Neural Computing and Applications (2023) 35:23959–23971
https://doi.org/10.1007/s00521-023-09001-1

ORIGINAL ARTICLE

Automated detecting and severity grading of diabetic retinopathy using transfer learning and attention mechanism

Maryam Dinpajhouh[1] · **Seyyed Ali Seyyedsalehi**[1]

Received: 14 June 2022 / Accepted: 22 August 2023 / Published online: 11 September 2023

Abstract

Diabetic retinopathy (DR) is a common retinal complication led by diabetes over the years, considered a cause of vision loss. Its timely identification is crucial to prevent blindness, requiring expert humans to analyze digital color fundus images. Hence, it is a time-consuming and expensive process. In this study, we propose a model named Attention-DenseNet for detecting and severity grading of DR. We apply a pre-trained convolutional neural network to extract features and get a hierarchical representation of color fundus images. What is essential for the correct diagnosis of DR is to recognize all the retinal lesions and discriminative regions. However, convolutional neural networks may overlook some tiny lesions of color fundus images. So, we use an attention model to solve this issue, which helps the model focus more on distinctive areas than others. We use APTOS 2019 dataset and fivefold cross-validation to assess the model's performance. The method achieves an overall accuracy of 98.44%, an area under receiver operating characteristic curve of 99.55%, and quadratic weighted kappa of 96.88% for the detection task, and an overall accuracy of 83.69%, an area under receiver operating characteristic curve of 97%, and quadratic weighted kappa of 89.26% for grading task. Our experimental results indicate that the model is superior to recent studies and can be suitable for DR classification in real life, especially for DR detection.

Keywords Diabetic retinopathy · Convolutional neural network · Attention · Transfer learning · APTOS 2019

1 Introduction

Diabetic retinopathy (DR) is among the most common retinal diseases, led by diabetes. It is one of the most significant causes of partial or complete blindness in 20–35 years. Approximately 33% of people with diabetes have DR symptoms, of which 10% will suffer from a vision-threatening stage of DR.

Increased blood glucose levels in people who have diabetes can cause serious damage to the retinal blood capillaries, causing them to leak fluid or blood. Such fluid leaks are likely to form various lesions in the retina, like microaneurysms, hard exudates, soft exudates, and hemorrhages. Specialists generally classify DR disease into the Non-Proliferative DR (NPDR) and Proliferative DR (PDR) stages. The NPDR stage can have various severities heavily dependent upon retinal lesions and damages, including mild, moderate, and severe. The mild severity is only characterized by the presence of microaneurysms [1]. In the moderate stage, extensive microaneurysms and hemorrhages are present. It may be seen in some cotton wool spots [2]. In the severe stage, venous beading and at least 20 hemorrhages could occur in at least two and every four quadrants. In this stage, intra-retinal microvascular abnormality (IRMA) is present in at least one quadrant [3]. Moreover, some new abnormal blood vessels grow around the optic disk and other regions in the PDR stage, which have thin and fragile walls that could lead them to leak blood. This leaked blood is dangerous and threatens to harm individuals' vision [1, 2].

By early diagnosis of DR, visual impairment or blindness can be delayed or prevented. Clinical and conventional DR diagnosing approaches are usually manually dependent on expert ophthalmologists who analyze digital color fundus images. However, these methods are time-

✉ Seyyed Ali Seyyedsalehi
ssalehi@aut.ac.ir

Maryam Dinpajhouh
Maryam.dinpajhouh@gmail.com

[1] Department of Biomedical Engineering, Amirkabir University of Technology, 424 Hafez Avenue, Tehran, Iran

consuming and most likely prone to error. Based on recent studies, automatic systems are claimed to solve such problems.

In recent few years, different models founded on conventional machine learning and deep learning have been proposed to detect and grade DR. Machine learning algorithms learn from their experiences and enhance their knowledge. Such algorithms make a machine capable of analyzing, understanding, recognizing, and classifying raw data. However, conventional machine learning approaches are not resistant to changes in data, so they have a weak generalization level. In other words, they cannot accurately analyze big data, especially images, due to their low computational complexity.

Convolutional neural networks (CNNs) have been among the most successful deep artificial neural networks in different fields, particularly computer vision tasks. This structure has a high-level computational complexity, requiring sufficient annotated data to avoid serious issues like overfitting, weak generalization power, and divergence. CNNs are considered one of the best proposed powerful tools in computer vision due to the accessibility of extensively annotated databases and graphics processing units (GPUs) [4].

The application of CNN in the medical field is a big challenge because the available annotated data in this field are usually inadequate. In addition, if a CNN is trained from scratch, it is time-consuming since its initial parameters are random. The transfer learning technique is a practical approach to such problems. In CNNs, transfer learning is a method for transferring knowledge from a CNN that has already been trained on a source dataset for a task to another CNN trained using the target dataset for the desired task. Since the knowledge of pre-trained CNN is used as the starting point in the new CNN, the initial parameters are not random. As a result, the generalization power increases, and the training process is not time-consuming. Moreover, it is not essential to access many annotated data to train the model.

In CNNs, transfer learning can be used by two methods. Firstly, transferred weights from the pre-trained CNN to the new one are frozen and used to extract features of the new dataset. Secondly, the new CNN is initialized using the weights of a specific pre-trained CNN. Based on the correlation between source and target datasets, several or all-new network layers' parameters are refined in a supervised manner. However, if the correlation between these two datasets is low, all layers' parameters should be fine-tuned, called full fine-tuning. But, if the correlation is high, only several last layers' parameters require fine-tuning, called layer-wise fine-tuning. In addition, the lower layers of a CNN extract low-level features related to many vision tasks, and the deeper layers extract high-level and more complex features related to specific vision tasks. So, full fine-tuning is not always necessary for many tasks. Besides, it is also time-consuming, which could be addressed by layer-wise fine-tuning [4].

All images have some key regions on which the classifier must focus more than others because they play a vitally important role in classifying different images and distinguishing them from each other. However, CNNs hierarchically extract features belonging to all spatial regions and aggregate them using pooling layers to represent an image accurately. So, CNN treats all features from different image areas the same and cannot discriminate between informative and non-informative regions in an image. As a result, the application of CNN for fine-grained tasks could be challenging. One of the most practical and best-proposed approaches to such challenges is the computational attention model inspired by a biological mechanism. It can be said that humans usually tend to pay more attention to more critical regions or content of a scene, such as familiar faces and different textures. In recent years, many researchers in the artificial intelligence field have proposed various computational attention models based on human brain mechanisms, helping deep learning networks, like CNNs, capture discriminative regions of an image or video.

This paper proposed a model named Attention-DenseNet for DR screening and severity grading inspired by the methods presented in [5, 6]. Diabetic retinopathy datasets are usually limited and have hundreds or thousands of digital color fundus retinal images, making deep learning models face overfitting and divergence. To tackle this problem and boost the generalization power, we prefer to use a pre-trained CNN dubbed DenseNet121 to extract features and get a hierarchical representation of color fundus images. In the next step, as it is necessary to recognize key regions and small lesions in the image to diagnose DR far more accurately, we applied the attention block proposed in [5], which is founded on a recurrent neural network (RNN) and a soft attention mechanism. We made some modifications to make it appropriate for the DR classification task. It is worth mentioning that, in this paper, the applied RNN does not generate any sequences or outputs, unlike the RNNs of the presented models in [5, 6]. In other words, in our model, the RNN aims to help the attention mechanism recompute the positive attention weights and refine the context vector expressing all important and unimportant features of the fundus image.

