Published in final edited form as:
Intell Based Med. 2021 ; 5: . doi:10.1016/j.ibmed.2021.100045.

Automated diabetic retinopathy screening for primary care settings using deep learning

Alauddin Bhuiyan[a,b,*], **Arun Govindaiah**[a], **Avnish Deobhakta**[b], **Mohd Hossain**[c], **Richard Rosen**[b], **Theodore Smith**[b]

[a]iHealthScreen Inc, NY, USA

[b]New York Eye and Ear Infirmary of Mount Sinai, Icahn School of Medicine at Mount Sinai, NY, USA

[c]RiteCare Medical Office PC, Hollis, NY, USA

Abstract

Diabetic Retinopathy (DR) is one of the leading causes of blindness in the United States and other high-income countries. Early detection is key to prevention, which could be achieved effectively with a fully automated screening tool performing well on clinically relevant measures in primary care settings. We have built an artificial intelligence-based tool on a cloud-based platform for large-scale screening of DR as referable or non-referable. In this paper, we aim to validate this tool built using deep learning based techniques. The cloud-based screening model was developed and tested using deep learning techniques with 88702 images from the Kaggle dataset and externally validated using 1748 high-resolution images of the retina (or fundus images) from the Messidor-2 dataset. For validation in the primary care settings, 264 images were taken prospectively from two diabetes clinics in Queens, New York. The images were uploaded to the cloud-based software for testing the automated system as compared to expert ophthalmologists' evaluations of referable DR. Measures used were area under the curve (AUC), sensitivity, and specificity of the screening model with respect to professional graders. The screening system achieved a high sensitivity of 99.21% and a specificity of 97.59% on the Kaggle test dataset with an AUC of 0.9992. The system was also externally validated in Messidor-2, where it achieved a sensitivity of 97.63% and a specificity of 99.49% (AUC, 0.9985). On primary care data, the sensitivity was 92.3% overall (12/13 referable images are correctly identified), and overall specificity was 94.8% (233/251 non-referable images). The proposed DR screening tool achieves state-of-the-art performance

[*]Corresponding author. iHealthScreen Inc, NY, USA. bhuiyan@ihealthscreen.org (A. Bhuiyan).

Author contributions

A. Bhuiyan contributed to the design, evaluation, interpretation of the data, and writing of the manuscript. A. Deobhakta contributed to the writing, proofreading, and critical review of the manuscript. A. Govindaiah contributed to the design, evaluation, interpretation of the data, writing of the manuscript. T. Smith contributed to the writing, proofreading, and critical review of the manuscript. R. Rosen contributed to the writing, proofreading, and critical review of the manuscript.

The software tool 'iPredict-DR'

An online version of the prediction system is available at https://www.ipredict.health (Please contact the corresponding author for username and password setup).

Declaration of competing interest

Dr. Bhuiyan reports a patent on an Image-based Screening System for Early Diagnosis of Diabetic Retinopathy related to this paper. The USPTO patent application number is 62721153. Mr. Govindaiah and Dr. Bhuiyan are paid employees of iHealthscreen Inc. which developed the product for commercial use. No other disclosures were reported.

among the publicly available datasets: Kaggle and Messidor-2 to the best of our knowledge. The performance on various clinically relevant measures demonstrates that the tool is suitable for screening and early diagnosis of DR in primary care settings.

Keywords

Diabetic retinopathy; Deep learning; Fundus imaging

1. Introduction

Diabetic retinopathy is one of the leading causes of blindness (Fig. 1) in high-income countries. In the US, the number of patients suffering from DR is expected to reach 6 million by 2020 and 11.3 million by 2030 [1]. The total annual economic burden of eye diseases in the US is about $139B [1]. Worldwide, this cost could be 3 times or higher (comparable cost analysis can be found in Refs. [2–4]). Early detection of the disease is key to its effective treatment and subsequent reduction of associated economic burdens. Diabetic retinopathy is a diabetes complication that affects the eyes. It is caused by damage to the blood vessels of the light-sensitive tissue at the back of the eye (retina). Fig. 2 shows the fundus (retinal) images affected by different stages of DR. Heat maps have been generated using the Layer-wise Relevance Propagation method [5]. The grading is based on the following symptoms. A few microaneurysms without any other abnormalities indicates mild DR. Cotton-wool spots and hemorrhages indicate moderate DR. An eye with four quadrants with intraretinal hemorrhaging, two with venous beading or one with IRMAs indicates severe DR. The presence of neovascularization of the disc or elsewhere, or vitreous hemorrhage are indicative of proliferative DR. It is a method that identifies important pixels by running a backward pass in a neural network. In the backward pass, neurons that contribute the most to the higher-layer receive the most relevance from it. The final heat maps are obtained after averaging the individual maps from the five models in the deep learning (DL) ensemble. Fig. 2 also shows heatmaps associated with the images.

