Intelligence-Based Medicine 3-4 (2020) 100022

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets

Gaurav Saxena [*], Dhirendra Kumar Verma, Amit Paraye, Alpana Rajan, Anil Rawat

Raja Ramanna Centre for Advanced Technology, Department of Atomic Energy, Indore, 452013, M.P, India

ARTICLE INFO

Keywords:
Diabetic retinopathy
Automated screening
Fundus images
Early detection
Convolutional neural networks
Adjudicated datasets

ABSTRACT

Diabetic Retinopathy (DR) is one of the leading causes of preventable blindness in the working-age diabetic population in India and across the world. It may lead to permanent blindness if not detected in the early stages. The prevalence of DR among diabetics in India was 10% and 16.9% in 2014 and 2019, respectively. In 2019, the International Diabetes Federation estimated that Diabetic Mellitus will affect 101 million people in India in 2030; the largest number in any nation in the world. Our work is an attempt to speed up preliminary screening of DR to cater to the future requirement of such a huge amount of diabetic patients. We have trained and validated robust classification models on publicly available datasets for early detection of DR. We have applied state-of-the-art deep learning models based on Convolutional Neural Networks (CNN), to exploit data-driven machine learning methods for the purpose. We framed the problem as a binary classification for the detection of DR of any grade (Grade 1–4) vs No-DR (Grade 0). We used 56,839 fundus images from the EyePACS dataset for training the models. The models were tested on a test set from EyePACS (14,210 images), benchmark test datasets Messidor-2 (1748 images) and Messidor-1 (1200 images). The model has achieved an AUC of 0.92 on benchmark test dataset Messidor-2 with sensitivity and specificity of 81.02% and 86.09%, respectively. AUC, Sensitivity and Specificity on Messidor-1 are 0.958, 88.84% and 89.92%, respectively. In this paper, we also discuss challenges of automated ailment detection in medical images using CNNs, such as the use of public datasets for training, pre-processing methods, performance metrics for unbalanced classes and present our results and their comparison with leading studies. The developed preliminary automated screening system will act as an aid to the manual diagnostic process by referring DR patients to an ophthalmologist for further examination (if detected positive) well in time to reduce the risks of vision loss.

1. Introduction

Diabetic Retinopathy (DR) occurs due to high blood sugar levels in diabetic patients. It damages the retina of the patient by making the blood vessels abnormal (leak or swell) at the posterior pole (backside) of the eye. It can lead to permanent vision loss if not detected in the early stages [1].

India had 50, 62.4, 69.2 and 77 million cases of diabetes in the year 2010, 2011, 2015 and 2019, respectively. The numbers and growth rates are very high as compared to the global average [2,3,4]. Nearly one-third of this population is likely to have diabetes-related complications such as DR [5]. The prevalence of DR was 18% in the year 2009, 21.7% in 2014 and 16.9% in 2019 [6,5,7]. As per the various studies reviewed in [3], in 2015, the DR prevalence was 13–18% in the urban Indian population and 9–10% in the rural Indian population. The International Diabetes Federation has estimated (in 2019) that India will have around 101 million cases of diabetes in the year 2030. For the year 2045, the same estimate is about 134 million cases [4].

In India, the prevalence of sight-threatening DR is 3.6% and among diabetic patients, the prevalence of blindness and visual impairment is 2.1% and 13.7%, respectively. The main reason for such a high rate of prevalence is found to be poor awareness. Around 90% of the diabetic patients have never gone for fundamental evaluation for DR [7].

DR and diabetes are closely related. However, DR is not an infection but can be considered as a sign of uncontrolled diabetes. Both DR and diabetes are not mutually inclusive.

The severity of DR is graded as per the International Clinical Diabetic Retinopathy Scale [8]. The grades help to determine the need for referral, frequency of monitoring/screening, treatment, etc. As per the guidelines [9], short screening intervals (from six months to one year) are

[*] Corresponding author.
E-mail address: saxenagaurav@rrcat.gov.in (G. Saxena).

https://doi.org/10.1016/j.ibmed.2020.100022
Received 18 June 2020; Received in revised form 7 November 2020; Accepted 16 November 2020

recommended in most of the DR cases. These intervals are reduced to few weeks in the cases of severe or worse DR. Manual screening process consists of diagnosing DR through visual assessment of fundus either by direct examination (in-person dilated eye examinations) and/or by evaluation of digital colour fundus photographs of the retina. This process is time-consuming and expensive. The problem becomes worse in rural areas where access to trained clinicians is limited. Also, the number of growing DR patients raises concerns about the sustainability of manual screening in the long run.

Frequent screening of DR patients and their exploding growth rate in India advocates the requirement of an automated screening system for early detection of DR. Early detection, frequent screening, and timely treatment are the most critical components that should be addressed by an automated system to prevent vision loss.

Contemporary deep learning methods such as Convolutional Neural Network (CNN) appears to be the perfect choice for automated detection of ailments in digital medical images [10–13]. This happened after AlexNet [14] won the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge). As of January 1, 2020, the work has been cited more than 53,844 times. Since then CNNs have become the de-facto standard in solving image recognition tasks and won all the later versions of ILSVRC (from 2012 till 2019). The performance of CNNs has improved with the advent of supportive tools like activation functions such as Rectified Linear Units [15], Dropout [16] regularization, Batch Normalization [17], etc.

In this work, we have developed an agent for the classification of digital colour fundus images of the retina for detecting the presence of DR. The problem is framed as a binary classification. The agent comprises of multiple models that have been developed via data-driven approaches of Machine Learning. We have applied state-of-the-art CNN models such as InceptionV3, InceptionResNetV2, etc. and trained them over publicly available EyePACS dataset. The same has been validated over another set of publicly available images from EyePACS and Messidor-1 & 2 datasets.

The main outcomes of this work are insights at (a) pre-processing methods (b) hyper-parameters of models and their tuning (c) effectiveness of ensemble learning methods (d) trained model capable for detection of DR and (e) web-based prototype for end-user. We also discussed the relevance of different performance metrics while dealing with the problem of unbalanced classes while solving classification problems.

The results that we achieved are encouraging and stand at a very good position among other leading studies carried out worldwide on similar datasets.

2. Related works

We have reviewed studies that involve CNNs and contemporary deep learning-based methods to detect and classify ailments in fundus images. A few studies that exploit computer vision methods and make use of traditional machine learning for the purposes are also reviewed.

A deep learning-based system named IDx-DR X2.1 was developed by Abràmoff et al. in [11] using multiple CNNs trained over fundus images of publicly available datasets for automatic detection of lesions. The images in the dataset were graded by multiple human experts and adjudicated before training. The system provided four types of outputs including Negative (implying no or only mild DR present), rDR (implying referral DR is present), and vtDR (implying vision-threatening DR is present). The results achieved by the system include 96.8% sensitivity, 87% specificity, and AUC of 0.980 for the detection of rDR on the Messidor-2 dataset.

In [12] customized deep CNN model with the principle of deep residual learning was used. The output of the model was 0 for no DR or 1 for DR of any severity level. The pre-processing techniques including rotational invariance, contrast invariance and brightness adjustment were applied to the training set. A 5-fold stratified cross-validation was also applied during initial training and then average metrics were derived and fed during the final training. The model was able to secure 93% sensitivity and 87% specificity with an AUC of 0.94. Public dataset Messidor-2 was used as a validation set.

A deep learning system (DLS) for screening diabetic retinopathy and related eye diseases was developed by Daniel ShuWei Ting et al. in [18]. The system was trained over 76,370 retinal images obtained from a community-based national diabetic retinopathy screening program in Singapore (SIDRP 2010–2013). The primary validation set used was taken from the ongoing DR screening program SIDRP 2014–15 with a total of 71,896 images, including some poor quality ungradable images. The sensitivity of 90.5% and specificity of 91.6% with an AUC of 0.936 for detecting referable diabetic retinopathy was achieved.

Very good results were obtained by Varun Gulshan et al. in [13] by developing a deep CNN-based system trained over 1,28,175 retinal images for the development dataset. The performance of the algorithm was evaluated by two different validation sets which were EyePACS-1 and Messidor-2 including 9963 and 1748 images respectively. All the images from training and validation sets were graded multiple times by the panel of ophthalmologists. The AUC of 0.991 and 0.990 was achieved on Eyepacs-1 and Messidor-2 validation sets respectively with very high sensitivity and specificity.

