The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2634-1964.htm

Detecting and staging diabetic retinopathy in retinal images using multi-branch CNN

Worapan Kusakunniran and Sarattha Karnjanapreechakorn
Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand
Pitipol Choopong
Department of Ophthalmology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
Thanongchai Siriapisith
Department of Radiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, and
Nattaporn Tesavibul, Nopasak Phasukkijwatana, Supalert Prakhunhungsit and Sutasinee Boonsopon
Department of Ophthalmology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

Received 10 June 2022
Revised 23 July 2022
14 October 2022
9 November 2022
Accepted 15 November 2022

Abstract

Purpose – This paper aims to propose a solution for detecting and grading diabetic retinopathy (DR) in retinal images using a convolutional neural network (CNN)-based approach. It could classify input retinal images into a normal class or an abnormal class, which would be further split into four stages of abnormalities automatically.

Design/methodology/approach – The proposed solution is developed based on a newly proposed CNN architecture, namely, DeepRoot. It consists of one main branch, which is connected by two side branches. The main branch is responsible for the primary feature extractor of both high-level and low-level features of retinal images. Then, the side branches further extract more complex and detailed features from the features outputted from the main branch. They are designed to capture details of small traces of DR in retinal images, using modified zoom-in/zoom-out and attention layers.

Findings – The proposed method is trained, validated and tested on the Kaggle dataset. The regularization of the trained model is evaluated using unseen data samples, which were self-collected from a real scenario from a hospital. It achieves a promising performance with a sensitivity of 98.18% under the two classes scenario.

Originality/value – The new CNN-based architecture (i.e. DeepRoot) is introduced with the concept of a multi-branch network. It could assist in solving a problem of an unbalanced dataset, especially when there are common characteristics across different classes (i.e. four stages of DR). Different classes could be outputted at different depths of the network.

Keywords Diabetic retinopathy, Abnormality detection, Staging, Side-branches CNN

Paper type Full length article

© Worapan Kusakunniran, Sarattha Karnjanapreechakorn, Pitipol Choopong, Thanongchai Siriapisith, Nattaporn Tesavibul, Nopasak Phasukkijwatana, Supalert Prakhunhungsit and Sutasinee Boonsopon. Published in *Applied Computing and Informatics*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-06-2022-0150

1. Introduction

Diabetic retinopathy (DR) is one of the most commonly seen complications of diabetes. It could lead to blindness, especially when it is left untreated. DR is diagnosed into four stages (i.e. Stage 1 to Stage 4). Therefore, this could be considered a problem domain of classification on retinal images with five classes of four stages and one normal class (Stage 0). In addition, this paper also considers a classification domain with two classes for differentiating normal cases from DR cases of any stage.

This paper focuses on the challenges of detecting and grading DR in retinal images [1]. The proposed method addresses the difficulty of grading five stages of DR, where its traces could be tiny, especially in the early stage of the disease. The modified zoom-in/zoom-out augmentation with attention layers is then deployed to solve the problem. In addition, in a general case of classification, final decisions of all classes are made at a final output layer. However, particularly for this research question, some classes may be more straightforward than others. Thus deeper learning could lead to the overfitting problem. The proposed solution then allows different classes to be exited for decision-making at different levels of a CNN architecture. This technical contribution could improve the grading of DR in retinal images.

1.1 Literature review

In this section, our literature review is structured into two parts regarding two main ways of solving the problem of DR detection and staging.

1.1.1 The first type of approach: segmenting/detecting traces of DR for DR detection and staging. The first way is to detect/segment traces of DR. Microaneurysms are detected for identifying Stage 1, and exudate is segmented for identifying Stage 2 [2]. In addition, Stage 3 and Stage 4 could be identified using abnormalities of retinal blood vessels. For example, the method proposed by Ref. [3] segmented patches of microaneurysm in retinal images using the autoencoder-regularized neural network, while the feature-transfer network with local background suppression was proposed by Ref. [4] for microaneurysm detection. The microaneurysm is the earliest signal of DR, whose size is tiny (i.e. less than 2% of the entire image's size). Compared with CNN-based solutions, the segmentation-based solutions could achieve higher performances for identifying Stage 1 of DR.

In addition, there have been several publications of exudate segmentation/detection for identifying Stage 2 of DR. For example, using a conventional solution, a hybrid solution of instance learning (iterative graph cut) and supervised learning (neural network) was proposed for the segmentation [5]. Recently, the dual-branch network-based solution was proposed by Ref. [6], where one branch was designed to focus on a large-sized exudate, while another branch was designed to focus on a small-sized exudate. The paper [7] introduced the CNN-based solution emphasizing super-pixel multi-feature extraction. This technique also focused on solving a small-size challenge of segmentation. The exudate segmentation seems to be also crucial for the DR detection in Stage 2, due to a small-sized trace. For Stage 3 and Stage 4 [2], there was a paper [8] that attempted to detect hemorrhage for detecting an abnormality in diabetic patients. The solution applied the modified VGG19 to extract image features before using the extreme learning machine for pixel-based hemorrhage detection. However, based on our literature reviews for Stage 3 and Stage 4, it is more popular to work on stage classification directly instead of segmentation.

1.1.2 The second type of approach: image-level output solutions for DR detection and staging. The second way is to directly apply machine learning techniques such as CNN for classifying stages and abnormalities in retinal images. There are two main types of existing solutions: (1) classifying into two classes of normal and DR and (2) classifying into five classes of normal and four classes of four stages.