Several studies and publications have proposed DR screening methods and tools. These studies have found that existing DR screening techniques are of varying accuracy and performance [6,7].

Deep learning is a popular tool that has been recently used for DR screening. Deep learning [8] is a class of machine learning techniques that allows systems to learn features directly from images without having to specify any rules or conditions about predictive parameters if there are many labeled images as input. Deep learning has also been applied in medical applications for detecting various diseases such as macular degeneration [9] and melanoma [10], among others. Our study focused solely on DR screening models.

At the very basic level in deep learning, there are input images, feature detectors, and feature maps. The detectors are then applied to images block by block to generate feature maps through a process called convolving. The same is repeated with feature detectors and going “deeper”. High-level features such as shapes are learned in the first few layers and more abstract features (such as optic discs or the drusen) are learned at deeper levels. Subsequent improvements in deep learning architectures reducing the number of features, chances of overfitting, the complexity of the model, etc., resulted in highly efficient neural network architectures that can be exploited to be used in medical applications with high reliability and accuracy.

Gulshan et al. [11] published a paper proposing an algorithm for DR detection using deep learning that achieved a sensitivity of 97.5% and a specificity of 93.4% and concluded that further research was needed to determine the feasibility of applying that algorithm in a clinical setting. Abramoff et al. [12] proposed a similar algorithm to detect referable DR with an 87% sensitivity and a 90% specificity and showed the advantages of deep learning over other techniques. Ting et al. [13] proposed and validated a deep learning system (DLS) that used the results from AI to compare with those of professional human graders in detecting DR. The DLS, built using retinal images from multiethnic populations with diabetes, showed a sensitivity of 90.5% and a specificity of 91.6% for detecting referable DR. Gargeya et al. [14] proposed a similar deep learning-based model that has an AUC of 0.94, a sensitivity 93% and a specificity of 87% on a publicly available dataset.

Cloud-based imaging and telemedicine platforms have helped increase the rate of diabetic retinal exams, as seen in the case of Gateway Medical Associates whose retinal exam compliance rate rose from 37% to 87% in just one year after adopting a telemedicine solution [15]. The same study reports an additional 14% of DR patients (who would be undiagnosed) would have benefitted from telemedicine. More studies [16,17] have concluded that telemedicine-based screening can identify up to 25% more DR cases in the diabetic population. Studies also showed that telemedicine could save healthcare costs significantly [18]. Considering the advantages of this technique, we have proposed a DR screening tool that takes advantage of the secure HIPAA compliant telemedicine platform and permits the patient to be screened for DR in primary care settings with subsequent referral to larger centers with retina specialists should it be indicated.

In this paper, we have demonstrated the effectiveness of an automated telemedicine-based DR screening tool that performs the screening with high accuracy compared to a retinal specialist. The tool showed high concordance with the evaluation of the ophthalmologist. Several novel features are included in this screening tool:

- We have used the probability values of each class in every network as features for input to the next classifier based on Logistic Model Tree, unlike other ensemble approaches that classify based on average or maximum of probabilities.
- In order to increase robustness and avoid problems with scale invariance, we incorporated different input image sizes and architecture combinations for every different network in the ensemble.
- Unlike traditional transfer learning techniques, where some of the nodes are frozen, all weights in the networks were allowed to continue to update, which allowed us to train the network much faster.

The architectures were pre-trained on ImageNet [19] database, a popular image classification dataset often used as the gold standard for image classification problems. Initially, the tool was developed on the Kaggle DR (KG-set) dataset [20]. These results were then further validated with the external dataset, the Messidor-2 (MD-set) [21,22], and a pilot trial was conducted in a primary care clinic.

2. Methods

In this study, we used various deep learning and traditional machine learning techniques to build an accurate and deployable DR screening system, explained in the following paragraphs starting with data sources and continuing with explanations of individual architectures in the ensemble method, and finally describing the overall system deployed on a telemedicine platform.

2.1. Data sources

KG-set was used for building the training model. MD-set was used for further validation of the model.

The KG-set contained 88702 high-resolution fundus images of people with varying stages of DR disease taken from a wide variety of cameras and imaging conditions, furnished by EyePACS, the organization that both built the dataset and made it available. The Kaggle Diabetic Retinopathy Detection competition, the source of the dataset, was funded by the California Health Care Foundation. A clinician was recruited by EyePACS to grade each image on a scale of 0–4.0 refers to No DR, 1 Mild, 2 Moderate, 3 Severe, and 4 Proliferative DR.

The MD-set contained 1748 high-resolution fundus images from the Messidor research program funded by the French Ministry of Research and Defense within a 2004 TECHNO-VISION program. The retinal images were captured without pharmacological dilation, using a Topcon TRC NW6 non-mydriatic fundus camera with a 45-degree field of view. The Messidor-2 dataset is a collection of Diabetic Retinopathy (DR) examinations, each consisting of two macula-centered eye fundus images (one per eye).