In [19] the impact of adjudicated DR grades was shown by Krause J et al. Errors made by individual graders during DR screening of Retinal fundus images can be better quantified through adjudication. Kappa score was measured for quantifying agreement between different graders as well as between the graders and the algorithm on a nominal scale. It was observed that a small set of adjudicated DR grades allows substantial improvement in the performance of the DR screening program. The algorithm was found to improve in terms of AUC from 0.934 to 0.986 for moderate or worse DR cases.

A random forest-based classifier for segmentation of true haemorrhages as well as discriminating vessels from haemorrhages was presented in [20] by Garima Gupta et al. The classifier achieved 82% sensitivity with 10 fold cross-validations on 191 images obtained from 58 diabetic subjects having different degrees of pathological severity. The major observation was that sensitivity increases for candidates having big haemorrhages and the variability in morphological and other image features such as appearance, colour, texture improves the confidence score for identifying true haemorrhage candidates.

After examining above related studies the specific highlights of our work include the use of only CNN models right from training to producing the outcome of classification. We used InceptionResNetV2 as a base model to build the classifiers. The pre-processing methods we applied include only rotational invariance, random vertical flips and brightness adjustment. During the training of models random-split approach was followed instead of k-fold cross-validation. The resolution of the training images we have taken is 512 × 512 pixels. Finally, the largest difference between our work and others is that the labels used in our training and validation set are not adjudicated by any human experts.

3. Methodology

The problem has been defined as binary classification with the question being answered as "Whether the given fundus image has signs of Diabetic Retinopathy?" The problem has been addressed by developing an intelligent agent that can find patterns in fundus images and classify the images as "the presence of DR" (BDR Grade 1–4) or "No-DR" (BDR Grade 0). The patterns may range from counting, localization and/or detection of microaneurysms, hard exudates, haemorrhage, cotton wool spots, etc. [21,22].

Fig. 1 shows two sample images from EyePACS data available through Kaggle [21]. Image (b) has BDR Grade 4 while image (a) has BDR Grade 0.

The detection of DR as a binary classification problem means an image is classified as positive if it belongs to BDR Grades 1, 2, 3, or 4, and negative otherwise as depicted in Table 1.

We have followed the data-driven approach of machine learning for

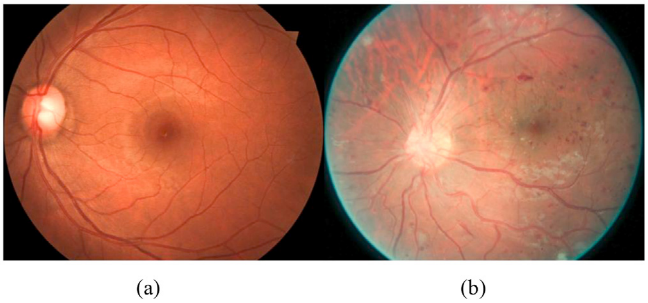


Fig. 1. (a) BDR Grade 0 vs (b) BDR Grade 4 fundus image [22]. The images are cropped (black border removed) and centered for illustration purposes.

Table 1
DR severity levels.

DR Grade	Severity Level	Expected outcome of our agent
0	No DR	Negative
1	Mild Non-proliferative Retinopathy	Positive
2	Moderate Non-proliferative Retinopathy	Positive
3	Severe Non-proliferative Retinopathy	Positive
4	Proliferate Retinopathy	Positive

developing the agent. We built multiple machine learning models (CNN) that were trained and validated on the publicly available dataset EyePACS. The dataset contains fundus images as well as their respective DR grades. Once the models were trained, i.e. their performances were up to the mark and were able to generalize their learning to never-seen-before examples; we put them under test from another set of publicly available datasets (Messidor-1 & 2). The complete process of model hyper-parameters tuning lies between optimization (overfitting/underfitting) and generalization. Fig. 3 depicts the complete process. The first part depicts the training process while the latter shows the working of the agent.

We used CNN model with a layered architecture. It comprises an Input Layer, followed by the blocks of Convolutional, Activation and Pooling Layers. The blocks contain Normalization and Dropout Layers between them. We treated RGB images as three-dimensional arrays, where every value represents pixel intensity. While training the model, we fed the images to CNN in batches. These images have undergone a series of successive transformations by the layers of the model. These transformations make use of the pixel values of the images and current weights of the model to produce predictions. Then Loss Function (binary cross-entropy) compares the predictions with True Labels and generates a Loss Score. This Loss Score is utilized by the Optimizer (RMSProp) to update the weights (Back-Propagation) of layers of the model. We have used Sensitivity, Specificity, Recall, F1 score and AUC to measure the performance of the model. We aim to improve the model's performance over the training set optimization and tune the model so that it can generalize over never-seen-before examples.

3.1. Data sources

3.1.1. EyePACS

The dataset is publicly available through Kaggle's data science competition [21]. The dataset is made available by EyePACS [22]. It comprises 88,702 fundus images, each labeled on the scale from 0 to 4 (0-No DR, 1-Mild, 2- Moderate, 3-Severe, 4-Proliferate) as per International Clinical Diabetic Retinopathy Scale [8]. It has been claimed by [23], that only 75% of the images in this dataset are gradable. Fig. 2 represents a few samples of non-gradable images from the dataset.

3.1.2. Messidor-1 & 2

These publicly available datasets [24,25] are provided by Messidor program partners. The dataset has been established to facilitate studies

