive datasets of retinal fundus images used to train the DNNs. In another study by Gulshan et al.[26], a retrospective development dataset of 128,175 retinal images was used. The images were graded in 2015 by 54 licensed ophthalmologists and senior ophthalmology residents in the US for diabetic retinopathy, diabetic macular edema, and image gradability. The study utilized a deep convolutional neural network to analyze the retinal images. The resulting method was verified in January and February 2016 using two distinct data sets evaluated by at least seven highly reliable, US-based ophthalmologists. Similarly, in a study by Brown et al.[27], a DCNN model was developed using 5,511 retinal images as training material. A RSD had previously been assigned to each image based on the agreement of image rating by three specialists and clinical diagnosis by one specialist. The RSD categories were normal, pre-plus disease, or plus disease. The model was assessed using 5-fold cross-validation and 100 photos from an independent collection. Eight universities involved in the Imaging and Informatics in ROP (i-ROP) cohort study provided the retinal images used in the study. On the other hand, few authors have employed pre-trained deep learning models to increase the prediction power of the model. Moreover, Zhnag et al.[28] developed a custom transfer learning strategy for DNN classifier training—first, a classifier used in preprocessing weeded out all the unflattering pictures. The pictures were then classified as showing ROP or not by pediatric ophthalmologists. Three Deep Neural Network (DNN) classifiers, AlexNet, VGG-16, and Google Net, were fine-tuned using a labeled training dataset containing 8090 positive and 9711 negative samples. These classifiers

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

were then evaluated on a separate dataset of 1742 samples and compared to the opinions of five pediatric retinal ophthalmologists.

The performance of the classifiers was assessed using metrics such as the Receiver Operating Characteristic (ROC) curve, the area under the ROC curve (AUC), and the precision-recall (P-R) curve. Also, Huang et al. used five types of deep neural networks as transfer learning targets25. After categorizing illness severity, the VGG19 model demonstrated a remarkable accuracy rate of 98.82%, with a sensitivity of 100% and a specificity of 98.41%, in predicting disease severity. To ensure the accuracy of the VGG19 model, the researchers employed 5-fold cross-validation on the datasets and found that it was highly influential in predicting ROP. These findings provide potential fuel for the expansion of computer-aided diagnostics.

Table 1. An analysis of the proposed study about recently conducted research.

Ref.	Year	Dataset	DenseNet	ResNet	EffecientNet	Fusion Classifier	Others
[20]	2018	5511	-	-	-	-	CNN
[21]	2018	4861	-	-	-	-	CNN
[22]	2019	6794	-	-	-	-	CNN
[23]	2020	5711	✓	-	-	-	U-Net
[24]	2017	3017	-	✓	-	-	VGG-16, Inception-V2
[25]	2018	20795	-	-	-	-	Id-Net Gr-Net
[26]	2016	128175	-	-	-	-	Inception-V3
[27]	2018	5511	-	-	-	-	U-Net CNN
[28]	2018	17801	-	-	-	-	VGG-16 AlexNet GoogleNet
[29]	2020	2351	✓	-	-	-	VGG-16 VGG-19 MobileNet Inception-V3
Proposed method	2022	2776	✓	✓	✓	✓	-

The above discussion and Table 1 show that researchers have used transfer learning and CNN-based algorithms for ROP classification are not surprising given their success in various computer vision tasks. Transfer learning allows for pre-trained models on large datasets, significantly reducing the required training data and improving model performance. However, the small size of the datasets used in ROP classification studies can limit model performance. It is well-known in machine learning that model performance is directly proportional to the amount of training data available. Therefore, increasing the dataset size could potentially improve the performance of ROP classification models. Moreover, while individual models have achieved high accuracy in ROP classification, combining multiple models through a fusion classifier could enhance overall performance. This approach has been successful in other classification tasks, such as cancer diagnosis, where various models are combined to increase accuracy and reduce the risk of misclassification. In conclusion, while using transfer learning and CNN-based algorithms for ROP classification is promising, there is still room for improvement in dataset size and model performance. Combining multiple models through a fusion classifier could be a potential solution to address these limitations and improve overall classification accuracy.

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Materials and Methods

Dataset

All images were captured in the Al-Amal Eye Center Private Clinic in Baghdad, Iraq. All images were taken by experienced professionals utilizing a RetCam3. This specialized facility has offered ROP screening services for a long time. Using ROP screening, 2776 fundus images were taken between 2015 and 2020. Several actions were taken to correct possible biases in the dataset. To begin, the dataset was preprocessed to standardize image dimensions and resolution, hence minimizing inter-image differences. Data augmentation techniques, such as random horizontal flipping and rotation, were employed to prevent overfitting and make the training dataset more diverse. The dataset used in this research was collected from various places, including pictures of patients of varying ages, sexes, and ethnicities. As a result, it is crucial to use broad and representative datasets for medical image analysis. Furthermore, it is essential to recognize and deal with potential biases in the dataset during the analysis and interpretation of results.