Part of the dataset (Messidor-Original) was kindly provided by the Messidor program partners. The remainder (Messidor-Extension) consists of examinations from Brest University Hospital. Some fundus images were obtained in pairs. Some others were single. In the original Messidor dataset, there were 1058 images from 529 examinations. In Messidor-Extension, diabetic patients were recruited in the Ophthalmology department of Brest University Hospital (France) between October 16, 2009, and September 6, 2010. Only macula-centered images were included in the dataset. Messidor-Extension contains 345 examinations (690 images, in JPG format). Overall, Messidor-2 has 874 examinations (1748 images).

Additionally, retinal images taken at the two primary care physician clinics (or PCP-clinics) were captured (241 images in one clinic and 23 images in the second clinic) using a non-mydriatic (no dilation needed) DRS automatic retinal camera from Centervue Inc.

with a 45-degree field of view. Only macula-centered images were included in the dataset. There are 13 referable images and 251 non-referable images, which were graded by an ophthalmologist. The grader and the algorithm were masked to each other. The grader had no information on what the algorithm automated grades were and vice versa.

Table 1 shows the number of images in the referable and non-referable DR categories. Further descriptions of KG-set and MD-set can be found in, [23].

The tool was built by using approximately 70% of the KG-set (i.e., training set), and the rest of the dataset was then used for testing (~12.5%) and validation (~17.5%), during the training phase.

2.2. Preprocessing

The retinal images from different sources and different imaging conditions have different sizes and qualities. To maintain uniformity, the images were rescaled so that circles have the same radius (500 pixels), after which the local mean color was subtracted. They were then mapped to 128 intensity levels. This preprocessing technique was found to be effective, through trial and error optimizing for 'loss' during training, when dealing with images from various sources taken under different lighting and environmental conditions. Preprocessing was done using the same python frameworks that the entire model is developed on, without using additional software or hardware. Fig. 3 shows an example of an image that is in its original RGB [24] form and the same image when preprocessed. For the experiment, we used both the original RGB images (i.e., color) and the preprocessed images.

2.2.1. Sample sizes—

The training and test data were taken from publicly available datasets. The overall incidence of the diseases in the US is 3.4%, and the incidence in our primary care dataset is 4.9%. The power calculations showed that the sample size would be 1281 with beta 0.2, alpha 0.05, and power 0.8.

2.3. Algorithms

A model for retinal fundus image classification must be robust in terms of both image variations and dataset variations. As an example, the features of a fundus image can be a small microaneurysm or a large soft exudate. Thus, the model should be capable of learning features on a wide scale in terms of size and location. Given the above, the selection of the image preprocessing techniques and neural network was made carefully.

Multiple different neural networks were used to learn features differently as one network may miss a feature that can be picked by the other network for the same image. In general, combining the results from different models to produce a final output is more effective in obtaining better performance than merely considering each of the constituent network architectures independently [25]. Five instances of networks were selected from three architectures (explained later) based on rigorous trial and error while optimizing for loss. A variety of combinations of network sizes, elements, input image sizes, and architectures were experimented with before deciding the final ensemble. While there is no definite underlying theoretical background for the chosen architectures, loss function was used to determine the best five architectures for the problem at hand. To increase the robustness, different input sizes for the networks were chosen. Also, two types of images were fed into the models. As referenced earlier, one type was a set consisting of regular RGB images, and the other consisted of preprocessed images.

A 5-point scale [26] (No DR, Mild, Moderate, Severe, and Proliferative DR) is usually used for grading DR based on the presence and extent of microaneurysms, exudates, hemorrhages, and other abnormalities in the retina. By definition, No DR and Mild are considered non-referable, and the other categories are considered referable. Table 2 has a per-class distribution of images in the development dataset as per the 5-point scale, and the number of images in training, testing, and validation during model development. It is hypothesized that due to using five classes in model building, each class representing a point on the scale can ultimately result in models that can better predict the image being in one of the referable and non-referable cases.

In our study, we make use of architectures proposed by Chollet et al. [27] (known as Xception), Szegedy et al. [28] (known as Inception-V3), and another by Szegedy et al. [29] (known as Inception-Resnet-V2). Fig. 4 shows an overview of the framework for the screening system.