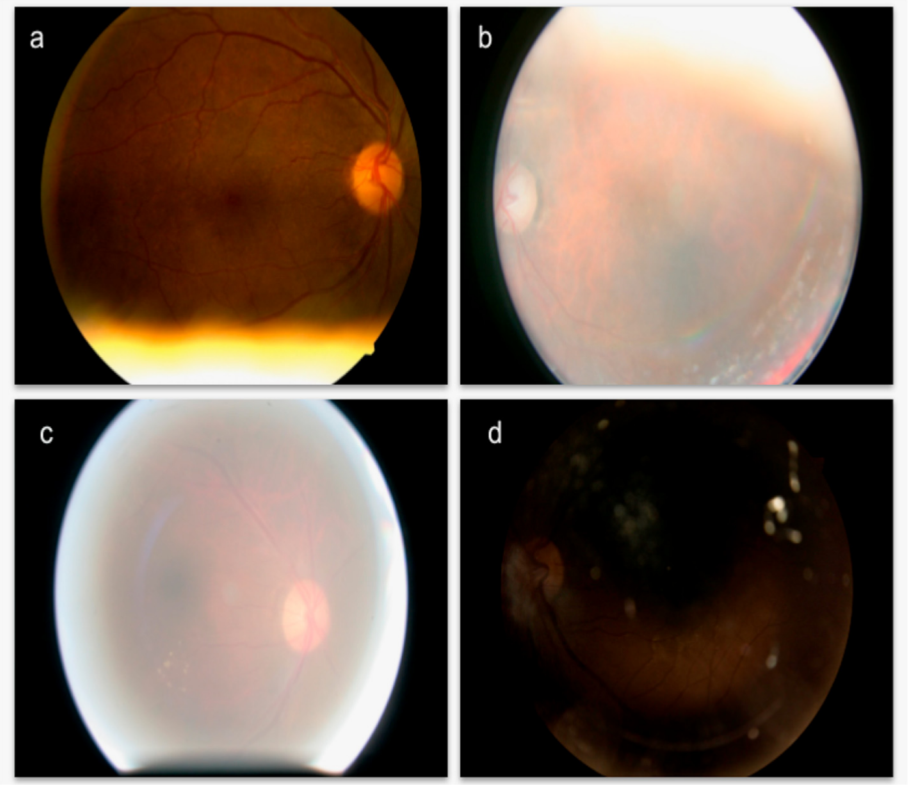
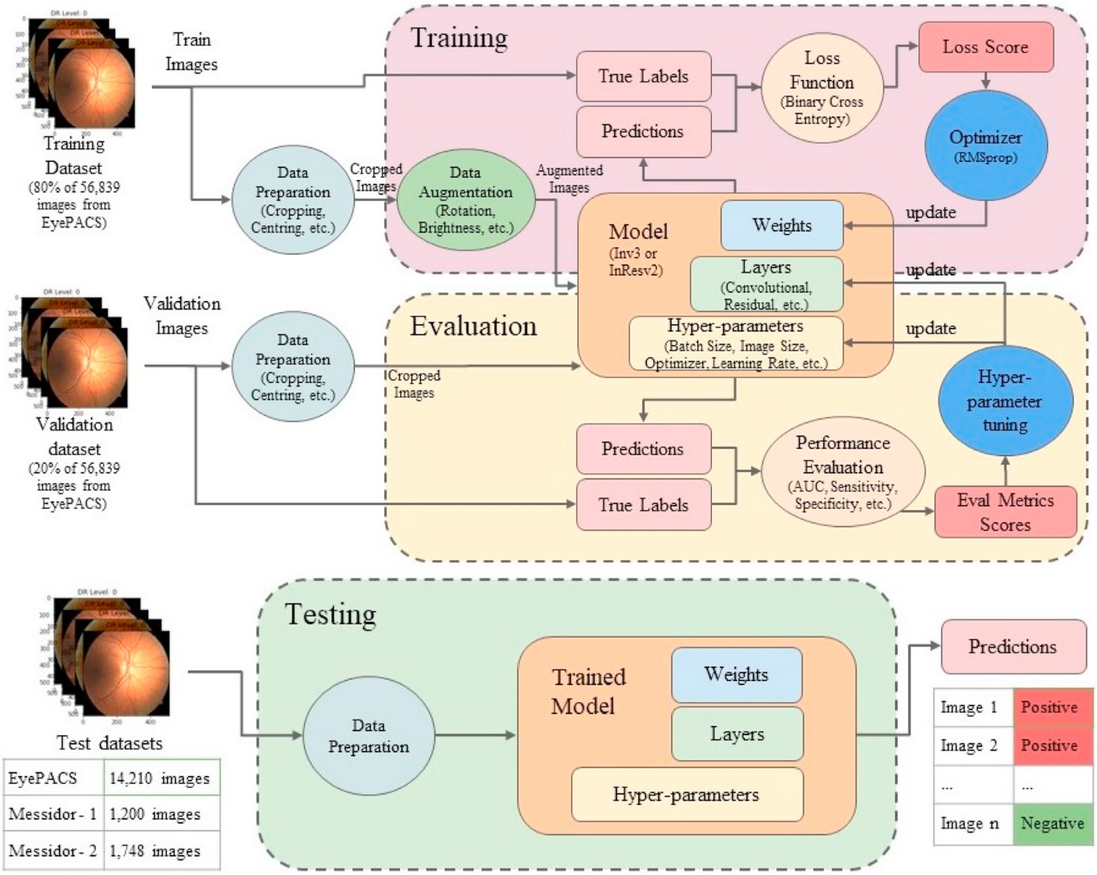


Fig. 2. Examples of non-gradable images from EyePACS. These images contain artifacts (a), out of focus (b), overexposed (c) and underexposed (d) images.


Training
Train Images
Training Dataset (80% of 56,839 images from EyePACS)
True Labels
Predictions
Loss Function (Binary Cross Entropy)
Loss Score
Optimizer (RMSprop)
Data Preparation (Cropping, Centring, etc.)
Cropped Images
Data Augmentation (Rotation, Brightness, etc.)
Augmented Images
Model (Inv3 or InResv2)
Weights
Layers (Convolutional, Residual, etc.)
Hyper-parameters (Batch Size, Image Size, Optimizer, Learning Rate, etc.)
update
update
update
Evaluation
Validation Images
Validation dataset (20% of 56,839 images from EyePACS)
Data Preparation (Cropping, Centring, etc.)
Cropped Images
Predictions
True Labels
Performance Evaluation (AUC, Sensitivity, Specificity, etc.)
Eval Metrics Scores
Hyper-parameter tuning
Testing
Test datasets
Data Preparation
Trained Model
Weights
Layers
Hyper-parameters
Predictions

EyePACS	14,210 images
Messidor - 1	1,200 images
Messidor - 2	1,748 images

Image 1	Positive
Image 2	Positive
...	...
Image n	Negative

Fig. 3. Workflow diagram of the intelligent agent. The above part depicts the training process while the latter shows the working of the trained agent.

on the computer-assisted diagnosis of diabetic retinopathy.

Messidor-1 has 1200 publicly available eye (RGB) images. These images were acquired by 3 ophthalmologic departments in which 800 images were acquired with pupil dilation and 400 without dilation. Two diagnoses (labels) have been provided (a) Retinopathy grade (0, 1, 2 & 3) and (b) Risk of macular edema (0, 1 & 2).

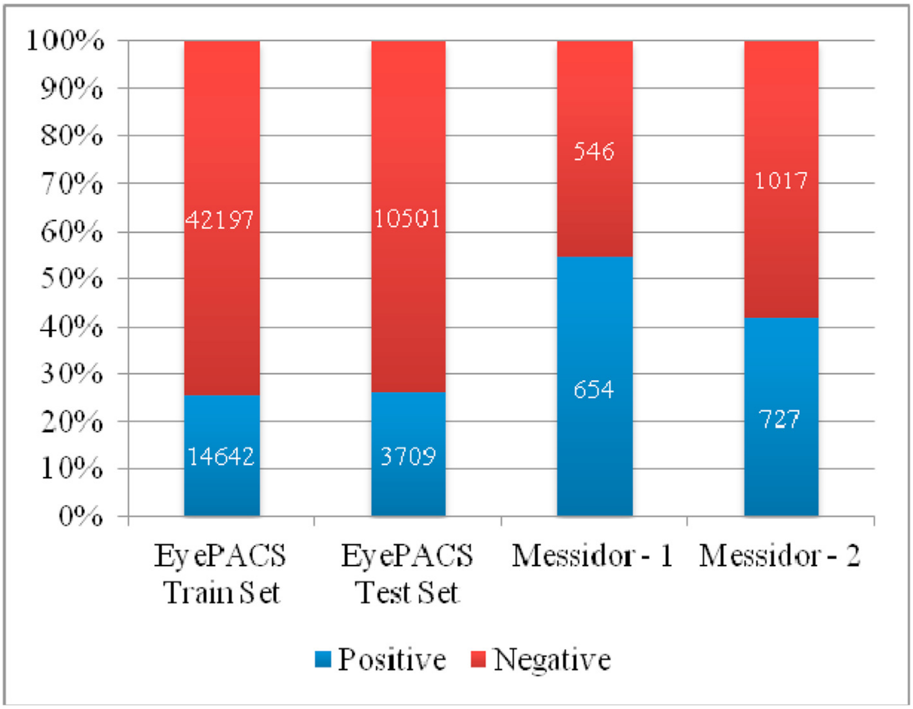


Fig. 4. Distribution of DR grades in datasets.

Messidor-2 is an addition over Messidor-1 with images from examinations from the Brest University Hospital, France. A total of 1748 images are provided, out of which, 1058 images are from the original Messidor dataset and 690 images were captured at the Ophthalmology department of Brest University Hospital (France) between October 16, 2009, and September 6, 2010. The images were taken from the non-mydriatic fundus cameras with a 45-degree field of view.

3.1.3. Exploratory data analysis

Fig. 4 shows the distribution of DR grades among different datasets. Out of 88,702 images from the EyePACS dataset provided by Kaggle, we selected 71,049 gradable images. The selection process was based on (a) gradability grades as supplied by [26], and (b) non-detection of the retinal circular disk by our algorithms for pre-processing. Out of the 71, 049 gradable images, we randomly sampled 14,210 images and kept aside as "EyePACS test set". We termed the remaining 56,839 images as "EyePACS train set". During the process of training, models had split further the "EyePACS train set" into the train (80%) and validation (20%) sets.

Both the sets that we have created from the EyePACS dataset contain about 26% of positive images (DR grades 1–4). Messidor-1 has 54% while Messidor-2 comprises 41% positive images. It is observed that almost all datasets are unbalanced.