The attention block of the Attention-DenseNet model receives the extracted features generated by the pre-trained DenseNet121 and determines the importance of all features of an image for DR classification. In addition, the model considers tiny lesions for DR diagnosis, which traditional CNNs probably overlook.

This attention block can be trained end-to-end, so we train the Attention-DenseNet model end-to-end. We evaluate the proposed model's effectiveness for binary (No DR and DR) and multi-class (NoDR, mild NPDR, moderate NPDR, severe NPDR, and PDR) classifications to have a more reasonable and acceptable comparison with other models. Our investigation of the APTOS 2019 dataset proves the effectiveness of the applied method and its superiority compared with other DR screening and grading models.

The chief contributions of the present article are as follows:

- We present an attention model and apply it to a CNN. It is based on an RNN which is used to improve the positive attention weights and the context vector (the attention-wise weighted sum of features) for a precise DR diagnosis. To this purpose, we make some modifications to the models proposed in [5, 6] to have an appropriate decoder (RNN) for attention weights refinement and image classification. To the best of our knowledge, this is the first work in which the RNN's role is to recompute and refine the attention weights, not generate a sequence in the output.
- Experiments on the publicly available Kaggle APTOS 2019 blindness detection dataset [7] indicate that the performance of the Attention-DenseNet model is superior to that of the alternative models for DR grading and screening, particularly it obtains the best results on the APTOS 2019 dataset for DR detection with an area under receiver operating characteristic curve (AUC-ROC) score of 99.55%.

2 Related works

Machine learning approaches have developed considerably in the last decades and have been deployed in various areas, ranging from NLP to image processing. It has been widely applied in medical image processing and diagnoses, particularly automatic diagnosis of diabetic retinopathy. These approaches are much less time-consuming and more accurate than manual and clinical methods. This section reviews some studies in which researchers presented various models based on machine learning techniques, including conventional methods and deep learning architectures for DR diagnosis.

2.1 Studies based on conventional machine learning methods for DR diagnosis

A wide range of research has focused on designing models that rely on traditional machine learning approaches to grade DR severity or detect retinal lesions. Rayudu et al. [8] introduced models using different machine learning algorithms like SVM, KNN, and LDA to classify the non-proliferative diabetic retinopathy severity. Their investigation of the Drive database proved that the model, including the SVM classifier, obtained 88.8% accuracy. Satyananda et al. [9] applied SVM, PNN, Bayesian classification, and K-Means Clustering to classify diabetic retinopathy into non-proliferative (NPDR) and proliferative DR (PDR) stages. They used 300 color fundus images to train and evaluate the proposed models and deployed some image processing techniques to extract features of fundus images. The authors observed that the SVM model with 97% accuracy is the best among the designed models. Kanimozhi et al. [10] proposed a novel model to detect retinal lesions. The model consists of four essential steps: enhancing contrast and luminosity, removing extracted optic disk and blood vessels, detecting lesions, and classifying dark and bright lesions. Their investigation showed that the model attained overall accuracies of 97.43%, 98.06%, and 96.98% for microaneurysms, hemorrhages, and exudates detection, respectively. Huda et al. [11] designed a model based on machine learning approaches to detect different retinal lesions of diabetic retinopathy. They applied a tree-based classifier to select the 30 critical features to decrease training time and avoid overfitting. After extracting the 30 most important features, they built classifiers based on SVM, KNN, Logistic Regression, and decision tree methods. They trained and evaluated models on the DIARETDB1 dataset and concluded that the performances of SVM-based and Logistic Regression-based models are better than other approaches. Chetoui et al. [12] designed a new model by extracting different texture features and using SVM with Radial Basis Function Kernel classifier to divide color fundus images into two categories (No DR and DR). The proposed model was trained on the MESSIDOR dataset and obtained an AUC-ROC score of 93.1%.

As conventional machine learning approaches have a low level of complexity, they cannot analyze and understand color retinal fundus images accurately. Moreover, these methods require accurate engineering to extract and select features. So, the models discussed in this section may have an inappropriate performance in DR detection and classification.

2.2 Studies based on deep learning and transfer learning methods for DR diagnosis

Nowadays, many researchers have designed different models using various deep learning algorithms to extract features automatically. These models are appropriate for interpreting and analyzing big data due to their computational complexity, especially for medical image processing and various diagnostic tasks, including detecting lesions and abnormalities [13], classification of lesions and abnormalities, grading the severity of a disease, and segmentation of images of humans' organs [13]. So, many deep learning-based models have been designed for these medical tasks, particularly for detecting and grading diabetic retinopathy, for which its timely identification is a big challenge. Doshi et al. [14] developed three different deep CNNs and an ensembled model made of them to diagnose DR severity automatically. Examining them on a dataset containing 35,126 color fundus images proved that the ensembled network achieved a 39.96% quadratic weighted kappa (QWK) metric and performed better than single proposed models. Ghosh et al. [15] proposed a model based on a six-layer CNN to screen DR and classify its severity levels. They trained and evaluated the presented model on over 30,000 color fundus images and observed that it attained 95% accuracy for the DR detection task and 85% for the DR severity grading task. Saranya et al. [16] proposed a CNN-based classifier for DR detection. They removed the optic disk and pre-processed color fundus images which are fed into the designed convolutional neural network. This CNN is made of three convolutional layers, three max-pooling layers, two dropout layers, fully connected, and classification layers. Finally, their analysis of the Messidor and IDRiD datasets revealed that the model achieved accuracies of 90.89% and 90.29%, respectively.

The unavailability of sufficient annotated color fundus images decreases the generalization power of deep learning-based models. However, applying transfer learning can solve this issue and improve the generalization power of such models. Hagos et al. [17] implemented a CNN model based on the pre-trained Inception-V3 structure for DR screening. The authors used the convolutional part of the pre-trained Inception-V3 as a feature extractor. They added a fully connected layer with the Relu activation function, followed by a softmax classifier to classify DR. The model was then trained and evaluated on the randomly selected color fundus images from EyePACS, obtaining a 90.9% accuracy. They concluded that the transfer learning approach is appropriate and helpful for detecting diseases with insufficient available annotated data, especially diabetic retinopathy. Gangwar et al. [18] deployed the pre-trained Inception-Resnet-V2. They removed the last layers of the pre-trained network- fully connected layers- to use it as a feature extractor and added a custom CNN block to hybridize it on top of the convolutional part of the pre-trained network. They also added a fully connected layer and a softmax classifier on top of the custom CNN to grade the DR severity. Finally, the proposed model obtained 72.33% and 82.18% test accuracy on the Messidor-1 and APTOS 2019 dataset, respectively.

It is important to note that although these studies show deep learning networks' great ability to grade and screen DR, they cannot emphasize discriminative regions and may overlook some small lesions playing important roles in DR diagnosis.