In the first type of the second way, for example, the paper proposed in Ref. [9] did not rely on deep learning-based techniques. A fusion of textural and ridgelet features was learned using Sequential Minimal Optimization (SMO) to classify DR. Similarly, in Ref. [10], a fusion of handcrafted features was also used but learned using Darknet53. Differently, the paper in Ref. [11] proposed a solution of feature extraction based on six convolutional layers. Then, SVM, AdaBoost, Naive Bayes, Random Forest and J48 were attempted to classify retinal images into normal or DR classes. In the work proposed by Ref. [12], the multi-task-based CNN was developed with three decoders: classification head, regression head and ordinal regression head. The regression head could be further used for the cut-off into multiple stages of DR. The method by Ref. [13] also relied on the CNN-based solution. In addition, unsharp masking was applied to enhance retinal images. Two channels were fed as an input of the CNN, including the green and entropy channels. Moreover, by Ref. [14], the well-known and pre-trained network of Inception-ResNet-v2 with an additional block of CNN layers was transferred for detecting DR in retinal images.

In the second type of the second way, retinal images are classified into different stages of DR. By [Ref. [15], the CNN-based solution was developed to identify intricate features for classifying stages of DR, such as microaneurysms, exudate and hemorrhages. The method proposed by Ref. [16] also developed the solution based on CNN. To enhance the performance of the stage grading, it applied the distances between stages of DR into loss function. The methods introduced by Refs [17–21] were also developed based on the newly designed CNN architectures for the DR staging. While the method proposed Ref. [22] applied well-known CNN architectures, including Resnet50, Inceptionv3, Xception, Dense121 and Dense169, for DR staging. Also, well-known Inception-v3, ResNet50, InceptionresNet50 and Xception were attempted by Ref. [23] for DR grading.

In existing works, different stages of DR were outputted from different output nodes in the final layer of the network. However, in this paper, the proposed CNN architecture is developed base on the assumption that different stages of DR are identified by different characteristics which should be extracted from retinal images at the different feature levels. This differs from a multi-level feature concept proposed in Ref. [24], where two levels were applied. The first level was a fusion of conventional image descriptors, including SIFT and GIST. While the second level referred to features extracted using CNN on the fused features from the first-level features. In contrast, the multi-level features proposed in our paper refer to features extracted at multiple depths of the CNN architecture.

1.2 Background knowledge: multi-branch network

1.2.1 Motivation. The main network of our proposed solution is developed based on the multi-branch network. It is a combination of sequential branches consisting of a convolutional layer, pooling layer and a fully connected layer. Even though the sequential CNN network can perform well on some problem domains, some tasks still get poor results, e.g. DR stage classification, people re-identification and medical image segmentation.

For example, in Ref. [15], the authors proposed a solution based on sequential CNN, a stack of convolution layers and three fully connected layers. Their proposed network is deep but not wide and has a large number of training parameters. Unfortunately, the reported result was poor in some DR's stages. This is one of the reasons why the multi-branch network is developed here in our work to overcome such complex tasks. These tasks require complex layer structures which can extract small and sparse features, i.e. micro-aneurysm, hemorrhage and small blood vessels in DR cases.

1.2.2 Existing networks. In addition, the method proposed by Ref. [25] was based on a multi-branch network for hyperspectral image classification. Typically, a hyperspectral remote sensing image (HSI) has a large data volume and high spectral resolution, with limited

labeled data and a small training dataset. This makes the classification very challenging. Therefore, they proposed the multi-branch fusion CNN-based network to overcome such problems. Instead of making one sequential network that goes deeper and wider which can lead to the high complexity of the network with a large number of parameters, they added additional branches. This technique provides excellent classification results on the training with the small-sized dataset. This is one of the multi-branch network benefits to extract very small features efficiently and to be convergently trained by a small dataset. So, the multi-branch network indicates that it is suitable for the DR stage classification problem.

One of the multi-branch network strengths is a high performance on small-sized feature extraction. For instance, the method by Ref. [26] proposed the LadderNet, which is a chain of multiple U-Net [27]. The purpose of the LadderNet is the same as U-net, semantic segmentation, but for better capability. In their experiments, the DRIVE dataset [28], a retinal dataset for blood vessel segmentation, was used in the evaluation. The segmentation results show that the LadderNet outperformed the previous networks, where U-net was one of them due to the multi-branch structure. This was because the LadderNet had the shared-weights residual block technique, which was the weight sharing among the branches. This technique significantly reduced the number of LadderNet's parameters.

Another example from Ref. [29], their experiments focused on multiple sclerosis lesion segmentation. Their proposed CNN included a multi-branch downsampling path which enables the network to encode information from different sources. Each branch of the network was the Resnet network [30]. Information on each branch was combined at each step of the encoding process with a filter size of 64, 256 or 512. So, the network could get more information than a single straight network, leading to more accurate segmentation. Therefore, their solution was among the best solutions for the ISBI challenge. These are examples of the multi-branch network key performance on small feature extraction.

The CNN architecture of a single straight branch structure has a stack of convolutional layers with different filter sizes and on top of the fully connected layers [31]. For simple problems, this CNN handles just fine. However, the result was not good on the complex tasks, e.g. in the DR stage classification [15], especially on Stage 1 and Stage 3 classification. This experiment indicated that a standard sequential CNN could not handle the DR stage classification problem.