To address the issue of class imbalance, we used appropriate metrics like sensitivity, specificity and AUC to measure the performance of models. These metrics quantify the ability of models to distinguish among different classes. The "true positive rate" (sensitivity) in conjunction with the "true negative rate" (specificity) indicates how sensitive (or specific) the models are toward the positive (or negative) class. The details in section 3.3 and Appendix.

We also used class weights for the said purpose. The loss function (binary cross-entropy) weighted by the class weights emphasize minority class to make classifiers learn equally from both the classes. The said weighting improved the AUC.

Algorithmic ensemble techniques improved our results too. We trained multiple models on different versions of train and validation sets extracted (via random splitting) from the EyePACS train set and combined their outcomes through an ensemble.

We also experimented with the architecture of the models to address the issue. We used CNNs as feature extractors and tested Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) and Artificial Neural Networks (ANN) as recognizers. We found that ANNs perform best in our case.

We found that over-/under-sampling did not improve results for this dataset.

3.2. Data preparation and pre-processing

The images in the dataset were taken from multiple camera models under different lighting conditions. It affects their visual appearances. The other variations include off-centered circular disk, inverted, under-/over-exposure and variations in their sizes; both, in terms of vertical (289–3456 pixels) and horizontal axis (400–5184 pixels). To make CNN models achieve a high level of recognition and generalization, we pre-processed the images such that there should not be any irrelevant information. Here, removal of irrelevant information means (a) removal of black border and centering the circular disk, (b) introducing variations in the images during the process of training. The variations helped models to learn to ignore irrelevant features based on lightning conditions (brightness, contrast, etc.), the position of the eye (left or right), rotation of the eye, exposure of the images, etc. and other dataset-specific features. In this way, the models focus their learnings on relevant features based on microaneurysms, hard exudates, haemorrhage, cotton wool spots, etc.

Pre-processing was carried out in two steps; before (offline) and during the process of training (online). While carrying out offline pre-processing, we cropped the images to square shape so that the circular disk comes at the center and black borders were removed. Then we resized them to multiple resolutions, viz. 256, 299, 512 and 598 pixels.

Online pre-processing was carried during the training process along with image augmentation. To introduce variations during image augmentation, we applied the following:

- Rotation of images by a random angle between 0 to 10°.
- Random vertical flips
- Variation of brightness by random (uniform distribution) factor β, represented by the following equalization.

  $$X'_{i,j} = X_{i,j} + \beta$$

  Where β ranges between -0.3 to 0.3 and $x'_{i,j}$ and $x_{i,j}$ represents the new and the old value of the pixel.
- Variation of contrast for every colour channel (R, G & B) separately by a random uniform distribution factor α, between 0.8 and 1.2; represented by the following equation.

  $$X'_{ch\ i,j} = \left(X_{ch\ i,j} - X_{ch}\ \mathrm{mean}\right) * \alpha + X_{ch}\ \mathrm{mean}$$

  Where $x'_{ch\ i,j}$ and $x_{ch\ i,j}$ represent the new and old value of the pixel respectively. x_{ch} mean is the mean of a particular channel.
- We normalized the images between *0.0* and *1.0.*

During the process of testing, only offline pre-processing was done. It includes cropping, resizing and normalizing. Hence, online pre-processing does not affect the inference time at the time of testing.

3.3. Performance metrics

In classification problems, there are four possible outcomes of predictions viz. True Positive (TP, predicted positive actual positive), False Positive (FP, predicted positive actual negative), True Negative (TN, predicted negative actual negative) and False Negative (FN, predicted negative actual positive). In most of the cases, the performance parameters such as Accuracy, Precision, Recall (a.k.a. Sensitivity) and F1 score [27] can tell about the model's performance. Since our datasets are imbalanced by nature, we have chosen sensitivity and specificity as a performance measure. Further details about the performance metrics are available in the Appendix.

For obtaining further insights into classification performance, we have analyzed the Receiver Operating Characteristic (ROC) curve [28]. ROC curve is a graph between Sensitivity (True Positive Rate) vs 1-Specificity (False Positive Rate). We found that Area under the ROC curve (AUC) is a de-facto standard in medical ailments detection/classification problems. Moreover, Sensitivity and Specificity are the preferred metrics to use in real clinical conditions.

Sensitivity and specificity are the yin and yang of the testing world. Sensitivity and specificity work in tandem. Together they can give critical information about the goodness of the test. A test with high sensitivity and low specificity will be able to detect (almost) all the positive cases correctly but has a problem of detecting more number of false-positive cases. It means that a number of negative people will be detected positive by the test. Similarly, for a test with high specificity and low sensitivity, more positive cases will go undetected. The classifier can be made more specific (or sensitive) by suitably choosing this cut-off as desired by clinical conditions. More on this in the Appendix.

It is expected that Sensitivity, Specificity, and AUC should be as close as possible to 1.0. For operating point selection on the ROC curve, we have used the shortest distance metric, i.e. the point on the curve which has minimum Euclidean distance from the point (0, 1).

3.4. Model and hyper-parameters

Initially, we trained Inception (v3) [29] based models. The performance of these models is superior as compared to traditional CNN models. The concepts such as multiple filter size at the same level, dimensionality reduction by *1 x 1* convolution, factoring *n x n* convolutions into *1 x n* and *n x 1*, RMSProp, Batch Normalization in Auxiliary Classifiers, Label Smoothing, etc. make Inception based models performance better, both in terms of speed and accuracy. We had an apprehension that Inception based models may not produce optimal results under limited resources. This is because the models exploit the concept of multiple size filters at the same level to cater to the requirement of the right kernel size due to variation in localization of features in an image. This is not a necessary case with fundus retina images. We observed that after necessary pre-processing the size of features in the images is almost the same across the dataset.

We have also trained InceptionResNet (v2) [30] based model to exploit the concepts of deep residual learning which is inspired by ResNet [31] and Inception (v4) [30]. The model contains hybrid inception modules in which residual connections add the output of the convolutional operation of the module to its input. It was claimed in [30] that the InceptionResNetV2 model can achieve higher accuracy and is easier to optimize.

A table containing the complete hyper-parameter space is available in the Appendix.

During the process of hyper-parameter tuning, with Grid Search, we have found optimal values of the parameters like image size: 512, optimizer: RMSProp, batch size: 16.

3.5. Training

From the entire collection of images from the EyePACS dataset, a test set was extracted and kept aside. This test set was used only after the training process was over. It contains 14,210 number of images. The remaining images in the dataset were split in the ratio of 80:20 to generate training and validation sets. This division follows a random-split approach for multiple training runs. Multiple models were trained using a multi-stage training process. For every run, training was performed using a training set and hyper-parameter tuning was performed using a validation set.

Training starts with an initial learning rate of 10^{-3}. Early stopping was used with a delta of 10^{-4} and patience 10. The model was checkpointed at every epoch saving the best weights at every epoch. The monitoring parameter was the validation set AUC. All the models were early stopped in the range 50–70 epochs. Giving too much training to the model may not necessarily improve its performance rather it may lead to other issues such as overfitting, exploding gradients, etc [32]. In the next stage, the learning rate was reduced by a factor of 10^{-1} and this process was repeated two times. For every running instance, the best weights in the previous stage were used to initialize weights in the next stage. The output of the training process was multiple trained models with multiple weights. The pseudocode for the process of training along with pre-processing and testing is available in the Appendix. A high performance computing cluster Kshitij-5 [33] has been used for the work. Refer Appendix for further details.

3.6. Ensemble learning

The individual best scores from each of the trained models were ensembled (by taking an average) to produce the final results. Among all the weights that an individual model has produced (by following "early-stopping", "check-pointing" and "reduce-learning"), the individual best score has been generated by searching for the weight that performs best on test sets. We identified a few best-performing weights from multiple models and ensembled them by averaging their output. The best combination of weights to produce an optimal ensemble was obtained with the help of a brute force search. One can also use a weighted average ensemble as an alternative to selecting the weights via a brute force search. More details on ensemble methods are presented in the Appendix.