Labeling

The research includes two senior ophthalmologists with over 15 years of experience treating ROP patients. The fundus pictures were classified as normal or diseased by these specialists. The two ophthalmologists organized the photos individually and then compared them to look for discrepancies in the labeling process to find whether the specialists allocated specific images to various labels. Following discussion among the experts, the labels were eventually organized collectively, and the photos were labeled.

Preprocessing

The fundus images had a resolution of 640 x 480 pixels but were reduced to 224 x 224 when input to our deep learning models. Data from 2,220 patients was utilized for training, including images that were indistinct, fuzzy, or dark were omitted from the research. We investigated fundus pictures showing numerous ROP phases in the same child, guaranteeing no overlap between training and test datasets. The data set was split into three sections: 20% for testing, 70% for training, and 10% for validation. The best model should be chosen based on how well it performs on the validation set to avoid overfitting, as shown in Table 2.

Table 2. ROP plus disease dataset.

Attribute	Normal	Plus disease
Train set (70%)	915	1027
Validation set (10%)	131	147
Test set (20%)	262	294
Total	1308	1468

Data Augmentation

During training, overfitting can occur if the model is trained with insufficient data. To resolve this concern, data augmentation used to supplement the existing training dataset with fresh retinal fundus images. New datasets were created with the help of data augmentation. Several various approaches to image enhancement were tried and tested, including rotation [3, 3], width [0.1, 0.1], height [0.1, 0.1], zoom [0.85, 1.15], and horizontal flip. Seven-fold scaling of the training dataset yielded a massive 22,178 test samples.

Deep learning models

One model was chosen among several backbone models in this work, including Res-Net50 from the ResNet group, DenseNet161 from the DenseNet group, and EfficientNetB5 from the efficient groups. In the end, the fusion models are combined using a fusion classifier to enhance the model's performance. ResNet50: it is a convolutional neural network model with 50 layers. ResNet is a type of Artificial Neural Network (ANN) in which residual blocks are stacked atop one another to build a network. Deep residual nets utilize residual blocks to improve the accuracy of the models further. These skip connections are capable of being used in either way. First, they address the gradient problem by designing a new route for the gradient to follow. With their assistance, the model may also learn an identity function. This ensures that the model's top layers do not experience a performance degradation relative to their lower equivalents. DenseNet161: This model is a member of the DenseNet family of image classification models. The primary distinction between this model and the dense net-121 model is the model's size and precision. The authors translated their Torch-trained

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

models to Caffe format. ImageNet was utilized during the earliest stages of training for each DenseNet model. The input to the model is a blob consisting of a single picture with the values 1, 3, 224, 224 in BGR order. When transferring the image blob into the network, the BGR mean values must be removed as shown: [103.94, 116.78, 123.68]. The values must also be split by 0.017. EfficientNetB5: This model is one of the image categorizations Efficient Net models. TensorFlow was utilized for model training. All Efficient Net models were educated using the ImageNet image database. It is a straightforward and very effective method that permits more principled scaling up of a basic ConvNet to any desired resource constraints while maintaining model efficiency.

Proposed Methodology

These DCNNs models were designated to accomplish our primary objective of recognizing ROP. After receiving the outcomes, a fusion classifier is developed to enhance performance and compare with previously published DCNN models. Data were collected to compare the models' outputs, and the best-performing model was selected for grading illness severity. With our approach, the Keras-provided pre-trained model was further weighted. Classifiers were combined by replacing the model's FC layers with four hidden layers. The size of the first and second FC layers was set to 250 in all DCNNs models, and a 20% drop rate was used for the dropout layers. The drop rate was 20% on the third dropout layer, whereas the third FC layer could hold 128 bytes. After an FC layer for image classification, a final soft max layer was added to classify fundus images. The ReLU activation function is utilized throughout all thick layers. The "Adam" optimizer was selected as the loss function of choice, and categorical cross entropy was used to classify data. A fusion classifier may predict based on the findings' average rather than selecting a single model. This fusion classifier receives the findings or results of all other classifiers and generates a single value 25, 28. Each model (from the prior three classification models) must predict every instance in the dataset for the fusion classifier to be effective in this research. More than half of the predictions for the final output have been disclosed. The model would be categorized using the majority label for its class. After preprocessing, A training set and a test set were created from the available data. Then, the model was refined using the enhanced training data. Following model deployment and testing on the test dataset for classification, we fine-tuned the model's hyperparameters to achieve optimal performance. The model's effectiveness was evaluated based on how effectively it predicted and differentiated between different data types. Furthermore, the problems presented by various models in categorizing ROP plus disease were demonstrated and the area under the curve (AUC) was computed to compare their performances. Finally, the outputs of the three DCNNs models are concatenated to build the fusion classifier. The overall categorization process workflow is depicted in Fig 2.