In Fig. 4, blocks of different colors are used to identify different stages in the overall architecture. For robustness, each of the five networks had a different combination of the type of input images (preprocessed images, RGB images), the type of architecture, and the input image size. The first value in the block refers to the name architecture (e.g., Xception), and the second value refers to the input image size (e.g., 699 × 699). The system consists of 5 neural networks made up of one of the three said architectures. The loss function used in each of the five networks was categorical cross-entropy. The learning rate was set at 0.0001. Images were randomly rotated, sheared, and zoomed for data augmentation in each of the five networks. The five models were trained individually for 500 epochs with an early stopping policy set to stop training when no improvement in validation loss is seen for 50 consecutive epochs. Artificial intelligence modules are built on machine learning libraries and technologies such as TensorFlow, Keras, and scikit libraries. The entire code is built using python language and platform. All the software used is open-source. The models were trained on NVIDIA Titan V and Tesla P100 GPUs for about seven days. Conventionally ensembling multiple models is done by averaging or taking the maximum or taking the most repeated value as the final output. In our study, we found through trial and error that we can enhance the overall model through further training by introducing a machine learning algorithm that combines the output probability values from the deep learning models and producing the final output. This approach consistently performed better in all metrics (specificity, sensitivity, accuracy) by about 1%–3% over other approaches such as averaging or taking maximum. Each of the five networks is a five-class classifier that gives probability scores for the five classes. Totally, 25 probability values were available from five architectures. These 25 values were concatenated to form a vector of length 25, which is then used as parameters (input variables) for a logistic model tree algorithm whose target variable is DR referable/non-referable.

2.3.1. Neural network architectures and logistic model trees definitions

2.3.1.1. Xception.: Xception is a convolutional neural network architecture based entirely on depth-wise separable convolution layers. Mapping of cross-channel correlations and spatial correlations in the feature maps of convolutional neural networks can be entirely separated, the authors of the proposed model found. The model is named Xception ("Extreme Inception") because it is a stronger version of the hypothesis underlying the Inception architecture. The Xception architecture has 36 convolutional layers forming the feature extraction base of the network. It is used extensively for image classification. The base formed with convolutional layers is followed by a logistic regression layer. Stochastic gradient descent [30] is used as the optimization algorithm in this neural network architecture.

2.3.1.2. Inception-V3.: Inception v3 is an architecture that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. It is based on the original paper: "Rethinking the Inception Architecture for Computer Vision" by Szegedy et al.

The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers. Batch normalization is used extensively throughout the model and applied to activation inputs. In total, the architecture has a depth of 59 layers with over 23 million parameters. The model is pre-trained on the ImageNet database. The stochastic gradient descent [30] is used as the optimization algorithm in this neural network architecture.

2.3.1.3. Inception-Resnet-V2.: Inception-ResNet-v2 is a convolutional neural network that combines the principles of the original inception architecture and Resnet architecture. It is pre-trained on more than a million images from the ImageNet database. As a result, the network has learned rich feature representations for a wide range of images. The network has a minimum image input size of 299-by-299. In our study, we have used an input image size of 799 × 799. The network is 164 layers deep. The stochastic gradient descent [30] is used as the optimization algorithm in this neural network architecture.

2.3.1.4. Logistic model trees.: A logistic model tree (LMT) [28] is a classification model with a supervised training algorithm that combines logistic regression and decision tree learning. The LMT is based on the idea of a model tree that is a decision tree that has linear regression models at its leaves to provide a piecewise linear regression model (where ordinary decision trees with constants at their leaves would produce a piecewise-constant model). The Logit Boost algorithm is used to produce a model at every node in the tree, and the node is then split. Each LogitBoost invocation is started from its results in the parent node. Finally, the tree is pruned. We used the numerical optimization algorithm that approaches maximum likelihood iteratively.

The models were trained to classify the images into one of the five classes. The output of a model was an array of 5 numbers. Each number in the array was a probability value of the image being in that class (with the class being represented by the array index of the value). The arrays from all five models are then concatenated to form an array of length 25 (5 probability values * 5 models) called the feature array. As a result, one feature array was created per labeled image. This array is then used as an input to a classifier with a binary output value – referable or non-referable DR. The classifier is based on Logistic Model Tree [31], a classification model with an associated supervised training algorithm that combines logistic regression [32] and decision tree learning [33].

The idea of ensemble methods is that they perform much better than their constituent methods. The methods can be combined in many ways, like averaging the outputs, taking their median values, the most recurrent output, and many more. We take a step further to introduce a machine-learning algorithm to combine the models' outputs. To compare the effect of ensembling the models and to set a baseline to assess our system, we analyze one single architecture – inception-V3 and measure its performance against the final system using the MD-set.

2.4. Telemedicine platform

The telemedicine platform (Fig. 5) integrates the server-side programs (the image analysis and deep-learning modules for DR screening) and local remote computer/mobile devices (for collecting patient data and images). The remote devices will upload images and data to the server to analyze and screen DR automatically. The telemedicine platform has been developed for web and Android platforms, the details can be obtained from Refs. [33–35]. The automatic analysis will be performed on the server, and a report will be sent to the patient/remote devices with an individual's DR stage, risk, and further recommendations to visit a nearby ophthalmologist as needed. The entire process from data entry to image analysis report is determined to take only a few minutes, depending on the experience in handling the equipment, which saves time for the doctor and the patient.