2.3 Attention mechanism

The attention mechanism is a highly effective method to capture fine-grained features. It has an extensive application in different tasks of computer vision, like image classification [19], semantic segmentation [20, 21], and object localization [21]. Regarding the attention mechanism transformers and the recurrent attention model are considered promising and successful models. Vaswani A [22] proposed an architecture founded on an attention (self-attention) mechanism, dubbed transformer, which nowadays has become one of the most standard and promising attention models. The transformer computes the representation of the input data without applying convolutions or RNNs. In fact, in this model, the multi-headed self-attention mechanism is used instead of the recurrent layers in encoder–decoder architecture and maps a sequence $(x_1, x_2, x_3, \ldots, x_n)$ to another sequence $(z_1, z_2, z_3, \ldots, z_n)$ of the same length with $x_i, z_i \in R^d$, like a hidden layer in an encoder or decoder. The experimental results on two machine translation tasks indicated that this model is more parallelizable and less time-consuming compared with recurrent or convolutional layers. Mnih et al. [23] presented a novel attention architecture based on recurrent neural network, called recurrent attention model (RAM). The model pays attention to different parts of the image selectively at every step and finally combines these extracted features to calculate a dynamic representation of the input (image or video). This model decides to attend to which part of the input image at the next step based on the previous extracted information. By virtue of RAM the entire image is not required to be processed, making the amount of computation independent to the size of input image. So, RAM is less computationally expensive compared to CNNs. Besides, they evaluated the model on several image classification tasks and concluded that this model is superior to the CNN.

2.3.1 Studies based on attention mechanism for DR classification

Generally, the attention mechanism is widely deployed in the medical image-based diagnosis area because all medical images have some key and informative regions necessary to be recognized for early and accurate identification, especially DR diagnosis. Zhao et al. [24] designed an architecture based on a CNN to simultaneously diagnose DR and localize suspicious areas of color fundus images using several high-resolution patches. The model is weakly supervised with the image-level label. It consists of three main networks: The main network, the Attention network, and the Crop network. Their investigation of EyePACS and Messidor showed their model superiority to other proposed models. Lin et al. [25] presented a novel framework based on the center-sample detector and Attention Fusion Network (AFN) to predict the probabilities of the lesions on the retina using bounding boxes of lesions and grade DR, respectively. The AFN consists of two CNNs extracting the feature maps of the original images and the lesion maps. In the next step, the weights between these feature maps are calculated, helping reduce the effect of unnecessary lesions for DR grading. The experimental results on the private dataset, Messidor, and EyePACS indicate that this model outperforms many other state-of-the-art models. Li et al. [26] developed an attention model called a cross-disease attention network (CANet) to grade DR and DME simultaneously. It consists of two modules: a disease-specific attention module, which learns features relevant to each disease, and a disease-dependent attention block learning the internal relationship between the diseases. They applied the pre-trained ResNet50 to extract features of both diseases simultaneously and then passed the encoded vectors through the attention block. Their experimental results on Messidor and IDRiD proved that the presented model outperforms other related models. He et al. [27] proposed a novel attention model dubbed Category Attention Block Network (CABNet), which helps CNN to learn discriminative features of digital fundus images for DR detection and grading tasks. They passed the encoded vector of pre-trained DenseNet121 through this block and evaluated the model on three different datasets, including DDR, Messidor, and EyePACS. This model obtained promising results.

These studies demonstrate that the attention mechanism improves models' performance in DR classification tasks. It allows conventional deep learning networks to concentrate more on informative regions and capture tiny lesions of color fundus images, like microaneurysms. So, in this paper, to detect these small lesions in the digital fundus images we propose an attention mechanism, inspired by [5, 6], and integrate it into a pre-trained CNN, which will be discussed in the following section.

3 Method and materials

This section highlights and explains all methods and materials we applied for the study, including the dataset, data pre-processing, proposed model, training process, and evaluation metrics.

4 Dataset

We used the Kaggle APTOS 2019 blindness detection dataset containing 3662 color retinal fundus images taken by fundus cameras in various imaging conditions [7].

An expert clinician categorized the images of the dataset into five following stages: No DR (stage 0), Mild DR (stage 1), Moderate DR (stage 2), Severe DR (stage 3), and Proliferative DR (stage 4). Figure 1 shows different categories of fundus images from the APTOS 2019 dataset.

From Fig. 2, we notice that the class distribution is highly imbalanced, and the most and least fundus images belong to the No DR and severe DR classes, respectively. However, we do not balance the class distribution by undersampling or oversampling. Instead, we use more reliable evaluation metrics besides accuracy, like QWK, to evaluate the model, which will be discussed in the next sections.

The dataset is classified into five levels of DR severity, which is appropriate for the DR severity grading task in the current paper. Moreover, we categorized the dataset into two NoDR and DR categories by unifying the mild, moderate, severe, and proliferative stages to generate the DR class and relabeling them with the same label.

4.1 Data pre-processing

We performed simple pre-processing techniques. The fundus images of the ATOS2019 dataset have different heights and widths in the range of [358,2848] and [474,4288], respectively. Since we used the pre-trained DenseNet121 as a backbone network to extract features, we resized all dataset images to 224*224 pixels to make their dimensions suitable for the input of this pre-trained

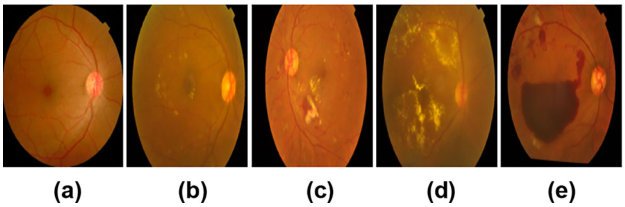

Fig. 1 Digital color fundus images belonging to APTOS 2019 dataset, showing various severities of diabetic retinopathy disease. **a** No DR **b**Mild DR **c** Moderate DR **d** Severe DR **e** Proliferative DR

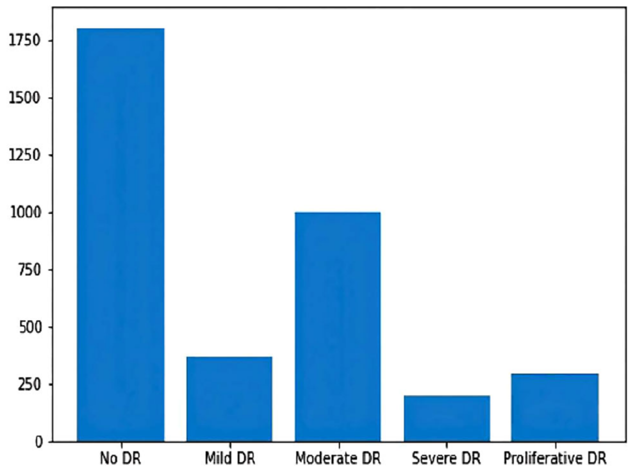

Fig. 2 Classes distribution of APTOS2019 dataset

network. Moreover, we standardized APTOS 2019 dataset by subtracting the pre-computed mean of ImageNet from the channels of all images (centering) and then dividing it by the standard deviation (scaling). We do these operations using the pre-process_input() function related to the DenseNet121 network in the Keras library.

The APTOS 2019 has insufficient color fundus images, leading to overfitting and divergence problems in deep learning. To address the issue and make the model robust, we augmented data by randomly rotating, zooming, and shearing images in the range of 0–360, 0–0.2, and 0–0.2, respectively, and applying the horizontal and vertical flipping augmentation techniques.