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

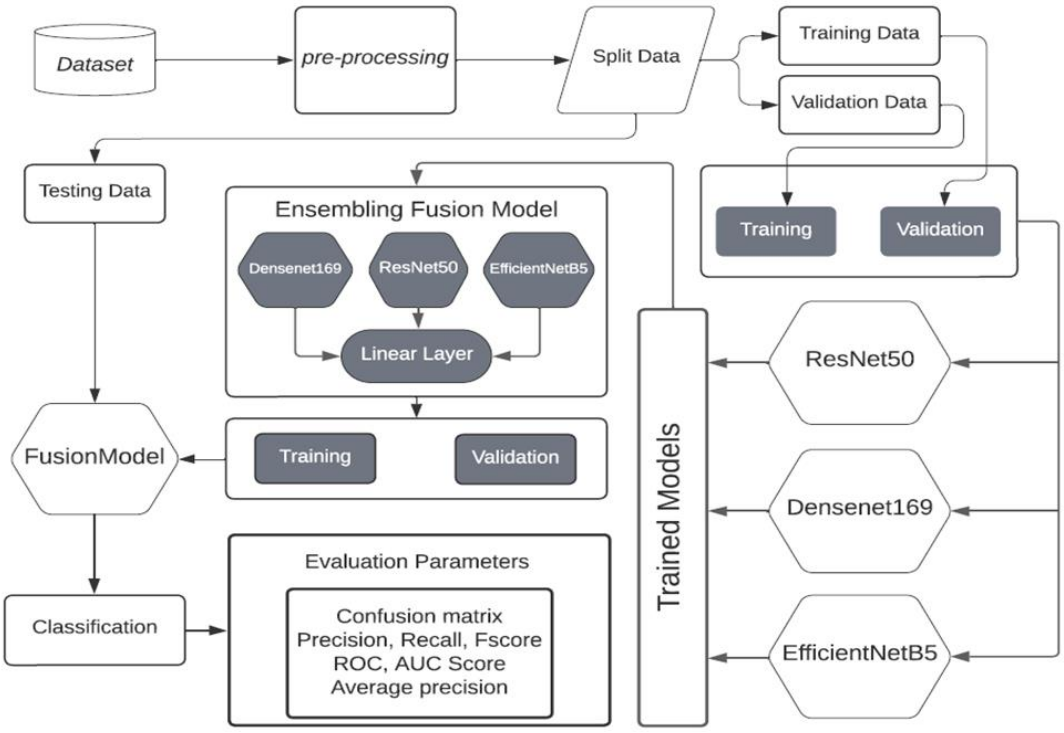


Figure 2. Proposed Model.

The results from the three different deep learning models were fused from the above-proposed method using a fusion classifier (ResNet50, DenseNet161, and EfficientNetB5). The fusion classifier takes the majority vote of the individual classifiers for the label predicted for each input image to make a final prediction. Classifiers generate a probability distribution for each input image over the classes it could potentially belong to. The probability distributions are averaged together, and the most likely category is chosen as the final forecast. A two-stage method was used to implement the fusion classifier in this investigation. Initially, each of the various classifiers made predictions for each input image. The second step was utilizing the fusion classifier to merge the results of the individual classifiers. Although it was not specified, a simple averaging or voting algorithm was likely used for the fusion classifier in this study, as described above.

Evaluation Metrics

Precision, Recall, F-measure, and area under the curve were used to compare each trained model's efficacy. After this investigation, all classifier data was merged into the fusion classifier to improve accuracy. This work assessed the trained model's precision and Recall to categorize the ROP plus disease. Whether these values are actual or not, precision is measured as a propensity toward correctness and discusses how closely two or more quantities are related. Precision is measured as a tendency toward correctness and discusses how closely two or more quantities are related[17]. Accuracy is calculated by dividing the number of true positives (TP) by the combined number of TP and false positives (FP) in Eq 1.

$$\text{Precision}=TP/(TP+FP) \quad \mathbf{1}$$

Recall: helps determine the number of relative duplicate data points in a dataset. Therefore, the situation is different from precision in the Recall. The Recall is calculated by dividing true positives by positive building blocks[18]. The positive category incorporates both the number of false negatives (FNs) and the number of true positives (TPs) in Eq 2.

$$\text{Recall}=TP/(FN+TP) \quad \mathbf{2}$$

F1-Score: The F-measure is utilized to strike a balance between precision and Recall for a fair evaluation of the model's performance on test datasets. F1 is calculated as the harmonic mean (HM) of Recall and precision[6]. The F score has a maximum possible value of 1 and a minimum potential value of 0. The F-measure is typically employed to maintain

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

a balance between Recall and precision. The F1 score equation is provided in Eq 3.