4. Results

Messidor-1 & 2 datasets are utilized by multiple studies carried out worldwide. They are extensively used in the comparison of classification algorithms as external independent benchmark datasets. We trained multiple models on the EyePACS train set. Each model has trained on a different train (80%) and validation sets (20%), selected from the EyePACS train set using random splitting. We combined the results of these models by using ensemble averaging methods. The ensemble of models has evaluated on separate datasets other than the dataset used for training; i.e. EyePACS test set, Messidor-1 & 2 test sets. Our algorithm scored an AUC of 0.958 and 0.92 on Messisor-1 and Messidor-2, respectively. We have used the shortest distance as the operating point selection strategy for the ROC curve. Sensitivity and specificity of 88.84% and 89.92%, respectively, have been achieved at this operating point for Messidor-1. We obtain Sensitivity and Specificity for Messidor-2 as 81.02% and 86.09%, respectively. For the EyePACS test dataset from Kaggle (14,210 images), our algorithm has scored an AUC of 0.927, Sensitivity and Specificity of 83.74% and 89.65%, respectively at the mentioned operating point. The ROC curves for the same are available in Fig. 5. The dotted line in Fig. 5(a), (b) and (c) represents a random chance classifier while the curve represents the model's trade-off between True Positive Rate and False Positive Rate. The dot on the curve represents the shortest distance threshold.

In Tables 2 and 3, we present the binary classification results of various studies. These studies have grouped the images with DR grades (0, 1, 2, 3, 4) into two classes. We have created binary classes as (0) vs (1, 2, 3, 4) for preliminary DR detection. Studies like [11,12,19,26] have created these classes as (0, 1) vs (2, 3, 4) for referral DR detection [11].

Table 2 represents the results of the studies (including our study) that used non-adjudicated datasets for training.

Table 3 shows the results of the studies that have used adjudicated datasets.

The studies like [13,26] have two pairs of values of sensitivity and specificity for each dataset. One pair is at the high-sensitivity operating point while the other one is at the high-specificity operating point.

We get interesting insights when we compare our results with the other studies listed in Table 2. While comparing sensitivity and specificity, we should take note that there is a tradeoff between these two metrics. Cutoff at the desired operating point governs this tradeoff. Sensitivity increases with a decrease in specificity and vice versa.

We find that for Messidor-1 and Messdor-2 datasets, our AUC values are higher than the closest results [26,34] by ~ 7.8% and ~6.4% respectively. Also, the values of our pair of sensitivity and specificity are better than the values of the corresponding pair of [26,34]. We can easily see that values of our sensitivity and specificity can be made pairwise higher against these studies if we shift our operating point. This is an indication of better classification performance of our models over benchmark datasets.

For the EyePACS test set, our models have scored an AUC of 0.927 which is similar to the corresponding values on benchmark test sets. The interpretation of these results is presented in the discussion section.

While comparing our results with the studies in Table 3, it is to be noted that we purposefully want to use non-adjudicated public datasets for training. We want to push the models to their limits and see what best can be achieved with these public datasets. Our AUC value (from Table 2)

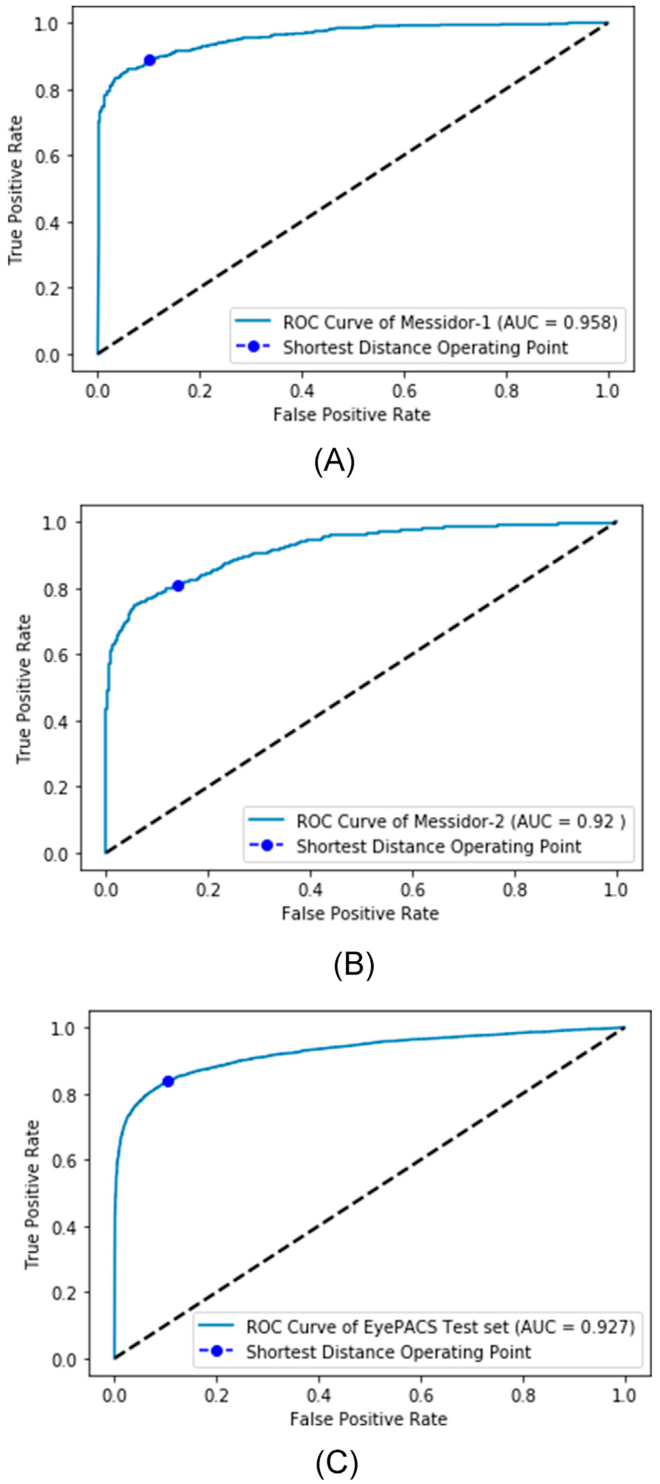


Fig. 5. Shows the ROC curves (a) Messidor-1, (b) Messidor-2 and (c) EyePACS test sets.

Table 2
Results of the studies that have used non-adjudicated datasets.

	Test Dataset	Sensitivity	Specificity	Area under the Receiver Operating Characteristic Curve
This Study	EyePACS (Kaggle)	84.74%	89.65%	0.927
	Messidor-1	88.84%	89.92%	**0.958**
	Messidor-2	81.02%	86.09%	**0.92**
Voets et al. [26] (2019)	Messidor-2	81.8% 71.2%	68.7% 88.5%	0.853
Seoud et al. [34] (2016)	Messidor-1	94%	50%	0.90

Table 3
Results of the studies that have used adjudicated datasets.

	Test Dataset	Sensitivity	Specificity	Area under the Receiver Operating Characteristic Curve
Abràmoffet. al [11]. (2016)	Messidor-2	96.8%	87%	0.98
Gulshan et al. [13] (2016)	Messidor-2	96.1% 87%	93.9% 98.5%	**0.990**
Gargeya et al. [12] (2017)	Messidor-2	93%	87%	0.94

for Messidor-2 is ~2.1% less than the nearest AUC value [12] and only ~ 7.0% less than the highest AUC value [13] in Table 3.

We expect that adjudication will increase the performance of our models as it reduces errors in DR grading. We present the detailed effects of adjudication on the performance of models in the discussion section.

5. Application runtime test

We established the suitability of the developed application for working in clinical settings by measuring its performance during runtime.

The purpose of this test is to check the feasibility of software to work as a preliminary automated DR detection tool in production environments. We tested its response time on commodity hardware (Laptop and Desktop PC with moderate configurations) that are available in any standard clinic. We have performed the same test on a standard server to check the suitability of its working in client-server mode. Table 4 presents the results of the tests.

We find that the response time of our agent is in the range of 0.5 – 2.0 s, which makes it suitable in clinical settings. Fig. 6 depicts the classification of fundus images by the developed web application.

Since we haven't deployed the application in the real clinical conditions yet; for a case study, we perceived the following few salient points w.r.t application's clinical usefulness, ease-of-use, domain knowledge of the operator, etc.