4.2 Proposed model

This section explains the Attention-DenseNet framework, inspired by the proposed models in Bahdanau et al. [5] and Xu et al. [6], which aim to improve the performance of the traditional encoder–decoder model for translation and caption generation for the input image, respectively, using an attention mechanism. Moreover, the modifications we have made to these models for classification tasks will be discussed.

We deployed a pre-trained CNN named DenseNet121 to extract the hierarchical features of the color fundus images. We extract the feature maps of the CNN's middle layers since they could correspond with different parts of the two-dimensional images [6]. It permits the attention mechanism to concentrate on an image's particular features or regions [6].

Generally, a CNN produces the output of size W*H*L in which the L is the number of feature maps, and W and H are the width and height, respectively. Our experiment considers the 310th middle layer of the pre-trained DenseNet121, generating 1024 feature maps of size 14*14, which has been selected after evaluating several low, middle, and deep layers of the network. We converted the extracted feature maps into 196 vectors expressed by $F = \{F_1, F_2, F_3, \ldots, F_{196}\}$, in which $F_j \in R^{1024}$, $j \in \{1, 2, 3, \ldots, 196\}$ consists of all jth features throughout all feature maps to prepare them as an appropriate input for the attention blocks- which calculate positive attention weights for each feature vector.

We also apply a gated recurrent unit (GRU) that contributes to refining calculated attention weights and the generated context vector in each attention block, which we will explain later. The equations of GRU are as follows:

$$r_i = \sigma\left(\boldsymbol{U_r}\mathrm{h}_{i-1} + \mathbf{C}_r\hat{X}_i + \boldsymbol{W_r}\hat{X}_{i-1}\right)$$

$$z_i = \sigma\left(\boldsymbol{U_z}\mathrm{h}_{i-1} + \boldsymbol{C_z}\hat{X}_i + \boldsymbol{W_z}\hat{X}_{i-1}\right)$$

$$\tilde{s}_i = \tanh\left(\boldsymbol{U_p}[r_i \cdot h_{i-1}] + \boldsymbol{C_p}\hat{X}_i + \boldsymbol{W_p}\hat{X}_{i-1}\right)$$

$$h_i = (1 - z_i) \cdot \tilde{s}_i + z_i \cdot h_{i-1}$$

In these equations z_i, r_i, $\tilde{s}_i$ and h_i are the update gate, reset gate, proposal hidden state, and the hidden state of the GRU, respectively. They are all computed based on the current context vector, previously calculated context vectors, and the previous hidden state. In addition, $\boldsymbol{U_r}$, $\boldsymbol{U_z}$, $\boldsymbol{U_p}$, $\boldsymbol{C_r}$, $\boldsymbol{C_z}$, $\boldsymbol{C_p}$, $\boldsymbol{W_r}$, $\boldsymbol{W_z}$ and $\boldsymbol{W_p}$ matrices include trainable weights learned during the training process. The context vector $\hat{X}_i$ represents the relevant and irrelevant regions or features of a retinal image, which will be discussed later.

The attention block helps us recognize more critical and relevant features or regions of the retinal images to grade and detect DR better and more accurately. Based on Fig. 3, the first step in the applied attention mechanism is to calculate the similarity between the hidden state of the previous step of the GRU (h_{i-1}) and each of the feature vectors F_j. Therefore, other inputs, like lesion maps, are not considered for calculating the similarity and the attention weights. It is unlike the presented model in [25], in which the attention weights are computed using the original image and lesion maps of DR.

$$e_{i,j} = Alignment\left(F_j, h_{i-1}\right)$$

We applied the alignment model e using a single-layer feedforward neural network presented in [5] to calculate this similarity, which is defined by Eq. (1).

$$Alignment = \boldsymbol{V_a} * \tanh\left(\boldsymbol{W_a}h_{i-1} + \boldsymbol{U_a}F_j\right) \tag{1}$$

Here $\boldsymbol{V_a}$, $\boldsymbol{W_a}$ and $\boldsymbol{U_a}$ are the matrices, including trainable weights.

Figure 3 shows that in the next step, the alignment scores $e = \{e_1, e_2, e_3, \ldots, e_{196}\}$ are normalized using a softmax function, and then the positive attention weights $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{196}\}$ are generated, the sum of which

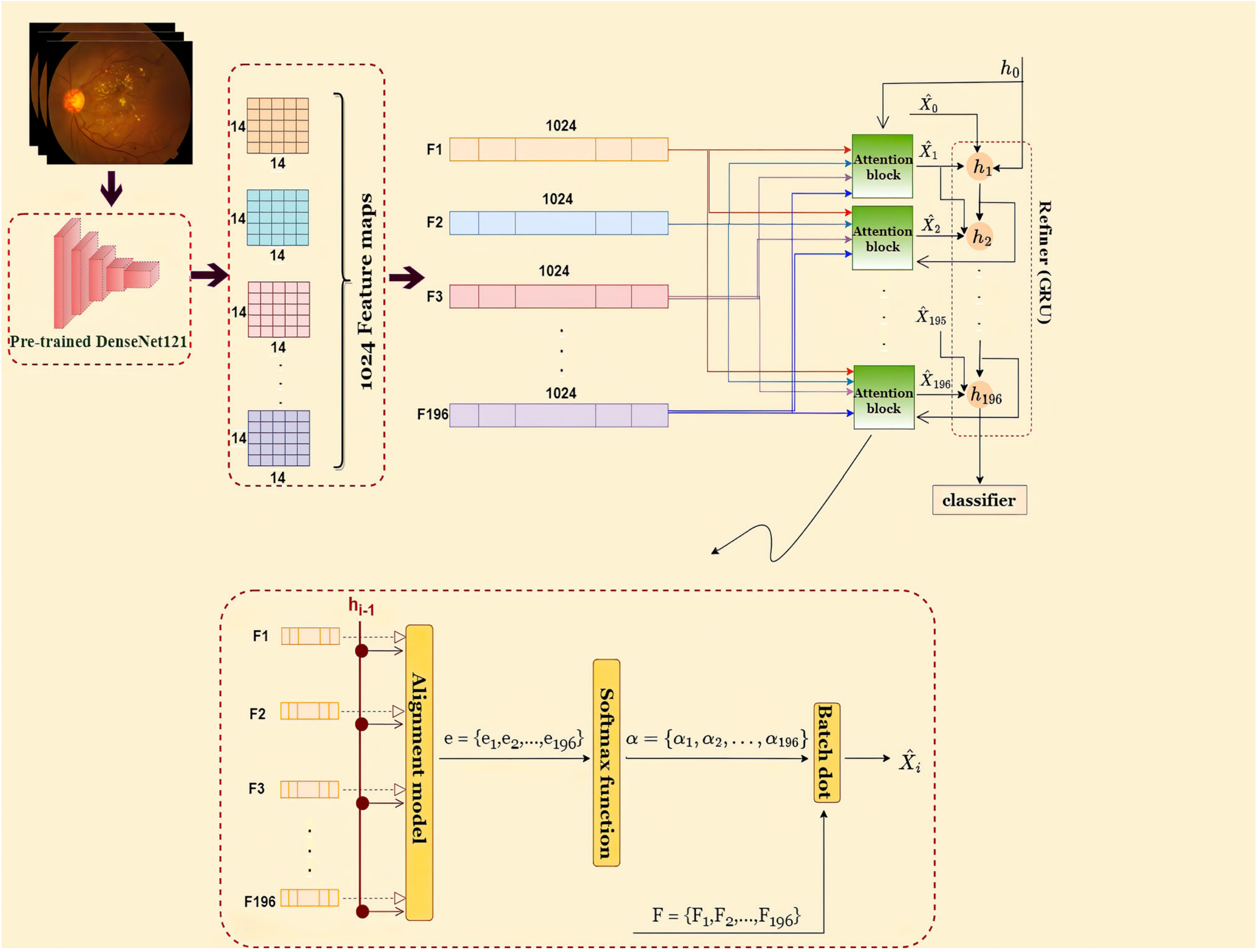

Fig. 3 The overall structure of the Attention-DenseNet network. It consists of three major parts, including the backbone, attention, and classifier. Notice that here GRU is used as a refiner and does not produce any sequence. The final hidden state is used for grading DR and sent to a classifier. It is important to note that we have two independent and distinct classifiers for grading and detecting DR and train them separately