Following login, the care provider at a remote location captures the retinal images from the patient and uploads the image(s) and clinical data into the webserver. The image is first analyzed for its quality by a proprietary AI algorithm developed by iHealthscreen Inc. The algorithm was built from a different set of fundus images than those used in this study. If the image is not of the desired quality or it is ungradable, the application throws an alert to retake the image. The client-side app will call the clinical decision support system to access the data, perform automated screening, and decide on a referral to an ophthalmologist if necessary. The automated evaluation of DR status and subsequent report generation is accomplished in under a minute and reported to the PCP clinic.

We have performed a cluster-bootstrap, biased-corrected, asymptotic 2-sided 95% CIs adjusted for clustering by patients were calculated and presented for proportions (sensitivity, specificity) and AUC, respectively. All hypotheses tested were 2-sided, and a P value of less than 0.05 was considered statistically significant. No adjustment for multiple comparisons was made because the study was restricted to a small number of planned comparisons. In addition to this, we have considered the entire dataset of Kaggle with 88702 images and Messidore 2 with 1748 images. The power calculation following the formula $\mathbf{n} = (\mathbf{Z}^2 *\mathbf{P} *\mathbf{Q})/\mathbf{e}^2$, shows that we need 233 subjects for validation on data taken prospectively [36]. The parameters: Where, **n** = sample size, **P** = proportion of actual referable cases with population, **Q** = proportion of actual non-referable cases with population, **e** = precision; **P** = = 0.19, **Q** = = 0.81, **e** = 0.05 (for 95% confidence interval, 5% plus or minus precision), **Z =**

1.96 based on the Z score with taking **α = 0.05, n = 237**. Thus, our primary care dataset was adequate with the numbers for external validation in a primary care dataset.

2.4.1. Measures—

The model was evaluated based on its performance of detecting referable DR in the KG-set (test data). It was further validated on the external public dataset, MD-set. The metrics calculated were the accuracy, the sensitivity, the specificity, and the AUC for detecting referable DR. The large size of the KG-set test dataset afforded us the ability to test two operating points: one for high sensitivity and another for high specificity. Table 3 shows detailed results with these metrics.

2.4.2. Ethics statement—

Informed consent was obtained from all participants in the primary care data (PCP). Mount Sinai institutional review board (IRB) approved our project. Images used in the figures are from the subjects in the publicly available datasets obtained upon request from the data providers for use in this research.

3. Results

As seen in Table 3, the screening system outperforms all existing screening systems and methods to the best of our knowledge on these public datasets. The AUC of 0.9992 (0.9981–1.0) on the MD-set is state-of-the-art (refer to Fig. 6). In detecting referable DR, the system shows an accuracy of 99.08% (98.52%–99.48%), sensitivity of 97.63% (95.55%–98.91%), and specificity of 99.49% (98.95%–99.79%) for the Messidor-2 dataset. High sensitivity and high specificity operating points resulted in the same values for the metrics for the Messidor-2 test set. In KG-set test data, the accuracy was 97.94% (97.61%–98.16%) and 98.21% (97.95%–98.45%) for high sensitivity and high specificity settings, respectively. The sensitivity in the high sensitivity setting was 99.21% (98.74%–99.55%), and the specificity in the high specificity setting was 99.22% (99.01%–99.39%).

We compared our model with a similar AI algorithm developed by iDX [37], which is recently FDA approved. We used the MD-set to compare these results with previously published results from iDX. On the three measures (Sensitivity, Specificity, and AUC) considered for comparison, our algorithm outperforms that of iDX. The proposed model achieves a specificity of 99% (Refer to Table 4), while iDX scores 87%. The AUC and the sensitivity are also higher compared to iDX's model.

In the primary care settings, the system achieved a sensitivity of 92.3% overall (12 out of 13 referable images are correctly identified) and an overall specificity of 94.8% (233 out of 251 non-referable images).

In the KG-set, the large dataset and the availability allow us to compare controls (normal) vs. mild DR differentiation in the model. We have 8130 controls (normal) and 720 mild cases of retinopathy. Our system correctly classifies 98.01% (7969 out of 8130) of the controls as such and 91.9% of mild as such (662 out of 720).

The baseline model (inception-v3 based) achieves a sensitivity of 0.934 and a specificity of 0.951 in the Messidor-2 (MD-set). The area under the curve (AUC) for this baseline model is 0.96. See Table 5 for more details on the comparison between the baseline model and the ensemble model.

The final ensemble method raises the sensitivity by about 6% points and the specificity by 2.5% points compared to the baseline single architecture model. The AUC is also better, with 0.99 against 0.96. The differences in the performances demonstrate the advantages of ensemble methods over single models.