One can deploy the application over a range of hardware types; from purpose-built hardware like smartphones [12], laptops, etc. to off-the-shelf hardware like desktop PC, servers, etc. It helps in reducing capital expenditure and increases the reachability of the application.

Such deployment scenarios enable the application to integrate with various online and offline image capturing system. It is suitable to work even with the camera of a smartphone [35] (supported with lenses). It enables the application to work on smartphones as an offline integrated system app.

Since the application works independently with the image capturing system, the requirement of domain knowledge of DR for the operator of the application is minimal. Transferring images to the system, uploading images and generating automated results required basic operating knowledge of computers. However, pupil dilation and image capturing still require domain skill expertise.

Batch processing of thousands of images for grading increases ease-of-use to an even greater extent. Generating Class Activation Mapping as a part of the application helps non-experts/training students to identify the

Table 4
Runtime application tests.

Hardware	Specifications	Response Time of agent
Laptop	Intel® Core™ i5-7200U Processor, 2.5 GHz, dual-core, 3 MB Cache, 8 GB Memory	1.8–1.9 s
Desktop PC	Intel® Core™ i5-4590 Processor, 3.3 GHz, quad-core, 6 MB Cache, 4 GB Memory	1.0–1.3 s
Server	2 x Intel® Xeon® Gold 6142 Processor, 2.6 GHz, 22 MB Cache, 384 GB Memory	0.5–1.0 s

problematic regions of the images.

6. Visualization and analysis of heatmaps

Fig. 7 (a) and (b) shows a heatmap of two fundus image. The images on the right are the heatmaps, while on the left are their corresponding original image.

Heatmaps, a.k.a class activation maps, help in debugging the models. It highlights the pixels that are most relevant in taking the final decision about the class under consideration. Through heatmaps, we can identify the portion of images where the model sees to make decisions.

DR Grade for the image in Fig. 7(a) is 4 and in Fig. 7(b) is 1. The same has been highlighted by their heatmaps. Certainly, the area of relevant portions in grade 4 is more than in the grade 1 image. We can easily identify a few microaneurysms and cotton wool spots in both images.

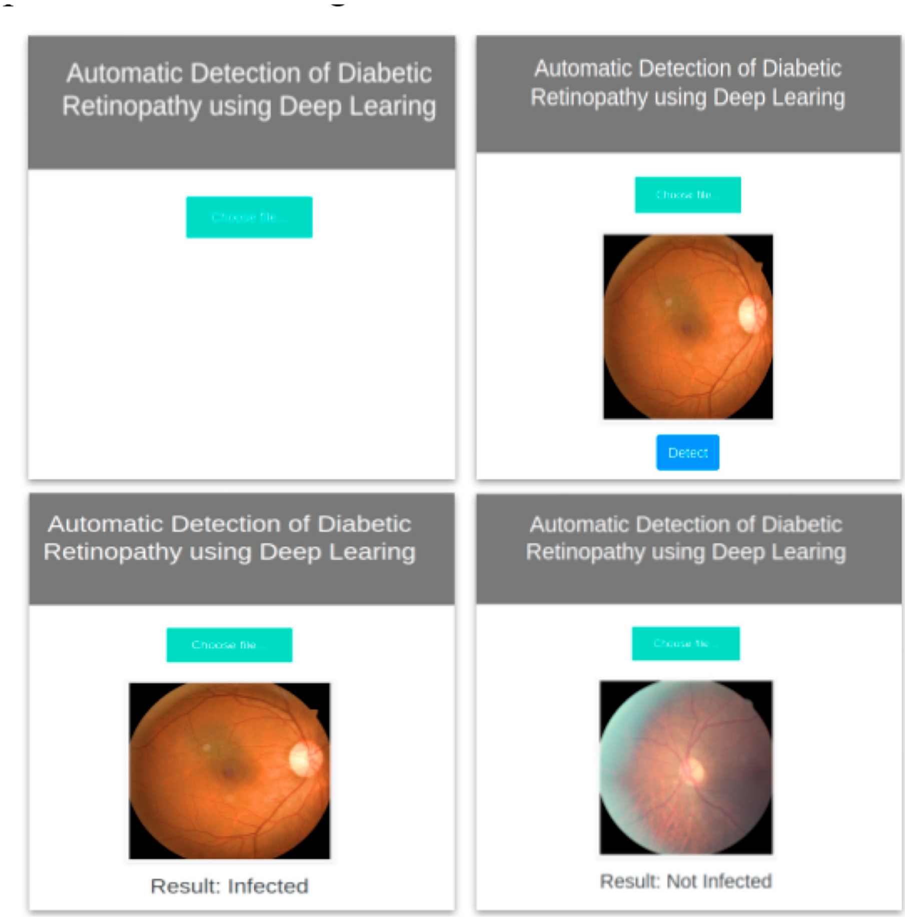
7. Discussion

7.1. Automated DR classification is hard

The problem of automated detection of DR grades in fundus images is a difficult task; at least for some pre-trained CNN models. They find it difficult to perform tasks like (a) extracting features from the images, (b) mapping the extracted features to DR grades (labels), etc. The complex definitions of DR grades; as defined in the International Clinical Diabetic Retinopathy Scale [8], may be held responsible for this. As per the definitions, features for classification of DR grades not only depends on the presence of abnormalities but also on their count, position, size, etc. E.g. fundus image with the presence of intraretinal haemorrhages should be categorized with severe non-proliferate diabetic retinopathy grade only if they are present in all four quadrants.

Most of the pre-trained models are built on the ImageNet dataset. The creators paid more attention to design the models in such a way that their categorization decision is based on the presence/absence of relevant features of the class. The count, position, size, etc. of these relevant features may or may not involve in the classification decision. For effective feature extraction, we must take appropriate measures and customize the models so that these features become relevant.

Complex definitions of DR grades confuse human graders too. When a set of images are graded by multiple human graders, an agreement between the graders can be measured by Cohen's Kappa Score. The value of this score typically lies in the interval from 0 (random agreement) to 1 (complete agreement). In [19], the authors pointed out that in many DR classification related studies, the intergrader kappa score ranged from 0.40 to 0.65, which shows a high level of disagreement among graders. Hence the quality of true labels obtained by human specialists may get


Automatic Detection of Diabetic Retinopathy using Deep Learing

Automatic Detection of Diabetic Retinopathy using Deep Learing

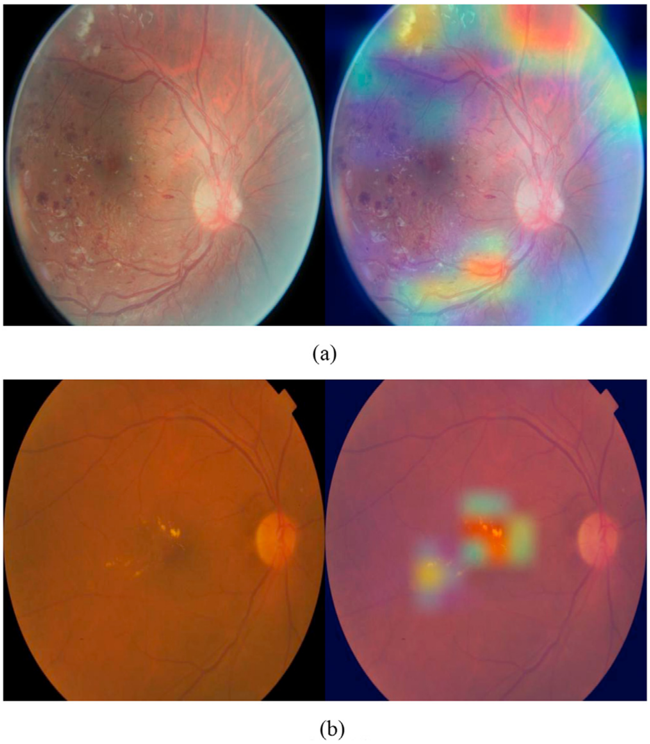
Automatic Detection of Diabetic Retinopathy using Deep Learing

Result: Infected

Automatic Detection of Diabetic Retinopathy using Deep Learing

Result: Not Infected

Fig. 6. Screenshot of web application of the preliminary automated DR detection system.