F1 Score=2 × (Precision. Recall)/ (Precision+ Recall) **3**

The area under the curve (AUC): ROC curve is called (Receiver Operating Characteristics). The area under the curve, or AUC, is a crucial performance indicator demonstrating how well the model can distinguish between multiple classes. Remember that the higher this area means, the better the model for detecting[7]. The ROC curve is determined by the true positive rate (TPR) and false positive rate (FPR) which are calculated using Eq 4.

FPR=FP/(FP+TN) **4**

Results and Discussion

The primary purpose of this study is to investigate the relationship between the plus illness of premature birth and the occurrence of ROP. To achieve this objective, three distinct classifiers were trained on the datasets, and precision, recall, F1 measure, and area under the curve were utilized to generate predictions. The results from each classifier were then combined using a fusion classifier to provide an overall accuracy result.

Experimental Setup

The models were trained for 100 epochs with a 256-batch size and a 0.001-per-iteration learning rate. Adam was used as the optimizer, and Cross Entropy Loss as the loss function. Training photos were augmented with data by random horizontal flipping and random rotation to boost the diversity of the training dataset and mitigate overfitting. To train the model quickly, we used a graphics processing unit. Separately, the ResNet50, DenseNet161, and EfficientNetB5 models were trained, and their results using accuracy, precision, recall, and the F1 score were compared. With the three models' highest accuracy, precision, recall, and F1 score, EfficientNetB5 emerged victorious. An Intel Core i7 PC operating at 2.7GHz with 8 GB RAM was used for processing to conduct our analysis. Scikit-learn, a Python-based open-source machine learning software, ran our analysis. To facilitate the generation, sharing, and collaboration of real-time coded reports, images, equations, and narrated text, Google Colab was used, which is free, web-based, open-source software.

Prediction of ROP Plus Disease by DL Models

Early and correct ROP diagnosis is critical for successful treatment. A prompt diagnosis like that would allow for the best possible care. Predictive models such as ResNet50, DenseNet161, and EfficientNetB5 were trained to do this. The algorithm was provided with test case photographs and assessed its performance based on its prediction.

Table 3. Comparing the efficacy of three distinct deep neural networks (DNN) models.

Method	Accuracy	Precision	Recall	F1-Score
ResNet50	69.78	69.7	69.77	69.73
DenseNet161	80.57	81.59	80.01	80.16
EffecientNetB5	81.29	82.38	81.81	81.26

EfficientNetB5 achieved the highest accuracy of 81.29 percent, followed by ResNet50's accuracy of 69.78 percent in our comparison. As shown in Table 3, EffecientNetB5 has the highest accuracy of any of the models, 82.38%, while Res-Net50 has the lowest, at 69.7. EffecientNetB5 also outperforms other models in Recall, with a score of 81.81% compared to VGG-19's lowest score of 69.77. Fig 3 exhibits the performance evaluation using ROC curves, which indicates the consistency of the deep learning models, with EffecientNetB5 showing superior performance than the other models with an F1-Score of 81.26% and ResNet50 scoring the lowest at 69.73.