is equal to 1. This softmax function is formulated as follows:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{196} \exp(e_{ik})}$$

The positive attention weight α_{ij} is between 0 and 1, calculated for each feature vector j in step i. In fact, every attention weight $\alpha_j, \quad j \in \{1, 2, 3, \ldots, 196\}$ expresses the relative importance of the jth feature vector by considering all 196 feature vectors. Then, the context vector is achieved by taking a weighted sum of all feature vectors as proposed in [5].

$$\hat{X}_i = \sum_{j=1}^{196} \alpha_{ij} F_j$$

As mentioned above, the context vector represents the weighted regions. In other words, it expresses the important and unimportant features and shows which parts of retinal images should be paid either more or less attention to for grading and detecting DR in a promising way.

What is important to note is that there are two main differences between our model (Attention-DenseNet) and those presented in [5, 6], which are as follows:

- In the Attention-DenseNet model, GRU does not produce any output or sequence. Indeed, it is used for attention weights refinement. However, in [5, 6], the RNN is applied to produce an output at each step for translation and caption generation tasks, respectively.
- Furthermore, from Eq. (2), it is clear that in the models proposed in [5, 6] the hidden states of the RNNs are not based on the previously computed context vector but on the output of the previous step (y_{i-1}). This is in contrast to the Attention-DenseNet model in which the hidden

state is dependent upon the previously computed context vector.

$$h_i = f\left(h_{i-1}, y_{i-1}, \hat{X}_i\right) \tag{2}$$

Thus, in our paper, the GRU's role is to refine the level of attention paid to each feature vector. In fact, in each step i, the importance of all regions is recomputed and improved. As a result, the context vector is dynamic and changes in each step i. The improved context vector is fed into GRU as an input, and the hidden state is updated, helping to refine the context vector in the next step. This mechanism continues to the last step and causes the model to be more able to concentrate on the key regions of retinal images compared with a traditional CNN.

This study aims to classify retinal images into five classes (multi-class classification) and two general categories (binary classification). To do this, we apply the attention mechanism separately for each task.

- We feed the last hidden state to a single-layer feedforward neural network consisting of five neurons that have a softmax activation function to grade DR to (1) No DR, (2) Mild DR, (3) Moderate DR, (4) Severe DR, and (5) Proliferative DR.
- We feed the last hidden state to a single-layer feedforward neural network consisting of two neurons, which have a sigmoid function to detect DR. (1) No DR, (2) DR

It should be noted that our preliminary results show that the model performs the same by using either the context vector or the last hidden state as the input to the classifier. So, the last hidden state of the GRU in the Attention-DenseNet framework can be as appropriate as the context vector for the input to the classifier. That is why we consider the last hidden state for the classification task.

The initial hidden state of GRU is calculated as that in [5]:

$$h_0 = \tanh(F_1.\boldsymbol{W}_i)$$

which $\boldsymbol{W}_i$ is a matrix including trainable weights learned in the training process.

4.3 Training process

We divided the APTOS 2019 dataset into five groups utilizing the fivefold cross-validation method to use all dataset images in the training and testing processes and have less biased results. In the training process, we fine-tuned the pre-trained DenseNet121. In other words, we retrained all its pre-trained parameters of all layers (from the 1st to the 310th layers) to adapt them to our data. We used the Adam algorithm [28] for optimization, Categorical_crossentropy (for severity grading task), and Binary_crossentropy (for screening task) loss functions for training the model. For the training phase of both binary and multi-class tasks, the batch size of 32, the learning rate of $6 * 10^{-5}$, and the epoch number of 40 were considered in all folds.

The model was trained on a Tesla P100-PCIE GPU with 16 GB memory, and the applied programming language is Python 3.7.12. Moreover, we used some modules of Python for implementing the model, including Keras 2.3.1, Tensorflow 2.2.0, Sklearn, SciPy, Imblearn, and Seaborn. For implementing the attention block, we were inspired by [29].

4.4 Evaluation metrics

This section describes the evaluation metrics by which we evaluate the model's performance in DR classification. As the APTOS 2019 dataset is highly imbalanced, accuracy is unreliable in assessing the model's performance, so we consider other evaluation metrics like sensitivity, specificity, precision, recall, and $F1$-score. The sensitivity and specificity metrics assess how well the classifier can estimate the positive and negative classes. The definition of recall is the same as sensitivity. Still, precision is the proportion of correctly predicted positives and total predicted positives, and the $F1$-score is the harmonic mean of precision and recall. These metrics are mathematically defined as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

$$\text{recall} = \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F1_{\text{Score}} = \frac{(2 * \text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

Here, TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Furthermore, we consider QWK, receiver operating characteristic (ROC), and precision–recall curve (PRC) to evaluate the model's performance. QWK measures the agreement or disagreement between predicted and actual labels [30]. The score can be in the range of [− 1, 1], in which the values of − 1, 1, and 0 mean the total disagreement, total agreement, and chance-based agreement, respectively [30]. QWK can be computed as Eq. (3):

$$K = 1 - \frac{\sum_{i=0}^{N}\sum_{j=0}^{N} w(i,j)c(i,j)}{\sum_{i=0}^{N}\sum_{j=0}^{N} w(i,j)e(i,j)} \quad (3)$$

In this formula, N is the number of classes and $w(i,j)$ is the element of the weighted matrix, which can be calculated as the following expression:

$$w(i,j) = \frac{(i-j)^2}{(N-1)^2}$$

$c(i,j)$ is also the element of the normalized confusion matrix, $e(i,j)$ is the element of the normalized expected matrix, which is the outer product of actual and predicted labels vectors.

Moreover, we plot ROC to evaluate the model's performance and determine its ability to differentiate between all stages of the disease. It is generated based on calculating the true positive rate versus the false positive rate for the proposed classifier at diverse thresholds. ROC summarizes the classifier's performance and provides a score assessing the model's power of differentiation. The AUC-ROC score can be in the range of 0 to 1, which a score of 1 shows that the model can distinguish diabetic retinopathy grades perfectly and possesses outstanding performance.

As APTOS 2019 dataset is highly imbalanced, we also plot PRC for a better evaluation of the predictive model. This curve shows the trade-off between precision and recall at various thresholds. Like ROC, we can summarize the information of PRC by calculating the area under precision–recall curve, (AUPRC), which is sensitive to the minority class. In this paper, all curves are plotted by averaging over fivefold with a total of hundred epochs.