Each branch receives a separate input in the CNN architecture of a multi-branch structure (e.g. Siamese network). Then, features generated from multiple branches are concatenated together at the end of the network. So, the final output will come from the concatenated features of the two branches. Since the multi-branch structure takes multiple inputs from multiple branches, this affects the training duration time. The network converges into the input dataset faster when compared with the one straight branch network, which has the same length of convolution layers.

So, this advantage of multi-branch can be used to add more branches into the network as long as a graphic card has enough memory. There is still one more point that should be mentioned in the multi-branch structure. Its CNN architecture has no connection between branches and no weight sharing. As noted by Ref. [29], the CNN architecture excepts at the end, which is the feature concatenation step. Even though the weight-sharing technique can make the network converges faster and deeper. But sometimes, it can cause confusion between branch's weights to the network if the structure of each branch is very different. This issue can be fixed easily by changing the structure of each branch to be the same. However, the network will lose its complexity and cannot achieve high performance. Therefore, the proposed CNN in this paper, DeepRoot, aims to overcome the multi-branch problem by changing the structure of normal multi-branch CNN and keeping its complexity.

1.2.3 Multi-branch applied to the proposed solution. Our proposed CNN architecture, DeepRoot, comprises one main branch and two side branches. The main branch is designed

for extracting the base features from retinal images. Then, connected from the main branch, it is split into two side branches designed using different details and purposes. The detailed technical explanations are described in Section 2 of the proposed method. The outputs of different stages are defined at different branches of the network. The proposed CNN architecture is trained and validated with the Kaggle dataset [32]. Then, the trained model is tested with the testing Kaggle dataset and unseen samples of self-collected retinal images from the real scenario of a hospital.

The main novelty of this paper is to propose a concept of combining outputs from multiple learned side branches for classifying each DR class independently. In addition, a zooming structure is also proposed for the main CNN structure for capturing small details of distinguishing DR classes. The validating process is also performed on cross-datasets where a test set was collected from real-world cases of a hospital.

The rest of this paper is organized as follows. Details of the proposed method are described in Section 2. Experiments and results are discussed in Section 3. Then, conclusions are drawn in Section 4.

2. Proposed method

This section explains the details of the proposed method. The proposed CNN architecture is mainly introduced to train a model for classifying retinal images into five classes of 1 normal class and 4 abnormal stages. The training retinal images with labels are fed into the architecture for learning the model. Therefore, this section mainly explains the details of the proposed CNN architecture, as described in the subsections below. In addition, some related supplementary materials of additional figures are located at https://github.com/worapanda/ACI_DiabeticRetinopathy-.git

2.1 DeepRoot network

In this paper, the proposed network, DeepRoot, is developed based on a combination of the main branch and two side branches. As shown in Figure 1, it could be seen why it is called DeepRoot. The original DeepRoot network comes from advanced convolution neural networks nowadays that have multiple branches. In addition, these branches will typically be concatenated together at some point in the network or the end. Even though it can increase the network performance, it could also waste an opportunity that this extra information can be used for another classification. Also, sometimes the extra classification can make the

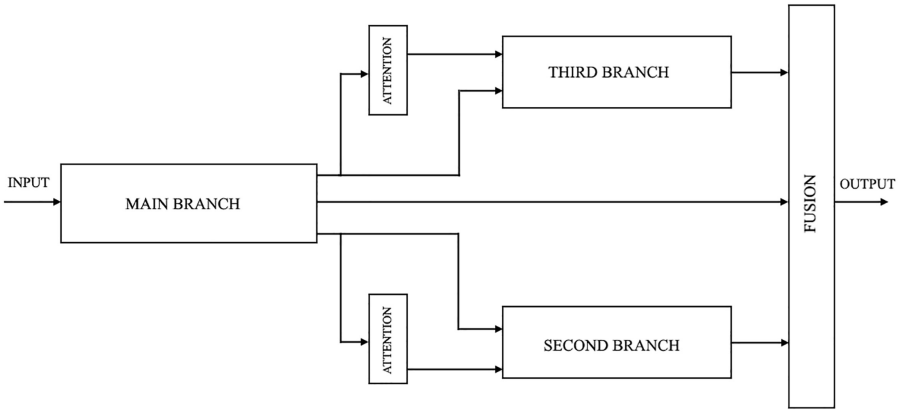


Figure 1.
The architecture of the proposed DeepRoot network

network converge quickly or be used as a combination of the classification. The structure of the DeepRoot network, as shown in Figure 1, consists of one main branch and two side branches. Descriptions of the three branches and their parameters are listed in Table 1.

In addition, in the proposed solution, each of the three branches (i.e. one main branch and two side branches) generates its own output classes. Therefore, in Figure 1, the fusion is referred to as score-level fusion, instead of feature-level fusion as in other existing methods of the multi-branches network. The learning loss is a combination of the three outputs of the three branches. More details on each branch are explained in the following subsections.

2.1.1 Main branch. The main branch of the proposed DeepRoot network acts as the primary feature extractor in both high- and low-level features. Then, the extracted features are passed to the side branches for more complex feature extraction. Thus, the structure of the main branch can be easily changed or even replaced by other well-known CNN architectures. This makes the DeepRoot network flexible and easily adapts to various problem domains. In this proposed method, the EfficientNet [31] is used as the main branch. At the final stage, an output from the main branch will be concatenated with the side branch's outputs. This can prevent diminishing gradients due to the extra information from the top layer.