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

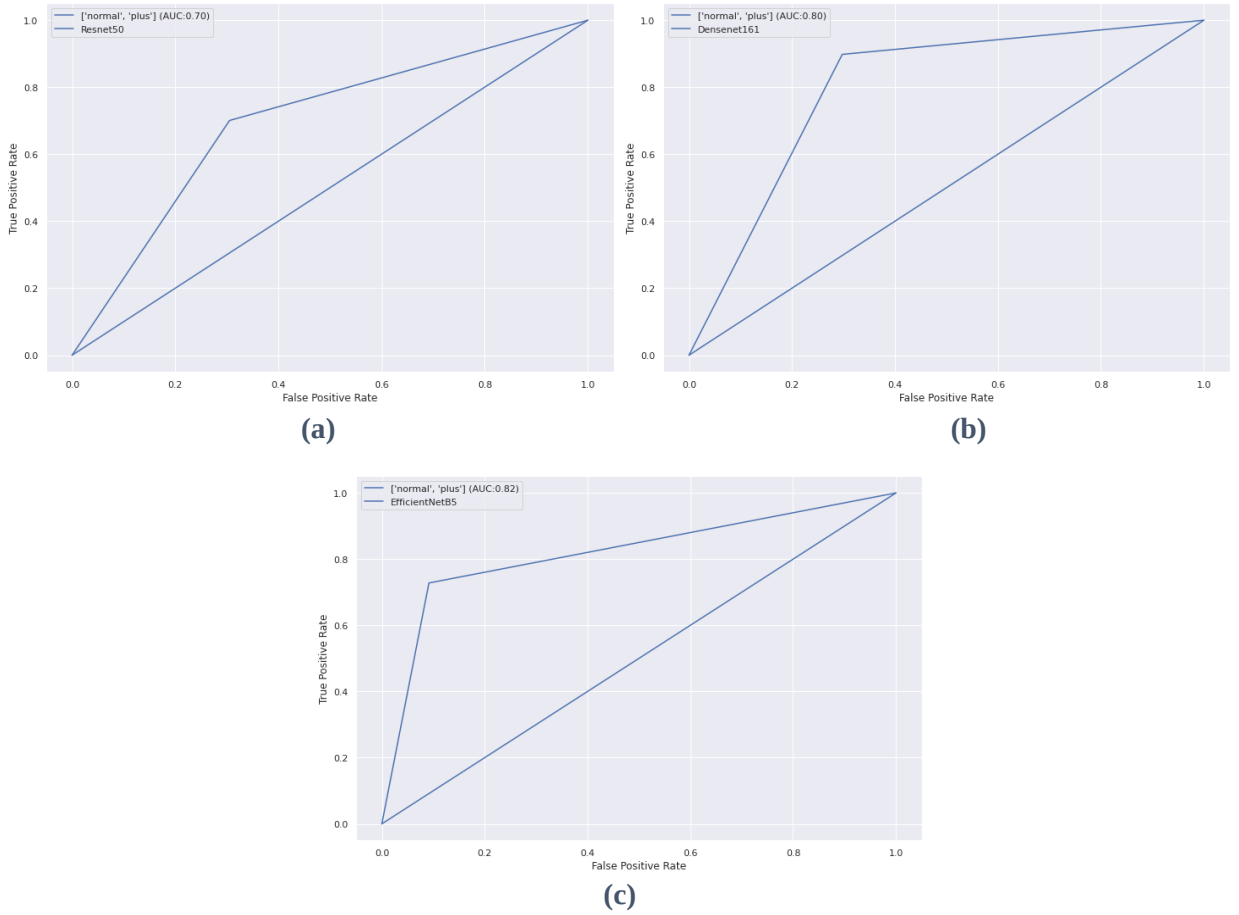


Figure 3. ROC-AUC curves for deep learning classifiers. (a) ResNet50, (b) DenseNet160 (c) EfficientNetB5.

The ensemble approach suggested a fusion classifier in the second stage. Multiple deep-learning classifiers were developed, combining their results to form a single verdict. A fusion classifier, a form of ensemble approach, was utilized in this work. This is accomplished using various matching DL algorithms on the same dataset[30]. In this study, a fusion Classifier was employed, which implies that each classifier (with the help of earlier categorization models) votes for each occurrence. The most recent output prediction has received more than half of all votes. The case would be labeled with the more common class label, and the models would be compared using the results tables.

Table 4. An evaluation of other methods and the proposed method (fusion classifier).

Ref.	Accuracy	Precision	Recall	F1-Score
Pour et al31	72.36	-	-	-
Coyner et al32	87.5	86	76	-
A grawal et al33	84.4	69.6	-	-
Proposed Method (Fusion Classifier)	90.28	90.37	90.15	90.23

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

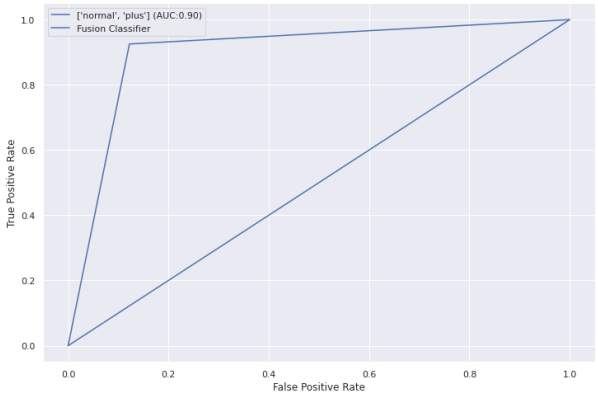


Figure 4. ROC-AUC curves of fusion Classifier.