4. Discussion and conclusion

Ophthalmologists and Optometrists, who screen for Diabetic Retinopathy (DR), are often limited geographically. Visits to eye specialists are also time-consuming, and many patients miss their appointments. Hence, an automated screening tool within the primary care settings would be ideal for mitigating these issues and providing better care for DR. We have built and tested a telemedicine-ready and AI-based, fully automated DR screening tool. The model was built in the KG dataset. The external validation was done using the MD dataset. The external validation demonstrated the consistently high accuracy of the proposed DR screening tool, comparable to that of human graders. We have also demonstrated the advantages of our ensemble method over a single neural network architecture. This AI system should be tested prospectively in primary care settings, with moderate cost and non-mydriatic fundus cameras, on the compatible telemedicine platform that we have constructed, with performance evaluated on the diagnosis of referable and non-referable DR.

We note that the test performance is relatively better on standardized test data such as Kaggle and Messidor-2 data compared to the data we gathered from PCP clinics. We believe the discrepancy may be a result of several factors. The variation in the camera models and the environment (e.g., lighting setup) are big factors. Another factor could be the variation because of human graders' input. The system is built on the Kaggle dataset, and that may result in performance skewed towards the Kaggle graders. This is a factor because there is a considerable difference in the way one grader might grade an image to another grader, as shown in the work by Krause et al. [38]. In that paper, it was shown that compared to the adjudicated grading (as opposed to individual grading or majority decision), for moderate or worse DR, the majority decision of ophthalmologists had a sensitivity of 0.838 and specificity of 0.981. Given this discrepancy, it is safe to the performance of our system is on par with human graders.

Our proposed system uses trial-and-error methods at different stages of system building. However, there is no standardized approach that has been proposed that we used in this study. Training a neural network takes an enormous amount of computing resources, and therefore, a systematic and exhaustive trial-and-error approach is impractical in this scenario. We trained multiple neural networks (about 40) that varied in structural differences such as input image size, optimization techniques, loss functions, etc. Five networks with the lowest losses were chosen for the next step. However, we do not propose a standardized approach based on our system for automation in the future, until such a time where computing resources are not major bottlenecks.

Although the performance of our proposed DR screening system is shown to be very high in terms of accuracy, we believe more extensive validations are needed, which include data taken prospectively with various conditions for imaging/camera and patient diversity. This will test the reliability of the system performance in a real-world application. The system is designed to look for diabetic retinopathy signs only, which will not indicate a referral for any other abnormality in the eye. This is in contrast with the human screening of the retina, where the doctor or healthcare worker might detect other non-DR signs that may require referral to an ophthalmologist. Therefore, care must be taken to ensure the patient is informed that the system only detects diabetic retinopathy signs.

The system has shown to be effective in screening for diabetic retinopathy, which is manifested in abnormalities that can be picked up by deep learning. There are several other retinal diseases such as macular degeneration, hypertensive retinopathy, etc., whose abnormalities in the eye can be picked up and graded by automated systems in the same way. As part of the future work, we will explore an "all-in-one" system for retinal abnormalities which will greatly improve referral systems and aid in mass screenings of several potential diseases from a single screening.

If the promising performance of the tool is confirmed in the context of real-world image acquisition, then deployment in primary care settings for early diagnosis of DR is warranted. The physical system and the telemedicine software have been tested for usability, convenience, and security. The software application is HIPAA compliant and is built with a design policy of minimum interaction with the interface. By using such a secure, fast, reliable, and low-cost system, millions of eyes can potentially be saved from preventable vision loss with significant healthcare savings.

Funding

This project is funded by NIH SBIR project number R44EY031202.

Data sharing

Messidor-2 (kindly provided by the Messidor program partners) and Kaggle datasets are publicly available and can be obtained upon request from here:

1. https://www.kaggle.com/c/diabetic-retinopathy-detection/overview
2. http://www.adcis.net/en/third-party/messidor2/

References

[1]. Eye Disease Statistics. National eye insititute 2014 [available from. https://nei.nih.gov/sites/default/files/nei-pdfs/NEI_Eye_Disease_Statistics_Factsheet_2014_V10.pdf.

[2]. Hex N, Bartlett C, Wright D, Taylor M, Varley D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. Diabet Med 2012;29(7):855–62. [PubMed: 22537247]

[3]. Heintz E, Wiŕehn A-B, Peebo BB, Rosenqvist U, Levin L-Å. Prevalence and healthcare costs of diabetic retinopathy: a population-based register study in Sweden. Diabetologia 2010;53(10):2147–54. [PubMed: 20596693]

[4]. Happich M, Reitberger U, Breitscheidel L, Ulbig M, Watkins J. The economic burden of diabetic retinopathy in Germany in 2002. Graefes Arch Clin Exp Ophthalmol 2008;246(1):151–9. [PubMed: 17406883]

[5]. Binder A, Bach S, Montavon G, Müller K-R, Samek W. Layer-wise relevance propagation for deep neural network architectures. Information science and applications (ICISA) 2016. Springer; 2016. p. 913–22.