2.1.2 Side branch. Each side branch of the DeepRoot network does not have to be the same structure or do the same things. One of the advantages of the multi-branch network with multiple outputs, like the DeepRoot network, is flexibility. Because each side branch can have a different shape due to its distinct purpose, it can be designed to do various tasks such as extracting finer feature detail, up-sampling feature size or grouping feature for global information. The DeepRoot network consists of two side branches designed to extract finer feature detail via a zoom-in/zoom-out [35] module and collect dominant features via an attention layer.

2.2 Zoom-in/zoom-out

In retinal images, many signs of diseases are very small, especially for DR. This can make the convolutional neural network struggle to converge on this data type. So, an additional module is needed for the network to achieve high performance. This is where a zoom-in/zoom-out module is added to the proposed network. The structure of zoom-in/zoom-out module is shown in Figure 2 (Left).

As shown in Figure 2 (Left), features extracted from this zoom-in/zoom-out block are the concatenation of features before and after the zooming process. This zooming process consists of four steps: zoom-in, convolution, zoom-out and convolution. The structure of the zoom-in/zoom-out module, as introduced by Ref. [35], is the process to re-size the feature. The zoom-in is for low-level feature extraction, whereas the zoom-out is for up-sampling back to the original size. This lets the network learn the low-level feature information and pass this extracted information to the up-sampling process. Then, the extracted data is re-sized to the original size for concatenation with the shortcut feature.

From this procedure, the network has extra information on tiny-size features for the classification at the terminal. But the typical structure of zoom-in/zoom-out still cannot converge well on the DR problem, especially in Stage 1 and Stage 2 of the disease. This is because features that define the symptoms in Stage 1 and Stage 2, i.e. microaneurysm and hard exudate, are hard to see due to their tiny sizes.

Table 1.
Descriptions of the three branches of the proposed architecture and their parameters

Parts in the proposed solution	Deployed architecture	Number of parameters
Main Branch	EfficientNet	12 M
Second branch	Inception Module-A with Zoom	2.5 M
Third branch	Inception Module-A with Zoom	2.5 M

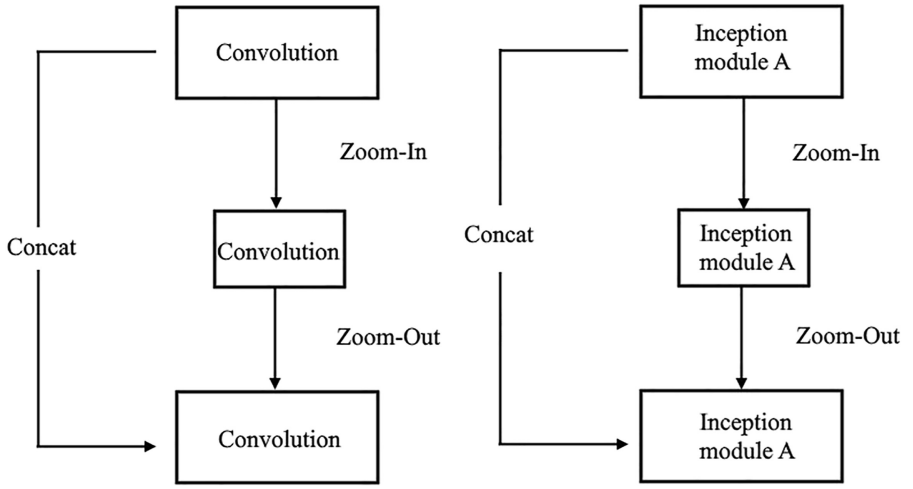


Figure 2.
The structure of zoom-in/zoom-out module (Left), and the improved structure of zoom-in/zoom-out module (Right)

Therefore, the proposed DeepRoot network improves the structure of zoom-in/zoom-out by changing the convolution2d layer to the Inception module A [36]. As a result, the proposed zoom-in/zoom-out is shown in Figure 2 (Right). This adds complexity to the module and has a high performance on small-sized features. The Inception module A is used inside the Inception V3 and V4, which are state-of-the-art classification models. By applying this module, the zoom-in/zoom-out has more receptor fields for extra signals.

2.3 Attention layer

To detect DR, only parts of DR' traces are helpful in the detection process. Thus, irrelevant information in retinal images must be ignored in the learning and inferencing processes. Attention layers [16, 37] would help emphasize traces of the disease, which would be learned from common features seen across input data samples. Particularly, in Stage 3 and Stage 4 of DR, it contains lots of small new blood vessels and various diffuse patterns. This type of layer can be used for enhancing the performance of detecting Stage 3 and Stage 4. Also, it could be a benefit for separating Stages 3 and 4 from Stages 1 and 2. In the proposed DeepRoot network, attention layers are applied after the main branch's output and before the two side branches, as shown in Figure 1. The adopted attention layers contain three sequential sets of normalization, ReLu and convolutional layers.

2.4 Output fusion

Since the DeepRoot network contains three outputs from three different branches, the classification process needs an extra step to combine all three outputs. Unlike other well-known CNN that combine branches together at some point in the network, the DeepRoot network uses all three outputs in the classification process with pre-defined fusion conditions. These conditions combine all outputs into one final output. The conditions can be changed for specific problems. The conditions for the DR stage classification are shown in Table 2.

3. Experiments and discussions

In the experiments, two datasets have been used in the evaluations. The first dataset is a well-known public dataset from Kaggle's DR competition dataset. The second dataset is a sell-collected dataset from the real environment of a hospital.