The results show that the fusion classifier improved the results and provided better prediction, as shown in Fig 4, with precision:90.73, Recall:90.15, F1 Score:90.23, and AUC: 90.15. The fusion classifier technique performs better because it combines two or more DL techniques, and the class that receives the most votes from the classifiers is declared the winner. A detailed comparison of the proposed method with other methods is shown in Table 4. This research suggested a deep learning-based method for detecting ROP in premature infants by fusing different categorization methods. Our approach was 90.28% accurate, better than the state-of-the-art methods in the previous section's literature study. The accuracy rate of the first paper[31] that was examined was improved by our method, which employed a deep-learning strategy for ROP detection. Both studies used deep learning models to diagnose ROP, but our method—which combined three models using a fusion classifier—had a higher accuracy rate. Newer deep learning models, such as the ones utilized in the research, have a reputation for being more effective than their predecessors. A multi-feature fusion method was also used in the second paper[32] to diagnose ROP. Unlike conventional machine learning methods, our approach relied on deep learning models. Our solution also outperformed the multi-feature fusion strategy, indicating that deep learning models are superior for ROP diagnosis. A machine-learning strategy for identifying ROP from color fundus photographs was presented in the next paper[33] examined Even though their approach was quite successful, our research used cutting-edge deep learning models proved more effective than their predecessors in machine learning. By combining numerous deep learning models with a fusion classifier, our method achieved a greater accuracy rate than the machine learning method. Our results show that a deep learning-based approach and a fusion classification technique can successfully identify ROP in premature infants. In comparison to current best practices, our technology has the potential to significantly improve the accuracy and timeliness of ROP diagnoses, which in turn will allow for earlier identification and treatment, reducing the risk of childhood blindness.

Limitations:

The dataset employed in this study is limited and comes from a single medical center in Iraq, which is one of the significant drawbacks. It's possible the results can't be extrapolated to the entire population of preterm babies worldwide. Also, while the proposed models show promise in performance, they may still fall short of entirely replacing human experts. While several symptoms and signs can point to ROP, the proposed models can only analyze fundus photographs. Finally, the proposed models were not validated on external datasets, so the results reported here may not reflect how the models perform on other datasets.

Conclusion

The results of this study show that deep learning approaches can potentially increase the accuracy of ROP diagnosis in premature newborns. The fusion classifier achieved an overall accuracy of 90.28 percent when using the three transfer learning models (ResNet50, DenseNet161, and EfficientNetB5) used in this study. The fusion classifier demonstrated increased performance across all evaluation parameters, including precision, recall, F1 measure, and AUC. The practical benefits of this research include providing a promising method for preterm baby ROP screening, particularly in regions with poor access to ophthalmologists. Automated deep learning models can improve the speed and precision of ROP diagnosis, enabling early disease detection and treatment. It is crucial to recognize the study's shortcomings, which include the dataset's modest size and the absence of external validation using additional datasets. Other deep-learning models and techniques for ROP diagnosis

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

need to be investigated, and more research is required to validate the results of this study on larger datasets. In conclusion, our study has added to the corpus of knowledge about applying deep learning to diagnosing ROP in premature infants. The results indicate that transfer learning models and a fusion classifier can enhance the precision of ROP diagnosis and offer a potential method for screening in regions with poor access to ophthalmologists. To increase the accuracy and effectiveness of ROP diagnosis, future research should concentrate on confirming these findings across more extensive datasets and investigating additional deep-learning techniques.

Acknowledgment

The authors are grateful to the Private Clinic of Al-Amal Eye Center in Baghdad, Iraq, for providing permission to use the data source.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in the Private Clinic of Al-mal Eye Center in Baghdad, Iraq.

Authors' Contribution Statement

Conceptualization, N.S, M.K, and N.H; methodology, N.S, M.K, and N.H.; software, N.S.; validation, N.S, M.K, and N.H.; formal analysis, N.S, M.K, and N.H.; investigation, N.S, M.K, and N.H.; resources, N.H, A.A, and S.A.; data curation, A.A, and S.A.; writing-original draft preparation, N.S.; writing-review and editing, N.S.; visualization, M.K, N.H, and D.B.H.; supervision, M.K, N.H, and D.B.H.; project administration, M.K. All authors have read and agreed to the published version of the manuscript.

References

1. Hong EH, Shin YU, Cho H. Retinopathy of prematurity: a review of epidemiology and current treatment strategies. Clin Exp Pediatr. 2022 Mar; 65(3): 115-126. https://doi.org/10.3345/cep.2021.00773.
2. Palmer EA, Flynn JT, Hardy RJ, Phelps DL, Phillips CL, Schaffer DB, et al. Incidence and early course of retinopathy of prematurity. Ophthalmology. 2020; 127(4 Suppl): S84-S96. https://doi.org/10.1016/j.ophtha.2020.01.034.
3. Committee for the Classification of Retinopathy of Prematurity. An international classification of retinopathy of prematurity. Pediatrics. 1984; 74(1): 127-133. https://doi.org/10.1016/s0031-3955(16)39015-3.
4. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. Arch Ophthalmol. 2005; 123(7): 991-999. https://doi.org/10.1001/archopht.123.7.991.
5. Chiang MF, Quinn GE, Fielder AR, Ostmo SR, Paul RV, Berrocal A, et al. International Classification of Retinopathy of Prematurity, Third Edition. Ophthalmology. 2021 Oct; 128(10): e51-e68. https://doi.org/10.1016/j.ophtha.2021.05.031.
6. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. Arch Ophthalmol. 2007; 125(7): 875–80. https://doi.org/10.1001/archopht.125.7.875.
7. Brady CJ, D'Amico S, Campbell JP. Telemedicine for Retinopathy of Prematurity. Telemed J E Health. 2020 Apr; 26(4): 556-564. https://doi.org/10.1089/tmj.2020.0010.
8. M. Ksantini, A. Ben Hassena and F. Delmotte, Comparison and fusion of classifiers applied to a medical diagnosis. 2017 14th Int Multi-Conf Syst Signals Devices (SSD), Marrakech, Morocco; 2017. p. 211-216. https://doi.org/10.1109/SSD.2017.8166985
9. Mosa Z M, Ghaeb N H, Ali A H. Detecting Keratoconus by Using SVM and Decision Tree Classifiers with the Aid of Image Processing.