Besides, we apply a confusion matrix that gives precise information about the type of errors in predicting sample categories. The numbers of correct and incorrect predicted samples in each category help us comprehend and analyze metrics more easily.

5 Experimental results

In this study, we implemented a fivefold cross-validation with stratified sampling, permitting us to average out metrics over fivefold.

5.1 Performance of the proposed attention model for DR severity grading

Table 1 compares the overall performance of the Attention-DenseNet model for the DR severity grading task with some other studies carried out for this task. From the table, our model obtained a sensitivity of 83.69%, a precision of 83.32%, and an F1-score of 83.04%, which are significantly greater than the same calculated metrics in other models, in particular the proposed model in [31]. In this system, t-SNE is used to reduce the dimension of the feature maps to improve the DR prediction accuracy. Besides, in terms of sensitivity and specificity, the Attention-DenseNet model outperforms the modified Xception architecture [32], founded on a combination of multilevel features in different layers of Xception architecture. Moreover, in terms of QWK, our model is superior to MobileNetV2 [33] which is a pre-trained network. It is important to notice that there is no attention mechanism in these mentioned architectures [31–33]. Therefore, we can conclude that an attention mechanism could be appropriate for an accurate DR severity recognition task.

The calculated AUC-ROC score of 97% in our model also indicates its more powerful differentiation ability than [34], which was designed using a gated attention mechanism to concentrate more on lesions in fundus images and less on the rest of the images.

Figure 4 presents the confusion matrix of the proposed classifier for DR grading, which is the average of matrices calculated in all fivefold. The confusion matrix shows that the misclassified retinal images of classes 0 and 2 are negligible. In addition, the misclassification of class 4 is much better than anticipated, although its available images are limited. We can also notice that almost 30% of retinal images of class 1 are predicted as class 2, and 24% of images from class 3 are predicted as class 2, owing to insufficient retinal images in these classes. Consequently, the attention model performs the best in classes 0, 2, and 4 and the worst in classes 1 and 3.

Figure 5 shows the averaged ROC of the attention model for DR severity grading over fivefold. From the figure, we notice that at the false positive rate of 0.1, the true positive rate is approximately 91%. In addition, we achieved an overall AUC-ROC of 97%, which indicates the outstanding model's power to distinguish various grades of the disease.

Figure 6 indicates the overall PRC of the Attention-DenseNet model for the DR severity grading task over fivefold. This figure highlights that the classifier has both high precision and recall, leading to a good AUPRC of 91.29%. This score means that the model is skillful and has a good performance for the DR grading task.

5.2 Performance of the proposed attention model for DR detection

Table 2 shows the overall performance of the Attention-DenseNet model for DR identification and some other research carried out for this task. Based on the table, the proposed model achieves an overall validation accuracy of 98.44%, which is significantly greater than other

Table 1 performance comparison of the Attention-DenseNet structure with other existing models designed for DR severity grading and trained on different DR datasets

Model	Dataset	Accuracy (%)	QWK	Sensitivity (recall)	Specificity	Precision	F1-score	AUC-ROC
Gated attention DNN [34]	APTOS 2019	82.54	–	83%	-	82%	82%	79%
Xception [32]	APTOS 2019	79.59	–	82.35%	86.32%	–	–	–
InceptionV3 [32]	APTOS 2019	78.72	–	63.64%	85.37%	–	–	–
MobileNet [32]	APTOS 2019	79.01	–	76.47%	84.62%	–	–	–
NASNet + t-SNE + SVM [31]	APTOS 2019	77.90		77%	–	76%	75%	–
MobileNetV2 [33]	APTOS 2019	78.47	81.23%	–	–	–	–	–
Hybrid Inception-ResNet-V2 [18]	APTOS 2019	82.18	–	–	–	–	–	–
Attention-DenseNet	APTOS 2019	83.69	89.26%	83.69%	94.72%	83.32%	83.04%	97%

"–" indicates no reported result

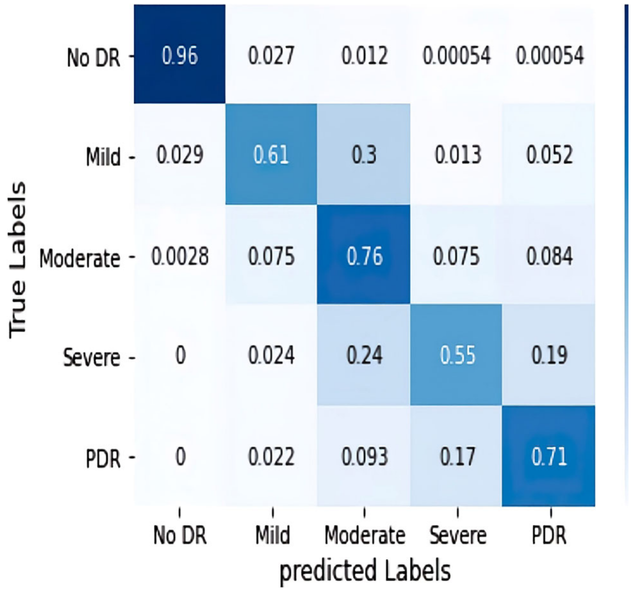

Fig. 4 Averaged confusion matrix over fivefold for DR severity grading using Attention-DenseNet structure

architectures in the table, whether they have been designed using an attention mechanism or not.

Figure 7 shows the averaged ROC curve of the proposed attention model for DR detection over fivefold. Based on the figure, we notice that at the false positive rate of 0.05, the true positive rate is nearly 99%. Furthermore, we achieved an overall AUC-ROC of 99.55%, which shows the model's promising power of differentiating data belonging to healthy (No DR) and unhealthy (DR) classes.

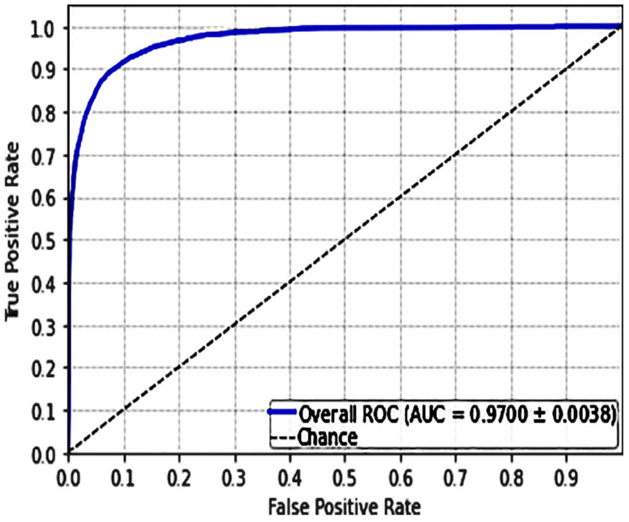

Fig. 5 Overall ROC curve over fivefold for DR severity grading using Attention-DenseNet structure

Figure 8 presents the confusion matrix of the proposed model for DR detection, which is the average of calculated matrices in all fivefold. It clearly shows that the model can perfectly identify DR disease with insignificant misclassification.

Figure 9 shows the overall PRC of the proposed model for the DR screening task over fivefold. Based on this curve, it is clear that the classifier has high precision and recall scores across the graph. In fact, these high scores of precision and recall express that the model has a low false positive rate and a low false negative rate, respectively.