The Kaggle DR dataset is very popular due to its purpose for the competition. Also, its large sized-dataset can be trained for a complex CNN. The training and testing datasets distributions are shown in Figure 3. Figure 3 indicates unbalanced datasets in both training and testing sets. In addition, Stage 4 images have the least amount of training and testing datasets. So, the numbers of images used in the experiments are 700 and 1,200 for training and testing processes, respectively, for each class. Then, 10% of the selected images of the training dataset are separated for a validating process. For noise and irrelevance details reduction, the datasets are applied with pre-processing techniques, including a low-pass filter on a green channel and central cropping. Figure 4 shows examples of original and pre-processed images. It could be noticed that the details of each image are reduced after the pre-processing step. However, the irrelevance information in each image is also significantly reduced. In the fourth column of Figure 4, some details of hemorrhages are removed (i.e. color information), but corresponding patterns of the lesion were remained in the image.

In addition, our self-collected dataset contains a limited number of images compared with the Kaggle dataset. However, it includes images from a real scenario in a hospital. The distribution of all stages is shown in Figure 3. This dataset is used for testing only to validate the generalization of the trained model when it must be applied to unseen data samples in the real scenario. Similarly, the pre-processing technique is applied to the Kaggle dataset, where sample images are shown in Figure 4.

These fundus images were captured with a 45°–50° field of view (FoV) and converted into a jpeg file format. The original size of images from the self-collected dataset is 1,604 × 1,206,

Table 2.
The conditions for outputs combination at the classification process

Conditions	Final output
Second branch output = Stage 0	Stage 0
Third branch output = Stage 3 or Stage 4	corresponding Stage 3 or Stage 4
Main branch output = Stage 1 or Stage 2	corresponding Stage 1 or Stage 2
Third branch output = Stage 0	Stage 0
Not in any conditions	uses third branch output

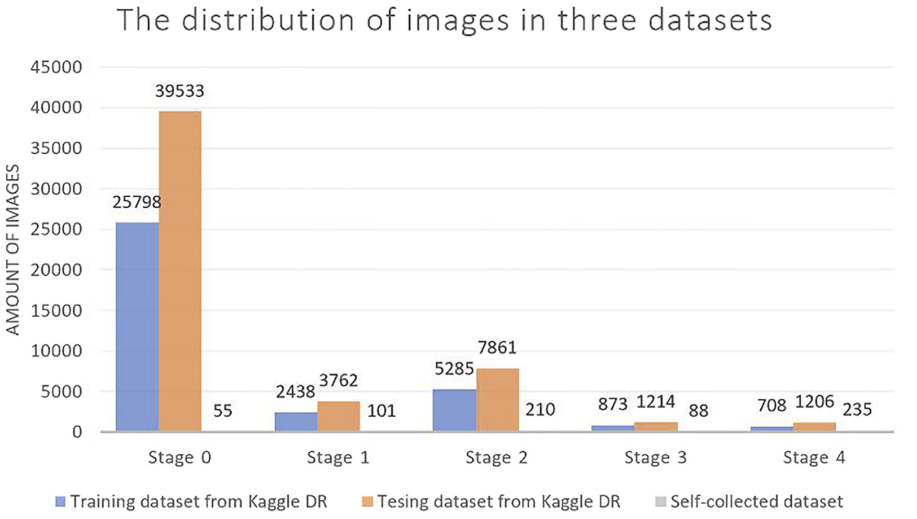
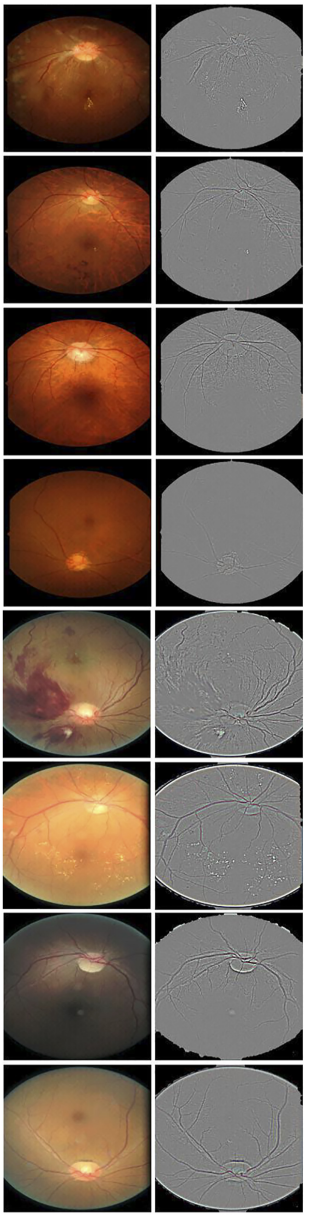


Figure 3.
The distribution of images in three datasets across five stages



Note(s): The first four columns are from in Kaggle dataset and the last four columns are from the self-collected dataset and the last four columns are from the self-collected dataset

Figure 4.
The examples of original retinal images (top row) and their corresponding pre-processed images (bottom row)

while the original size of images from the Kaggle DR dataset varies from 433 × 289 up to 5,184 × 3,456.