[6]. Xiao D, Bhuiyan A, Frost S, Vignarajan J, Tay-Kearney M-L, Kanagasingam Y. Major automatic diabetic retinopathy screening systems and related core algorithms: a review. Mach Vis Appl 2019;30(3):423–46.

[7]. Saha SK, Xiao D, Bhuiyan A, Wong TY, Kanagasingam Y. Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: a review. Biomed Signal Process Control 2019;47: 288–302.

[8]. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436. [PubMed: 26017442]

[9]. Govindaiah A, Hussain A, Smith R, Bhuiyan A. Deep convolutional neural network-based screening and assessment of age-related macular degeneration from fundus images. In: The proceedings of IEEE international symposium on biomedical imaging; 2017. p. 1525–8. 10.1109/ISBI.2018.8363863.

[10]. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542 (7639):115. [PubMed: 28117445]

[11]. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 2016;316(22):2402–10. [PubMed: 27898976]

[12]. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digit Med 2018;1(1):39. [PubMed: 31304320]

[13]. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, 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. [PubMed: 29234807]

[14]. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124(7):962–9. [PubMed: 28359545]

[15]. Cloud-based imaging, telemedicine platform increases rate of diabetic retinal exams: healio Primary care optometry news. Available from: https://www.healio.com/optometry/retina-vitreous/news/online/%7Bd48b2e4c-19f2-4d5f-96d0-3c80f9432514%7D/cloud-based-imaging-telemedicine-platform-increases-rate-of-diabetic-retinal-exams; 2018.

[16]. Gao X, Park CH, Dedrick K, Borkar DS, Obeid A, Reber S, et al. Use of telehealth screening to detect diabetic retinopathy and other ocular findings in primary care settings. Telemedicine and e-Health; 2018.

[17]. Naik S, Wykoff CC, Ou WC, Stevenson J, Gupta S, Shah AR. Identification of factors to increase efficacy of telemedicine screening for diabetic retinopathy in endocrinology practices using the Intelligent Retinal Imaging System (IRIS) platform. Diabetes Res Clin Pract 2018;140:265–70. [PubMed: 29649538]

[18]. 5 ways telemedicine is reducing the cost of healthcare. 2012.

[19]. Deng J, Dong W, Socher R, Li L-J, Li K, et al. ImageNet: a large-scale hierarchical image database. CVPR09; 2009.

[20]. Kaggle diabetic retinopathy detection competition [Available from, https://www.kaggle.com/c/diabetic-retinopathy-detection.

[21]. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: the Messidor database. Image Anal Stereol 2014;33(3):231–4.

[22]. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA ophthalmology 2013;131(3):351–7. [PubMed: 23494039]

[23]. Kaggle diabetic retinopathy detection competition [Available from, https://www.kaggle.com/c/diabetic-retinopathy-detection.

[24]. Gonzalez RC, Woods RE. In: Digital image processing. Pearson Prentice Hall; 2008.

[25]. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscipl Rev: Data Min Knowl Discov 2018;8(4):e1249.

[26]. Wilkinson C, Ferris III FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110(9):1677–82. [PubMed: 13129861]

[27]. Chollet F Xception: deep learning with depthwise separable convolutions. arXiv: 1610.02357v3 [cs.CV], https://arxivorg/pdf/161002357pdf; 2017.

[28]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, editors. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

[29]. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Szegedy C, Ioffe S, Vanhoucke V, Alemi AA, editors. The proceedings of thirty-first AAAI conference on artificial intelligence; 2017. file:///C:/Users/abhui/Downloads/14806-66795-1-PBpdf - last accessed on Aug 16, 2019.

[30]. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent 2012;14(8). Cited on.

[31]. Landwehr N, Hall M, Frank E. Logistic model trees. Machine Learning 2005;95 (1–2). 161–205, https://wwwcswaikatoacnz/~ml/publications/2005/LMTpdf Logistic Model tree. https://enwikipediaorg/wiki/Logistic_model_tree.

[32]. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.

[33]. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660–74.

[34]. Laravel. Laravel - a PHP framework (last accessed on September 18, 2019).2018 (Sept 2, 2018, https://laravelcom/docs/.

[35]. API R. What is REST API', REST API Tutorial (last accessed on September 18, 2019), https://restfulapinet.

[36]. Cochran WG. Sampling technique. second ed. New York: John Wiley and Sons Inc.; 1963.