Baghdad Sic. J. 2019; 16(4(Suppl.), 1022. https://doi.org/10.21123/bsj.2019.16.4(Suppl.).1022
10. Chakraborty S, Jana GC, Kumari D, Swetapadma A. An improved method using supervised learning technique for diabetic retinopathy detection. Int J Inf Technol. 2019. 12: 473-477. https://doi.org/10.1007/s41870-019-00318-6
11. Hasan A M, Qasim A F, Jalab H A, Ibrahim R W. Breast Cancer MRI Classification Based on Fractional Entropy Image Enhancement and Deep Feature Extraction. Baghdad Baghdad Sic J.2023; 20(1), 0221. https://doi.org/10.21123/bsj.2022.6782
12. Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput. Sci. 2021; 2(3): 420. https://doi.org/10.1007/s42979-021-00815-1.
13. Salih N, Hussein N. Human corneal state prediction from topographical maps using a deep neural network and a support vector machine. Int. J. Curr. Res. 2018; 10(11): 75461-75467. https://doi.org/10.13140/RG.2.2.30320.94726
14. Salih N, Ksantini M, Hussein N, Halima DB, Razzaq AA, Mahmood SA. Detection of Retinopathy of Prematurity Stages Utilizing Deep Neural Networks. 7th Int Cong Info Commun Technolo. 2023: 771-782. Springer. https://doi.org/10.1007/978-981-19-1607-6_62
15. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell. 2022; 15(1): 1-22. https://doi.org/10.1007/s12065-020-00540-3.
16. Barriada RG, Masip D. An Overview of Deep-Learning-Based Methods for Cardiovascular Risk Assessment with Retinal Images. Diagnostics. 2023; 13(1): 68. https://doi.org/10.3390/diagnostics13010068.
17. Ait Nasser A, Akhloufi MA. A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography. Diagnostics. 2023; 13(1): 159. https://doi.org/10.3390/diagnostics13010159.
18. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data. 2022; 9: 102. https://doi.org/10.1186/s40537-022-00652-w.
19. Rothe S, Kudszus B, Söffker D. Does Classifier Fusion Improve the Overall Performance? Numerical Analysis of Data and Fusion Method Characteristics Influencing Classifier Fusion Performance. Entropy (Basel). 2019 Sep 5; 21(9): 866. https://doi.org/10.3390/e21090866.
20. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. JAMA Ophthalmol. 2018; 136(7): 803-810.https://doi.org/10.1001/jamaophthalmol.2018.1934.
21. Redd TK, Campbell JP, Brown JM, Kim SJ, Ostmo S, Chan RVP, et al. Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity. Br J Ophthalmol. 2019; 103(5): 580-584. https://doi.org/10.1136/bjophthalmol-2018-313156.
22. 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(6): 23. https://doi.org/10.1167/tvst.8.6.23.
23. Mao J, Zhang Y, Zhang Y, Guo Q, Zhang L, Dai R, et al. Automated diagnosis and quantitative analysis of plus disease in retinopathy of prematurity based on deep convolutional neural networks. Acta Ophthalmol. 2020; 98(3): e352-e359. https://doi.org/10.1111/aos.14264.
24. 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(1): 269–79. https://doi.org/10.1109/TMI.2018.2865357.
25. Wang J, Ju R, Chen Y, Zhang L, Hu J, Wu Y, et al. Automated retinopathy of prematurity screening using deep neural networks. EBio Medicine. 2018; 35: 361–8. https://doi.org/10.1016/j.ebiom.2018.07.028.
26. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc. 2016; 316(22): 2402–10. https://doi.org/10.1001/jama.2016.17216.
27. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RP, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. J Am Med Assoc. 2018; 136(7): 803–10. https://doi.org/10.1001/jamaophthalmol.2018.1934.
28. Wang J, Ju R, Chen Y, Zhang L, Hu J, Wu Y, et al. Automated retinopathy of prematurity screening using deep neural networks. EBio Medicine. 2018; 35: 361-368. https://doi.org/10.1016/j.ebiom.2018.08.033
29. Huang Y-P, Li Y-C, Liu C-C, Wu J-Y, Huang S-H, Chen Y-J, et al. Deep learning models for automated diagnosis of retinopathy of prematurity in preterm infants. Electronics. 2020; 9(9): 1444. https://www.mdpi.com/2079-9292/9/9/1444.
30. Zaidi SAJ, Tariq S, Belhaouari SB. Future prediction of COVID-19 vaccine trends using a voting classifier. Data. 2021; 6(11): 112. https://doi.org/10.3390/data6110112.
31. Pour EK, Pourreza H, Zamani KA, Mahmoud A, Sadeghi AM, Shadravan M, et al. Retinopathy of prematurity-assist: Novel software for detecting plus disease. Korean J Ophthalmol. 2017; 31(6): 524-532. https://doi.org/10.3341/kjo.2017.0012.
32. Coyner AS, Campbell JP, Ostmo S, Kim SJ, Jonas KE, Paul RV, et al. Machine Learning for Prediction of Retinopathy of Prematurity Fundus Image Quality from Clinical Data. *Investig Ophthalmol Vis Sci*. 2019; 60(9): 1525-1525. https://doi.org/10.1167/iovs.19-27311.