In the experiments, the proposed CNN network is trained using four GPUs of Nvidia A100 with VRAM of 40 GB and system memory of 1,024 GB. The proposed network is trained using the Kaggle training dataset for 300 epochs, with an image's size of 1500 × 1500 pixels. The training time is approximately 21 h. Adam Optimizer is used in the optimization process, where a batch size is 16, and a learning rate is dynamically adjusted, starting from a value of 0.001 and reduced by 1/10 in a period of epochs. Also, early stopping and data augmentation are applied to prevent overfitting. The augmentation includes vertical flips, linear contrast and rotation.

Then, the Kaggle validation dataset is used at the end of every epoch for measuring the model performance. The epoch checkpoint with the highest performance is chosen as the final model. Finally, the Kaggle test dataset and self-collected dataset are used for the testing. A confusion matrix is used to demonstrate the corresponding results. It is a tool to show correct and incorrect classification results to explain which classes testing images are miss-classified [38].

In Figure 5a, it could be seen that the trained model can achieve good performances in Stage 0 (normal) and Stage 4, where there is no disease at all in Stage 0 and a significant trace of disease in Stage 4. It obtains moderate performances in Stage 2 and Stage 3, where traces of the disease are still sufficiently significant and could be noticed. In contrast, it achieves a low performance on Stage 1 since it could be noticed using traces of microaneurysm, which is very small (less than 2% of the total size of a retinal image). Particularly, the used dataset from Kaggle seems to be very difficult since it contains many challenges due to its first objective of being used in the competition. Compared with the recent technique using the same dataset on the five-classes scenario, the proposed solution achieves accuracies of 0.83, 0.19, 0.53, 0.48 and 0.72. In contrast, the method in Ref. [34] achieves the accuracies of 0.98, 0.54, 0.84, 0.35 and 0.29 for Stages 0–4, respectively. The proposed solution outperforms for detecting Stages 3 and 4.

Then, Figure 5c shows the confusion matrix of classifying two classes (normal and abnormal classes). The sensitivity is 79.64%, whereas the specificity is 83.08% for unseen data samples. However, the trained model of the proposed CNN architecture could perform

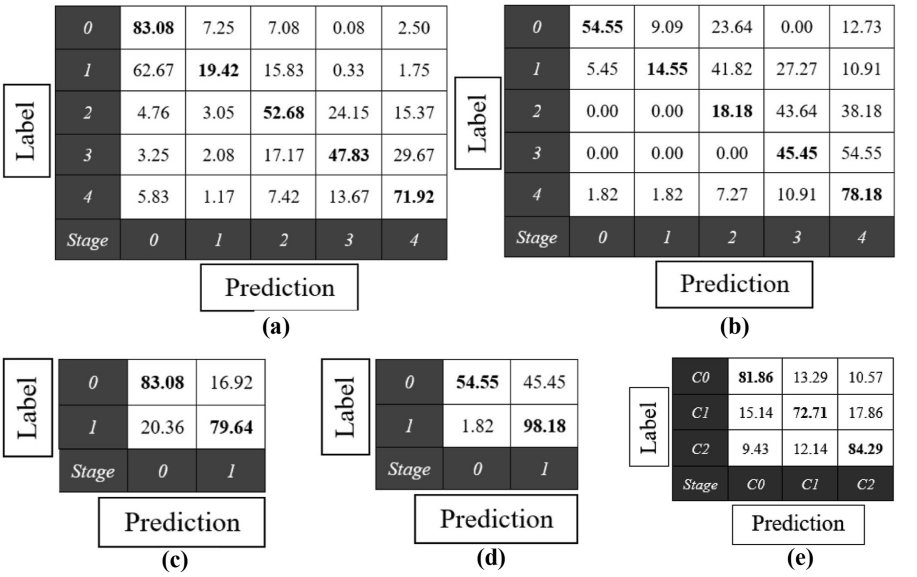


Figure 5.
Confusion matrix (%) on: (a) the Kaggle test dataset of five classes, (b) the self-collected dataset of five classes, (c) the Kaggle test dataset of two classes (normal and abnormal), (d) the self-collected dataset of two classes (normal and abnormal), (e) the self-collected dataset of three classes (C0 contains Stage 0, C1 contains Stage 1 and Stage 2, and C2 contains Stage 3 and Stage 4)

better on unseen samples that were collected from the real scenario in a hospital. The results are shown in Figure 5b.

In addition, the ablation study is performed on a scenario of two classes (normal and abnormal) based on the Kaggle dataset. Four components of the proposed solution are investigated. The experimental results are shown in Table 3. It is proved that all additional three components of side-branch, zoom and attention can enhance the performance on top of the main branch. So, all components of the proposed solution are applied for the rest of the experiments.

In Figure 5b, the trained model is tested with unseen data samples, which were collected from a real scenario in a hospital, where test images were classified into five classes, including a normal class and four abnormal classes of four stages. Then, Figure 5d shows the confusion matrix of classifying two classes (normal and abnormal). The sensitivity is 98.18%, where the specificity is 54.55 % for unseen data samples. Compared with testing on Kaggle, the trained model could detect the abnormality (i.e. high sensitivity) better on the unseen data of the self-collected dataset. This could be because the retinal images collected from a hospital are of better quality when compared with the retinal images in the Kaggle dataset designed for a competition.

The trained model developed in this paper is also compared with other existing techniques under both scenarios of two classes and five classes with four stages. The results are shown in Table 4. For the two-classes scenario, the performances are reported in terms of sensitivity, specificity and accuracy. While for the five-classes scenario, the performances are reported using a weighted average of accuracies from the five classes.