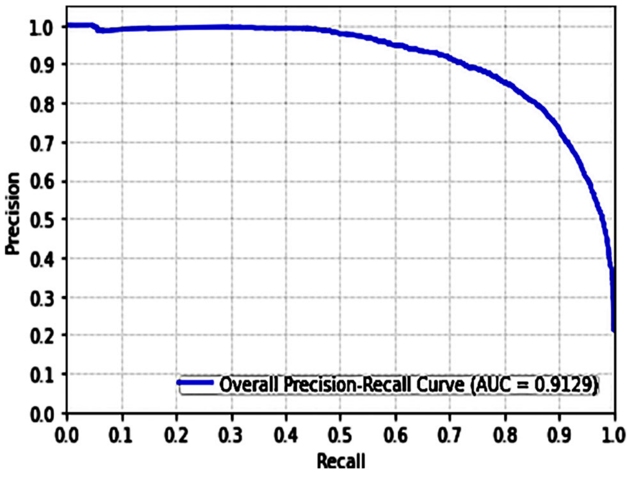

Fig. 6 Overall PR curve over fivefold for DR severity grading using Attention-DenseNet structure

Moreover, an AUPRC score of 99.49% is obtained, showing the model is perfect for the DR detection task.

6 Discussion

This paper demonstrated the presented Attention-DenseNet architecture's ability for DR grading and detecting tasks.

The experimental results prove that the presented model for detecting and grading tasks outperforms other recent models mentioned in Tables 1 and 2 regarding the accuracy, sensitivity, specificity, QWK, precision, AUC, and $F1$-score. Figures 4 and 8 show that the misclassification of the model is negligible, especially for the screening task. It is important to note that the usage of the attention mechanism in the proposed model plays a vitally important role in such enhancement. It contributes to the model distinguishing between key and unimportant regions of color fundus images, which are necessary to diagnose diabetic retinopathy accurately. Regarding the applied attention mechanism in our model (Attention-DenseNet), what should be highlighted is that the attention weights are recomputed in every step of the RNN to refine the context vector and stress the relevant features of the digital fundus images more than in the previous steps. This is unlike the models proposed in [5, 6], in which the calculated attention weights are not refined.

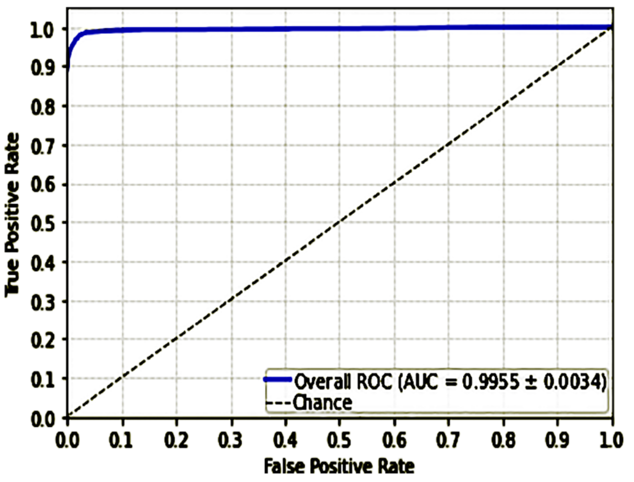

Fig. 7 Overall ROC curve over fivefold for DR detection using Attention-DenseNet structure

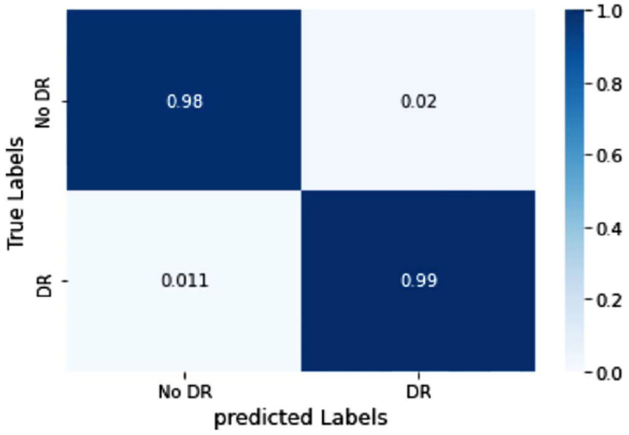

Fig. 8 Averaged confusion matrix over fivefold for DR detection using Attention-DenseNet structure

Table 2 Performance comparison of the Attention-DenseNet structure with other existing models designed for DR detection and trained on different datasets

Model	Dataset	Accuracy (%)	QWK	Sensitivity (Recall)	Specificity	Precision	F1-score	AUC-ROC
ResNEt34 [35]	APTOS 2019	96.35	–	–	–	–	–	–
DenseNet121 [35]	APTOS 2019	84.05	–	–	–	–	–	–
DetNet [35]	APTOS 2019	93.99	–	–	–	–	–	–
Gated attention DNN [34]	APTOS 2019	97.82	–	–	–	98%	98%	98%
Attention-DenseNet	APTOS 2019	98.44	96.88%	98.44%	98.45%	98.45%	98.44%	99.55%

"–" indicates no reported result

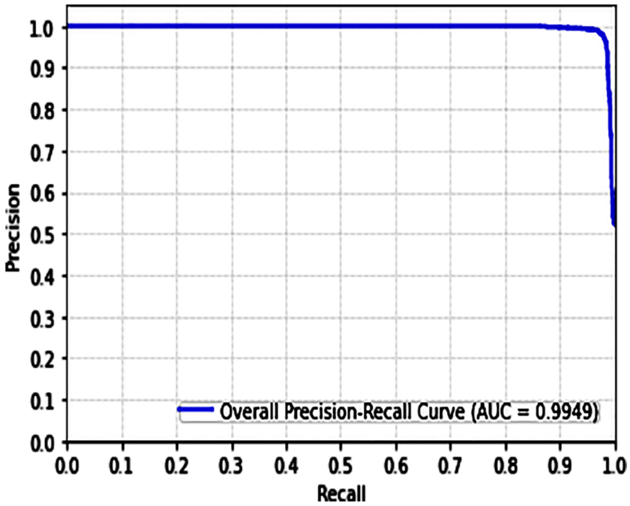

Fig. 9 Overall PR curve over fivefold for DR detection using Attention-DenseNet structure

In addition, the proposed model's applied pre-processing techniques are simple, making the model less time-consuming compared with various clinical ophthalmology methods.

Consequently, the Attention-DenseNet model can be considered a robust, accurate, and time-saving method to diagnose diabetic retinopathy disease in real life, particularly for its identification.

7 Conclusions and future direction

The main aim of this study was to develop a robust deep learning model to detect and grade diabetic retinopathy disease. Since almost all available DR datasets do not have enough color fundus images, we utilized a pre-trained CNN dubbed DenseNet121 to get a hierarchical representation of the images to avoid overfitting and weak generalization power. However, traditional CNNs may overlook some key and small lesions, like microaneurysms of a color fundus image, and cannot focus more on informative regions than others. To solve such issues, we applied an attention block founded on GRU and soft attention mechanism, through which we passed the encoded vector (output of DenseNet121). The attention block allows the model to learn distinctive features. The Attention-DenseNet architecture can be trained in an end-to-end manner. The experimental results on the APTOS 2019 dataset reveal the effectiveness of the applied attention mechanism in DR detecting and grading tasks and the model's superiority compared with the other existing studies.