2024, 21(5): 1729-1742
https://dx.doi.org/10.21123/bsj.2023.8747
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

33. Agrawal R, Kulkarni S, Walambe R, Kotecha K. Assistive framework for automatic detection of all the zones in retinopathy of prematurity using deep learning. J Digit Imaging. 2021; 34(4): 932-947. https://doi.org/10.1007/s10278-021-00455-2.

نماذج التعلم العميق وتقنيات التصنيف المدمج للتشخيص الدقيق لاعتلال الشبكية الخداجي عند الأطفال الخدج

نزار صالح[1،2]، محمد قسنطيني[2]، نبراس حسين[3]، دنيا بن حليمة[2]، علي عبد الرزاق[4]، صهيب احمد[4]

[1] المدرسة الوطنية للإلكترونيات والاتصالات، جامعة صفاقس، تونس.
[2] معمل التحكم وإدارة الطاقات (CEM-Lab)، المدرسة الوطنية للمهندسين بصفاقس، جامعة صفاقس، صفاقس، تونس.
[3] قسم الهندسة الطبية الحيوية، كلية الخوارزمي للهندسة، جامعة بغداد، العراق.
[4] مستشفى ابن الهيثم التعليمي للعيون، بغداد، العراق.

الخلاصة

اعتلال الشبكية الخداجي (ROP) هو السبب الأكثر شيوعًا لعمى الأطفال الذي لا رجعة فيه، ويعتمد تشخيصه وعلاجه على الدرجات الذاتية بناءً على سمات الأوعية الدموية في شبكية العين. ومع ذلك، فإن هذه الطريقة شاقة وعرضة للخطأ، لذا فإن الأساليب الآلية مرغوبة لمزيد من الدقة والإنتاجية. تهدف هذه الدراسة إلى تطوير نهج قائم على التعلم العميق للتشخيص الدقيق لمرض اعتلال الشبكية الخداجي الزائد عند الأطفال الخدج باستخدام نماذج التعلم التحويلية وتقنية تصنيف الاندماج. زودنا مركز الأمل للعيون في العيادة الخاصة في بغداد، العراق، بـ 2776 صورة لقاعدة فحص مرضى اعتلال الشبكية الخداجي بين عامي 2015 و2020، واستخدمنا هذه الصور لتدريب ثلاثة نماذج للشبكات العصبية التلافيفية العميقة (ResNet50 وDensenet161 وEfficientNetB5). تم استخدام نهج مصنف الاندماج لدمج النماذج الثلاثة من أجل تشخيص شامل ودقيق. تتميز النماذج الثلاثة بمعدلات دقة نسبية تبلغ 69.78٪ و80.57٪ و81.29٪ في تصنيفاتها الخاصة. ومع ذلك، زادت الدقة الإجمالية إلى 90.28 في المائة عند استخدام مصنف الاندماج. هذا يدل على أن الطريقة المقترحة مفيدة لتحديد اعتلال الشبكية الخداجي عند الخدج. تشير نتائج الدراسة إلى أن الطريقة المقترحة لديها القدرة على تعزيز الدقة والسرعة التي يتم بها تشخيص اعتلال الشبكية الخداجي بشكل كبير، مما قد يؤدي بدوره إلى الكشف المبكر عن المرض وعلاجه وتقليل احتمالية الإصابة بعمى الأطفال.

الكلمات المفتاحية: الذكاء الاصطناعي، التعلم العميق، مصنف الاندماج، صور قاع العين، اعتلال الشبكية الخداجي.