In Table 4, for a fair comparison, the performances of our proposed method are also evaluated based on the dataset without the test set balancing as done for results reported in Figure 5. It can be seen that the proposed model is somehow comparable with the other

Components	Combination#1	Combination#2	Combination#3	Combination#4	Combination#5
Main-branch	x	x	x	x	x
Side-branch		x		x	
Side-branch with zoom			x		x
Attention				x	x
	65.61/67.50	73.98/71.08	78.73/79.83	78.81/76.75	79.64/83.08

Note(s): The performances of each combination are reported using sensitivity (%)/specification (%)

Table 3. The ablation study on four components of the proposed solution, using a scenario of two classes (normal and abnormal) based on Kaggle dataset

Methods	2 classes	5 classes
CNN with 12 layers [15]	sensitivity = 0.95, accuracy = 0.75	–
VGG-D [18]	accuracy = 0.82	accuracy = 0.51
ResNet-50 [16]	–	accuracy = 0.47
Bi-ResNet [16]	–	accuracy = 0.49
RA-Net [16]	–	accuracy = 0.47
BiRA-Net [16]	–	accuracy = 0.54
VGG16 [17]	–	accuracy = 0.51
Two-stages CNN [19]	–	accuracy = 0.54 (approx. from a result graph)
Multi-task CNN [33]	sensitivity = 0.64, accuracy = 0.82	–
Deep feature extraction [34]	–	accuracy = 0.51
Proposed method (DeepRoot)	specificity = 0.83, sensitivity = 0.80, accuracy = 0.80	accuracy = 0.55

Table 4. Comparisons under both scenarios of two classes and five classes with four stages

methods. The main objective of this paper is also to introduce the new CNN-based architecture, as explained above. Also, the trained model is validated to be sufficiently generalized on unseen samples of the self-collected dataset, with a sensitivity of 98.18% under the 2-classes scenario.

In addition, it is sometimes useful to classify the retinal image into three classes in a real-world scenario, where Stages 1 and 2 are grouped and Stages 3 and 4 are grouped. Thus, this experiment further validates the proposed method in this scenario of three classes. Figure 5e shows that the accuracy across the three classes becomes more stable at 81.86%, 72.71% and 84.29%. The sensitivity and specificity values of classifying normal from abnormal cases also balance, as 81.86% and 78.5%, respectively.

The advantage of the proposed solution is that it could distinguish DR from non-DR cases with high accuracy of over 80%. It could elaborate the sensitivity up to 98% with the lower specificity. The early stages (i.e. Stage 1 and Stage 2) could also be differentiated from the severe stages (i.e. Stage 3 and Stage 4) with high accuracy of 80% on average. However, the proposed solution's main limitation is grading individual stages. It still suffers from a low performance of separating Stage 1 from Stage 2 since the traces of the early stages of DR are very small (i.e. cover less than 2% of the whole image). In future work, the proposed solution can be used to classify Stage 1 and Stage 2 into the same category. Then, the two classes could be further split using microaneurysm, exudate and hemorrhage detections.

4. Conclusion

This paper proposed a new CNN architecture, DeepRoot, to detect and grade DR in retinal images. DeepRoot was designed to cope with fine-level features for detecting tiny traces of DR, such as microaneurysm and exudate. The staging outputs were determined at different locations in different layers of DeepRoot. DeepRoot was then trained and validated with the retinal images dataset from Kaggle. The trained model was also tested with unseen data samples, i.e. self-collected from a hospital. The model could achieve a very high sensitivity of 98.18% for the scenario of classifying into two classes of normal and DR. It could also be seen from the confusion matrix that the model could handle well with severe Stages 3 and 4 due to the advantage of the added attention layers. However, the performance significantly dropped in the early stages. In future work, the techniques of DR traces segmentation and CNN-based solution of DR staging should be combined. The segmentation-based solution should be employed for detecting Stage 1 and Stage 2, by segmentation microaneurysm and exudate. Then, Stage 3 and Stage 4 should be detected using the CNN-based classifier.

References

1. Salamat N, Missen MMS, Rashid A. Diabetic retinopathy techniques in retinal images: a review. Artif Intelligence Med. 2019; 97: 168-88.
2. Gangaputra S, Lovato JF, Hubbard L, Davis MD, Esser BA, Ambrosius WT, Chew EY, Greven C, Perdue LH, Wong WT., Condren A, Wilkinson CP, Agrón E, Adler S, Danis RP. Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity. Retina (Philadelphia, Pa). 2013; 33(7): 1393-1399.
3. Kasantikul R, Kusakunniran W. Improving supervised microaneurysm segmentation using autoencoder-regularized neural network. 2018 Digital Image Computing: Techniques and Applications (DICTA). IEEE; 2018: 1-7.
4. Zhang X, Wu J, Meng M, Sun Y, Sun W. Feature-transfer network and local background suppression for microaneurysm detection. Machine Vis Appl. 2021; 32(1): 1-13.
5. Kusakunniran W, Wu Q, Ritthipravat P, Zhang J. Hard exudates segmentation based on learned initial seeds and iterative graph cut. Comput Methods Programs Biomed. 2018; 158: 173-83.