The limitation of the model is its poor performance in diagnosing mild and severe grades, which may be due to insufficient available data in these classes. In fact, a limited dataset is likely to make the model overfitted, so the model's weak generalization power in these two categories could be hard to avoid. For the future direction, we would like to utilize some highly effective augmentation methods or additional data courses to expand the dataset to boost the model's performance for grading various levels of DR disease more accurately. In addition, we will apply other pre-trained CNN as the backbone network of the proposed architecture, including ResNet, MobileNet, GoogleNet, and VGG16, instead of DenseNet121, for the desired tasks.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-023-09001-1.

Funding Partial financial support was received from Amirkabir University of Technology (Tehran Polytechnic).

Data availability The APTOS 2019 dataset analyzed during the current study is publicly available at https://www.kaggle.com/c/aptos2019-blindness-detection.

Code availability The code files supporting our published claims are in supplementary files.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Alghadyan AA (2011) Diabetic retinopathy: an update. Saudi J Ophthalmol 25:99–111
2. Faust O, Acharya UR, Ng EYK et al (2012) Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. J Med Syst 36:145–157. https://doi.org/10.1007/s10916-010-9454-7
3. Wang X, Lu Y, Wang Y, Chen WB (2018) Diabetic retinopathy stage classification using convolutional neural networks. In: Proceedings: 2018 IEEE 19th international conference on information reuse and integration for data science, IRI 2018. Institute of Electrical and Electronics Engineers Inc., pp 465–471
4. Li X, Pang T, Xiong B et al (2018) Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In: Proceedings: 2017 10th international congress on image and signal processing, biomedical engineering and informatics, CISP-BMEI 2017. Institute of Electrical and Electronics Engineers Inc., pp 1–11
5. Bahdanau D, Cho KH, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. In: 3rd international conference on learning representations, ICLR 2015–conference track proceedings
6. Xu K, Ba J, Kiros R et al (2015) Show, attend and tell: neural image caption generation with visual attention, pp 2048–2057
7. APTOS 2019 Blindness Detection | Kaggle. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed 1 Jun 2021
8. Rayudu M, Pendam S, Dasari S (2020) Prediction of severity of non proliferated diabetic retinopathy using machine learning

techniques. J Comput Theor Nanosci 17:4219–4222. https://doi.org/10.1166/jctn.2020.9049

9. Satyananda V, Anithalakshmi KC (2019) Diagnosis of diabetic retinopathy using machine learning techniques and embedded systems. Perspect Commun Embed Signal-Process 2:2566–932. https://doi.org/10.17485/ijst/2015/v8i26/81049
10. Kanimozhi J, Vasuki P, Roomi SMM (2020) Fundus image lesion detection algorithm for diabetic retinopathy screening. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02417-w
11. Huda SMA, Ila IJ, Sarder S et al (2019) An improved approach for detection of diabetic retinopathy using feature importance and machine learning algorithms. In: 2019 7th international conference on smart computing and communications, ICSCC 2019. Institute of Electrical and Electronics Engineers Inc
12. Chetoui M, Akhloufi MA, Kardouchi M (2018) Diabetic retinopathy detection using machine learning and texture features. In: Canadian conference on electrical and computer engineering. Institute of Electrical and Electronics Engineers Inc
13. Azer SA (2019) Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: A systematic review. World J Gastrointest Oncol 11:1218–1230
14. Doshi D, Shenoy A, Sidhpura D, Gharpure P (2016) Diabetic retinopathy detection using deep convolutional neural networks. In: 2016 international conference on computing, analytics and security trends (CAST). pp 261–266
15. Ghosh R, Ghosh K, Maitra S (2017) Automatic detection and classification of diabetic retinopathy stages using CNN. In: 2017 4th international conference on signal processing and integrated networks, SPIN 2017. Institute of Electrical and Electronics Engineers Inc., pp 550–554
16. Saranya P, Prabakaran S (2020) Automatic detection of non-proliferative diabetic retinopathy in retinal fundus images using convolution neural network. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02518-6
17. Hagos MT, Kant S (2019) Transfer learning based detection of diabetic retinopathy from small dataset. arXiv
18. Gangwar AK, Ravi V (2021) Diabetic retinopathy detection using transfer learning and deep learning. In: Advances in intelligent systems and computing. Springer, pp 679–689
19. Peng Y, He X, Zhao J (2018) Object-part attention model for fine-grained image classification. IEEE Trans Image Process 27:1487–1500. https://doi.org/10.1109/TIP.2017.2774041
20. Li R, Zheng S, Duan C et al (2021) Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2021.3063381
21. Choe J, Lee S, Shim H (2020) Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.2999099
22. Vaswani A (2017) Attention Is All You Need arXiv:1706.03762v5. Advances in neural information processing systems, 2017-Decem 5999–6009
23. Mnih V, Heess N, Graves A et al (2014) Recurrent models of visual attention. In: Advances in neural information processing systems, p 27
24. Zhao Z, Zhang K, Hao X et al (2019) Bira-net: bilinear attention net for diabetic retinopathy grading
25. Lin Z, Guo R, Wang Y et al (2018) A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 11071 LNCS, pp 74–82. https://doi.org/10.1007/978-3-030-00934-2_9/COVER
26. Li X, Hu X, Yu L et al (2020) CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular Edema grading. IEEE Trans Med Imaging 39:1483–1493. https://doi.org/10.1109/TMI.2019.2951844
27. He A, Li T, Li N et al (2021) CABNet: category attention block for imbalanced diabetic retinopathy grading. IEEE Trans Med Imaging 40:143–153. https://doi.org/10.1109/TMI.2020.3023463
28. Kingma DP, Ba JL (2014) Adam: a method for stochastic optimization. In: 3rd international conference on learning representations, ICLR 2015: conference track proceedings
29. Ahmad Z (2017) Zafarali Ahmed—Medium. https://medium.com/@zafarali. Accessed 11 Dec 2021
30. Ben-David A (2008) Comparison of classification accuracy using Cohen's weighted Kappa. Expert Syst Appl 34:825–832. https://doi.org/10.1016/J.ESWA.2006.10.022
31. Dondeti V, Bodapati JD, Shareef SN, Naralasetti V (2020) Deep convolution features in non-linear embedding space for fundus image classification. Rev d'Intell Artif 34:307–313. https://doi.org/10.18280/ria.340308
32. Kassani SH, Kassani PH, Khazaeinezhad R et al (2019) Diabetic retinopathy classification using a modified xception architecture. In: 2019 IEEE 19th international symposium on signal processing and information technology, ISSPIT 2019. Institute of Electrical and Electronics Engineers Inc.
33. Wang L, Schaefer A (2020) Diagnosing diabetic retinopathy from images of the eye fundus. cs230.stanford.edu
34. Bodapati JD, Shaik NS, Naralasetti V (2021) Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02727-z
35. Adriman R, Muchtar K, Maulina N (2021) Performance evaluation of binary classification of diabetic retinopathy through deep learning techniques using texture feature. Procedia Comput Sci 179:88–94. https://doi.org/10.1016/J.PROCS.2020.12.012

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.