(a)

(b)

Fig. 7. Heatmaps for (a) DR Grade 4 image and (b) DR Grade 1 image.

affected; if not adjudicated.

The quality of true labels governs the performance of models. During training, models learn to map extracted features to true labels. If true labels contain noise, the models get confused in selecting the relevant features for the classification decision. E.g. If a relevant feature specific to DR grade 3 is present in grade 2 and/or grade 4 labeled images, models learn to reduce the impact of such features in the classification decision of grade 3 images. This reduces overall performance.

In DR classification, noise from the labels can be removed via adjudication consensus among multiple graders. Removal of this noise has a positive impact on performance. The authors of [19] have claimed that the value of AUC has improved from 0.934 to 0.986 for moderate or worse DR when they used a small tuning set of adjudicated consensus grades and higher resolution input images. In [26], the authors report a 36% decline in performance in their “original study” when only one grade per image (without adjudication) for ground truth was used.

7.2. CNN models

We have experimented with different versions of CNNs by training multiple models. These models were based on Inception and Resnet. We have found that InceptionResNetV2 outperforms all the other versions of CNNs. We consider the following to be the reason. Since Inception models are general purpose in nature. They have multiple kernels with different sizes at the same level. They are designed to work well in the conditions where relevant features are of varying size in images across the datasets (ImageNet). This is not the case with fundus retina images, where the relevant features are of similar size in images across the dataset. If we apply Inception models to DR datasets, only a few layers (and parts of layers) of these models may be used for the classification decision. A major portion of the models accounts for the wastage of computing resources. This advocates our intuition that custom models built for the specific purpose (ailments detection in medical images) having less number of layers and fixed-sized kernel may also perform at par with Inception models under limited resources. Our assumption was right and supported by two facts. First, most of the leading solutions in Kaggle competition have a very limited number of layers (approx. 20 including Pooling). Second, in our experiment, InceptionResnetV2 outperforms all others. ResNet applies identity shortcut connections and creates residual blocks. These connections map the output of lower layers directly to the input of the higher layers. It helps in addressing the problem of vanishing gradient and enables models to converge fast. It also helps deeper models to train quickly. Here, reduced training time is due to the quick convergence rather than the reduction in computational time. The input is still processed by all the layers. Moreover, skip connections reduces the number of effective layers. Here, the effective layer means the layer that takes part in the classification decision. It may happen that majority of the decision part come from a few numbers of layers. We also noted that InceptionResNetV2 was published a few months after the Kaggle competition was over.

7.3. Input image size

The size of the input image (resolution) governs the decision-making capabilities of CNN models. It decides the number of parameters of the CNN model to be optimized. This affects the time taken for training and predictions. It also governs the size of kernels at each layer and acts as one of the leading factors for structuring CNN models. In medical images like fundus retina images, higher resolution images will explode the number of CNN parameters. This will make the models computationally more intensive and make convergence slow while training. On the other hand, in low-resolution fundus images, some of the relevant features may vanish. This happens because the important features like lesions, haemorrhages, exudates, etc. cover a very small portion of fundus images [36].

7.4. Comparison with other studies

When we compare our approach with the approaches of other studies listed in Table 2, we consider the difference in (a) Structure of models, (b) Input image size, (c) Methods for splitting train dataset, (d) Methods for image pre-processing and data augmentation, and (e) Ensemble methods, are the significant reasons for improvement shown in our results.

Our approach gives the best results with InceptionResNetV2, while Voets et al. have used InceptionV3. InceptionResNetV2 achieves higher accuracy and easier to optimize as compared to InceptionV3, as claimed in [30]. It may be because InceptionResNetV2 inherits the features of ResNet and InceptionV4.

We have used the image size of 512 × 512 pixels, which is about 193% bigger than the image size of 299 × 299 pixels used by Voets et al. We think that the 299 × 299 pixel image may diminish some of the DR related features as compared to its 512 × 512 pixel counterpart. Thus, it may be easier for kernels at a few initial layers of our models to pick up these features and pass onto higher layers for further processing. Using the image size of 299 × 299 pixels, our models have scored an AUC of 0.91 and 0.86 on Messidor-1 and Messidor-2 dataset, respectively. These AUC values are low in comparison with their 512 × 512 counterpart. Experimental details are presented in the Appendix.

Further, we have noticed that in the aforementioned studies (along with ours), the authors have created train, validation and test sets from the EyePACS dataset provided by Kaggle. The images in these sets may contain dataset-specific features such as cameras, lighting conditions, patient characteristics, etc. Due to these dataset-specific features, the train set of these studies are more closer/similar to their corresponding test set from EyePACS, in comparison to the reference benchmark test sets; i.e. Messidor-1 and Messidor-2. It is expected that a well-generalized model should ignore these dataset-specific features and produce similar AUC values on the EyePACS test set and reference benchmark test sets. For our study, the said differences between the AUC values of the EyePACS test set with Messidor-1 and Messidor-2 are quite small; ~0.03 and ~0.007, respectively. Producing similar AUC values across different datasets indicates that our models are prone to over-fitting.

Given the above, we can conclude that our models have learned just the right set of features; specific to DR and not specific to the EyePACS dataset. We consider random splitting of the EyePACS dataset, minimalistic set of image pre-processing and data augmentation methods, and methods for the ensemble as the significant reasons.

Here we compare our study with the studies listed in Table 3. We present below a few of our insights for these studies regarding the difference in (a) train, validation and test data sets, (b) pre-processing methods, (c) model structures, (d) ensemble techniques, etc.

Regarding the training dataset, we and Rishab et al. have taken public datasets from EyePACS with labels as single DR grade per image. Both the studies defined the problem as a binary classification to distinguish among two classes (DR grade 0 vs 1, 2, 3 & 4). Of course, all the studies listed in Table 3 taken consensus labels prepared through adjudication among multiple human experts. We used the Labels as provided by the EyePACS (Kaggle).

The train set of Gulshan et al. consisted of 1,28,157 fundus images collected from EyePACS (not public) and three eye hospitals from India. The images were graded by multiple ophthalmologists and contained consensus grades for DR, diabetic macular edema, and image quality. Their models had multiple binary outputs for the presence/absence of (a) DR grade 2, 3, or 4, (b) DR grade 3 or 4, (c) macular edema, and (d) were fully gradable.

Abràmoff et al. obtained the images for the train set from the EyeCheck project and the University of Iowa. The dataset had contained multiple consensus labels per image regarding DR grades, macular edema, etc. They trained the classifier to detect/classify (a) No DR; DR grade 0 or 1, and no macular edema, (b) referable DR; DR grade 2, 3 or 4, or macular edema, (c) vision-threatening DR; DR grade 3 or 4, or macular edema, (d) Low exam quality.

Gulshan et al. and Abràmoff et al. are the two most scoring models in Table 3. Regarding their model's input and output, we noted the following as one of the possible reasons for this. Both of them used grades of DR and macular edema in conjunction to determine multiple related binary outputs. It certainly provides more information to the models to learn. Also, it may help in creating more informed feature space so that the models can converge fast and score higher.

Regarding pre-processing methods, we and Rishab et al. have applied a few pre-processing steps such as cropping, rotation/flips, change in brightness, change in contrast, etc. We both have cropped the images to 512 × 512 pixels. In addition to our pre-processing methods, Gulshan et al. used random hue and saturation changes. They used an image size of 299 × 299 pixels.

We have used InceptionResNetV2 as our base model. Rishab et al. used a custom-built model in conjunction with the concept of deep residual learning. Gulshan et al. used InceptionV3 based models while Abràmoff et al. used Alexnet based model to build the classifiers. Among all the standard models stated above, InceptionResNetV2 converges fast.

We, and Gulshan et al. have used an ensemble of models with linear averaging for combining the results of multiple classifiers. We have trained the ensemble members using random splitting of train dataset for improving the skills of models. Rishab et al. used 5-fold stratified cross-validation.

While comparing our results with the studies in Table 3, one should note that we have used single non-adjudication grades per image for training and validation. Still, by using such images, we have obtained AUC which is slightly lower by 2% and 7% as compared to the nearest AUC [12] and highest AUC values [13], respectively. Considering obtaining such close results and expectations to improve it further as per section 7.1, we can say that our results stand competitive with these studies.