6. Liu Q, Liu H, Zhao Y, Liang Y. Dual-branch network with dual-sampling modulated dice loss for hard exudate segmentation in colour fundus images. IEEE J Biomed Health Inform. 2021; 26(3): 1091-102.
7. Huang C, Zong Y, Ding Y, Luo X, Clawson K, Peng Y. A new deep learning approach for the retinal hard exudates detection based on superpixel multi-feature extraction and patch-based CNN. Neurocomputing. 2021; 452: 521-33.
8. Maqsood S, Damaševičius R, Maskeliūnas R. Hemorrhage detection based on 3d CNN deep learning framework and feature fusion for evaluating retinal abnormality in diabetic patients. Sensors. 2021; 21(11): 3865.
9. Ramasamy LK, Padinjappurathu SG, Kadry S, Damaševičius R. Detection of diabetic retinopathy using a fusion of textural and ridgelet features of retinal images and sequential minimal optimization classifier. PeerJ Comput Sci. 2021; 7: e456.
10. Lal S., Rehman SU, Shah JH, Meraj T, Rauf HT, Damaševičius R, Mohammed MA, Abdulkareem KH. Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. Sensors. 2021; 21(11): 3922.
11. Gayathri S, Gopi VP, Palanisamy P. A lightweight CNN for diabetic retinopathy classification from fundus images. Biomed Signal Process Control. 2020; 62: 102115.
12. Tymchenko B, Marchenko P, Spodarets D. Deep learning approach to diabetic retinopathy detection. arXiv Preprint. arXiv:2003.02261. 2020; 1-9.
13. Pao S-I, Lin H-Z, Chien K-H, Tai M-C, Chen J-T, Lin G-M. Detection of diabetic retinopathy using bichannel convolutional neural network. J Ophthalmol. 2020; 2020: 1-7.
14. Gangwar AK, Ravi V. Diabetic retinopathy detection using transfer learning and deep learning. In: Evolution in computational intelligence. Springer; 2021: 679-89.
15. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Proced Comput Sci. 2016; 90: 200-5.
16. Zhao Z, Zhang K, Hao X, Tian J, Chua MCH, Chen L, Xu X. Bira-net: bilinear attention net for diabetic retinopathy grading. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE; 2019: 1385-9.
17. Bravo MA, Arbeláez PA. Automatic diabetic retinopathy classification. In: 13th International Conference on Medical Information Processing and Analysis, 10572. International Society for Optics and Photonics; 2017. 105721E.
18. Kwasigroch A., Jarzembinski B., Grochowski M. Deep cnn based decision support system for detection and assessing the stage of diabetic retinopathy. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE; 2018: 111-16.
19. Yang Y, Li T, Li W, Wu H, Fan W, Zhang W. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2017: 533-40.
20. Qureshi I, Ma J, Abbas Q. Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning. Multimedia Tools Appl. 2021; 80(8): 11691-721.
21. Sridhar S, PradeepKandhasamy J, Sinthuja M, Minish TS. Diabetic retinopathy detection using convolutional neural networks algorithm. In: Materials Today: Proceedings; 2021.
22. Qummar S, Khan FG, Shah S, Khan A, Shamshirband S, Rehman ZU, Khan IA, Jadoon W. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access. 2019; 7: 150530-9.
23. Reguant R, Brunak S, Saha S. Understanding inherent image features in CNN-based assessment of diabetic retinopathy. Scientific Rep. 2021; 11(1): 1-12.
24. Naeem H, Bin-Salem AA. A CNN-LSTM network with multi-level feature extraction-based approach for automated detection of coronavirus from ct scan and x-ray images. Appl Soft Comput. 2021; 113: 107918.

25. Gao H, Yang Y, Lei S, Li C, Zhou H, Qu X. Multi-branch fusion network for hyperspectral image classification. Knowledge-Based Syst. 2019; 167: 11-25.
26. Zhuang J. Laddernet: multi-path networks based on u-net for medical image segmentation. arXiv Preprint. arXiv:1810.07810. 2018; 1-4.
27. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2015: 234-41.
28. Staal J, Abràmoff MD, Niemeijer M, Viergever MA, Van Ginneken B. Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging. 2004; 23(4): 501-9.
29. Aslani S, Dayan M, Storelli L, Filippi M, Murino V, Rocca MA, Sona D. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage. 2019; 196: 1-15.
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016: 770-8.
31. Tan M., Le Q, Efficientnet. Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR; 2019: 6105-14.
32. Kaggle. Diabetic retinopathy detection, identify signs of diabetic retinopathy in eye images; 2015. Available from: https://www.kaggle.com/c/diabetic-retinopathy-detection
33. Majumder S, Kehtarnavaz N. Multitasking deep learning model for detection of five stages of diabetic retinopathy. IEEE Access. 2021; 9: 123220-30.
34. Sungheetha A, Sharma R. Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. J Trends Comput Sci Smart Technol (TCSST). 2021; 3(02): 81-94.
35. Bai J, Ren J, Yang Y, Xiao Z, Yu W, Havyarimana V, Jiao L. Object detection in large-scale remote-sensing images based on time-frequency analysis and feature optimization. IEEE Trans Geosci Remote Sensing. 2021; 60: 1-6.
36. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016: 2818-26.
37. Wang Z, Yin Y, Shi J, Fang W, Li H, Wang X. Zoom-in-net: deep mining lesions for diabetic retinopathy detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2017: 267-75.
38. Ullah F, Moon J, Naeem H, Jabbar S. Explainable artificial intelligence approach in combating real-time surveillance of COVID-19 pandemic from CT scan and x-ray images using ensemble model. J Supercomputing. 2022; 78(17): 19246-71.

Corresponding author
Worapan Kusakunniran can be contacted at: worapan.kun@mahidol.edu

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**