8. Conclusion

We have engineered, trained and validated deep learning-based models for the development of a preliminary automated detection system for the detection of Diabetic Retinopathy in colour fundus images of the retina. We structured the models based on Convolutional Neural Networks and trained them over publicly available datasets. Our models achieved notably better results as compared to the previous studies on the same public datasets. It makes Convolutional Neural Networks a leading choice for tasks like automated detection of ailments in digital medical images. Further, less model inference time (~1.5s) and batch processing of images make the said automated system a suitable choice for preliminary screening of a large number of patients. It makes the screening process efficient, addresses the requirement of frequent screening of patients and in-time treatment which prevents vision loss.

There is a scope of improvement regarding the quality of datasets. The quality of public datasets is not-so-good because they may contain noise in images and labels. The images may contain artifacts, are out of focus, over-/under-exposed, etc. Noise in the labels comprises of missing/wrong/inconsistent information. With the help of consensus (through adjudication), one can make the labels consistent and free from noise. However, pre-processing and visual inspection helps to remove noise from images. Datasets used in most of the studies were cleaned and graded by multiple ophthalmologists (consensus). It makes them free from noise in images and labels, thereby enabling classifiers to train and score better on them.

We have experimented with multiple pre-processing methods and hyper-parameters. We found that pre-processing methods like intelligent cropping (remove black border and centering), random brightness change (~12–13%) and random contrast change (~20%) give better results. Also, the optimal image size is nearly 512 × 512 pixels. Lastly, we have found that residual connections in Inception based models significantly improve the results.

9. Future works

We can enhance the performance of classifiers by making custom models built for specific purposes. Our study uses general-purpose models. The creators of these models designed them to work excellent for the conditions where the image contains varying size features. This is not the case with digital medical images. The size of the features in these images is almost the same across datasets. Here, a general-purpose model may or may not perform better under limited resources. So initially, one can use general-purpose models to get optimum size kernels and later can build custom models using those kernels. The depth of the model can also be optimized similarly. This will reduce the number of parameters (weights) to be trained, enhance targeted learning, increase the model's performance and provide greater insights into the problem.

We claim that our classifier worked as a preliminary automated detection system for early detection of DR. We have achieved this by making a classifier to distinguish the images with DR grades 1, 2, 3, 4 vs grade 0 (binary classification). We can enhance the classifier and strengthen the claim in two ways. First, we can train a binary classifier that will distinguish the DR grades 1 and 2 vs grade 0. This will capture the features for early detection in the image more prominently. Here, early detection is the key to prevent vision loss. Second, we can train a multi-class classifier to predict the DR grade. This will make the system usable in conditions other than early detection. We conducted experiments for (a) binary classification of DR grade 1 and 2 vs grade 0, and (b) multi-class classification, and presents our preliminary results on the Messidor-2 data set. For binary classification, we obtained an AUC value of 0.91, and for multi-class classification, an AUC value of 0.95 has been obtained. The experimental details are presented in the Appendix.

Lastly, we think that the generation of heat maps can bring out the most relevant portions of the image that cater to the classification decision. This will help novice users and patients to better understand the problems in images. This will also help in the training of medical practitioners.

Declarations of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2020.100022.

References

[1] Boyd, Kierstan. "What Is Diabetic Retinopathy?" *American Academy of Ophthalmology*. 2020. Web. 10 June 2020. <http://www.aao.org/eye-health/diseases/what-is-diabetic-retinopathy>.

[2] Anjana RM, et al. Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research–India Diabetes (ICMR–INDIAB) study. Diabetologia 2011;54(12):3022–7.

[3] Raman Rajiv, et al. Diabetic retinopathy: an epidemic at home and around the world. Indian J Ophthalmol 2016;64(1):69.

[4] International Diabetes Federation. IDF diabetes atlas, 9th ed. Brussels, Belgium. 2015 & 2019. Available at: http://www.diabetesatlas.org.

[5] Gadkari Salil S, Maskati Quresh B, Kumar Nayak Barun. Prevalence of diabetic retinopathy in India: the all India ophthalmological society diabetic retinopathy eye screening study 2014. Indian J Ophthalmol 2016;64(1):38.

[6] Raman Rajiv, et al. Prevalence of diabetic retinopathy in India: sankara Nethralaya diabetic retinopathy epidemiology and molecular genetics study report 2. Ophthalmology 2009;116(2):311–8.

[7] Ani. Prevalence of diabetic retinopathy in India is 16.9%: survey. Business Standard, Business-Standard; 10 Oct. 2019. www.business-standard.com/article/news-ani/prevalence-of-diabetic-retinopathy-in-india-is-16-9-survey-119101000997_1.html.

[8] Wilkinson CP, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110(9):1677–82.

[9] Chakrabarti Rahul, Harper C Alex, Jill Elizabeth Keeffe. Diabetic retinopathy management guidelines. Expet Rev Ophthalmol 2012;7(5):417–39.

[10] Ardila Diego, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–61.

[11] Abràmoff Michael David, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci 2016;57(13):5200–6.

[12] Gargeya Rishab, Leng Theodore. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124(7):962–9.

[13] Gulshan Varun, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 2016;316(22):2402–10.

[14] Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E. Imagenet classification with deep convolutional neural networks. Communications of the ACM 2017;60(6):84–90. https://doi.org/10.1145/3065386.

[15] Nair Vinod, Hinton Geoffrey E. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10); 2010.

[16] Srivastava Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929–58.

[17] IIoffe Sergey, Szegedy Christian. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. arXiv preprint arXiv:1502.03167.

[18] Ting Daniel Shu Wei, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Jama 2017;318(22):2211–23.

[19] Krause Jonathan, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125(8):1264–72.

[20] Gupta Garima, Ram Keerthi, Kulasekaran S, Joshi Niranjan, Sivaprakasam Mohanasankar, Gandhi Rashmin. Detection of retinal hemorrhages in the presence of blood vessels. In: Chen X, Garvin MK, Liu JJ, editors. Proceedings of the ophthalmic medical image analysis first international workshop, OMIA 2014, held in conjunction with MICCAI 2014, boston, Massachusetts, September 14; 2014. p. 105–12. https://doi.org/10.17077/omia.1015.

[21] "Diabetic Retinopathy Detection." *Kaggle*. July 2015. Web. 09 May 2020. <http://www.kaggle.com/c/diabetic-retinopathy-detection>.

[22] Diabetic retinopathy screening. EyePACS; 8 May 2019. www.eyepacs.com/.

[23] Rakhlin Alexander. Diabetic Retinopathy detection through integration of Deep Learning classification framework. bioRxiv 2018:225508.

[24] Decencière Etienne, et al. Feedback on a publicly distributed image database: the Messidor database. Image Anal Stereol 2014;33(3):231–4.

[25] Abràmoff Michael D, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol 2013;131(3):351–7.

[26] Voets Mike, Møllersen Kajsa, Bongo Lars Ailo. Reproduction study using public data of: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PloS One 2019;14(6).

[27] Van Rijsbergen, Joost Cornelis. Information retrieval. 1979.

[28] Hanley James A, McNeil Barbara J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.

[29] Szegedy Christian, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

[30] Szegedy Christian, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence; 2017.

[31] He Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

[32] Joshi Soumya, et al. Issues in training a convolutional neural Network model for image classification. In: International conference on advances in computing and data sciences. Singapore: Springer; 2019.

[33] Team, Web. "Scientific Computing Facility." *RRCAT - Raja Ramanna Centre for Advanced Technology, Indore*. 2020. Web. 10 June 2020. <http://www.rrcat.gov.in/technology/accel/indus/lil/comp/scientific.html>.

[34] Seoud Lama, et al. Red lesion detection using dynamic shape features for diabetic retinopathy screening. IEEE Trans Med Imag 2015;35(4):1116–26.

[35] Khanamiri Nazari, Hossein, et al. Smartphone fundus photography. JoVE : JoVE 6 Jul. 2017;125. https://doi.org/10.3791/55958. 55958.

[36] Koziarski Michal, Cyganek Boguslaw. Impact of low resolution on image recognition with deep neural networks: an experimental study. Int J Appl Math Comput Sci 2018;28(4).