[37]. Abràmoff MD, Lou Yiyue, Erginay A, Clarida W, Amelon R, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investig Ophthalmol Vis Sci 2016:5200–6. [PubMed: 27701631]

[38]. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125(8):1264–72. [PubMed: 29548646]

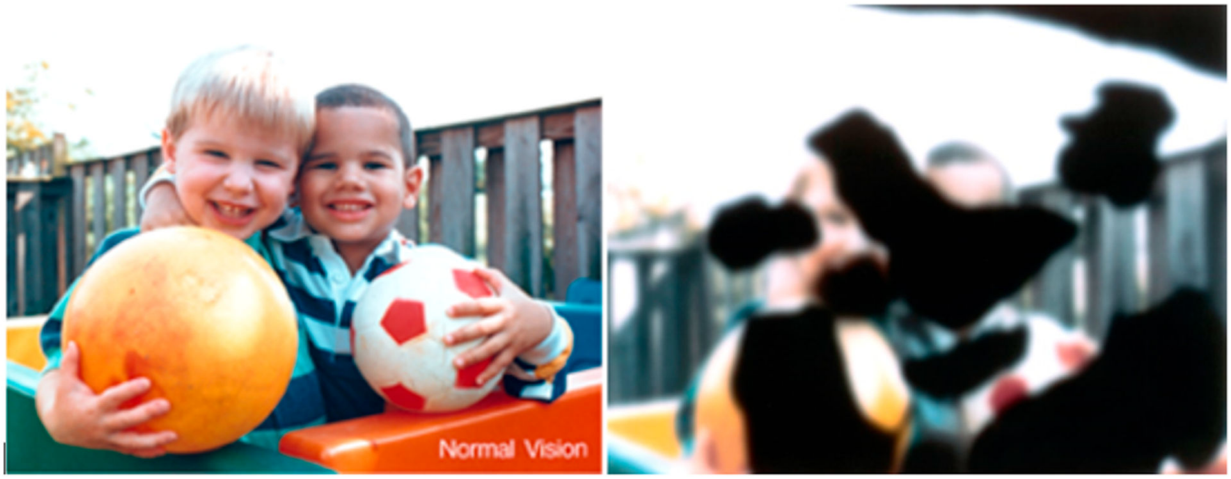

Fig. 1.

Normal vision (left) and what DR patient sees (right) (Credits: ÓNEI, Source: https://nei.nih.gov/health/examples).

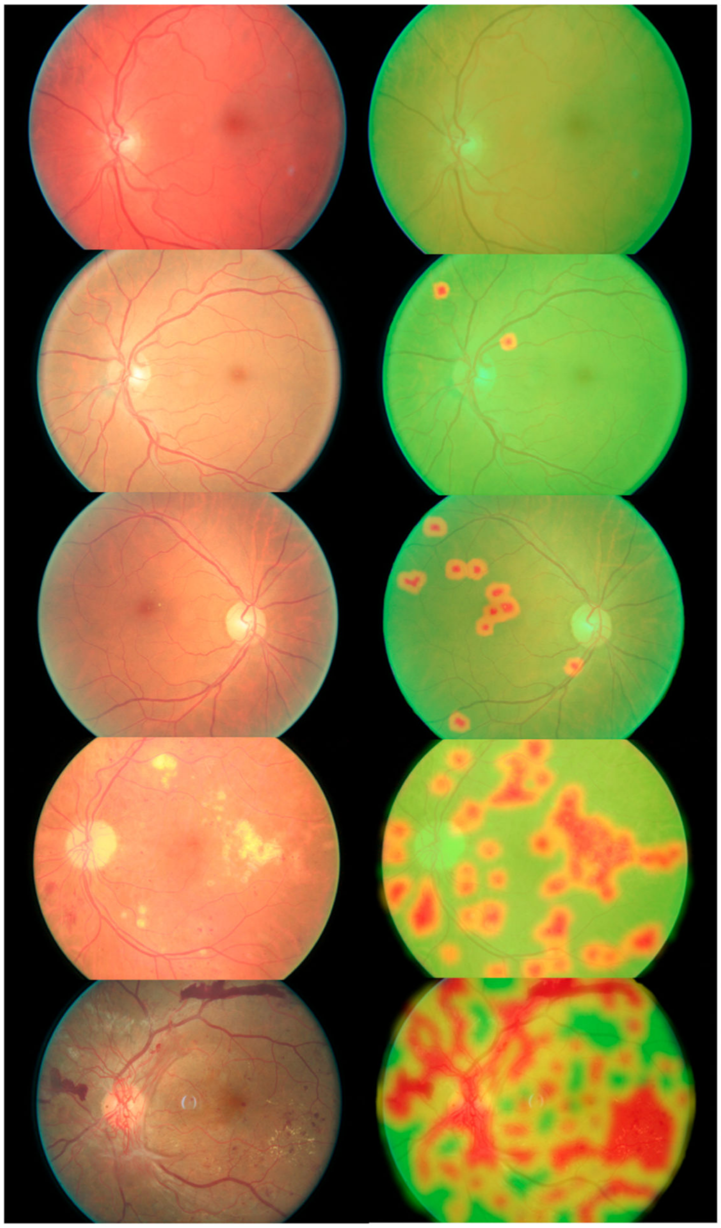

Fig. 2.

Five stages of DR progression with examples from the Kaggle Dataset. (A) Normal, (B) Mild, (C) Moderate, (D) Severe, and (E) Proliferative DR. Severe and Proliferative DR includes the heat map of the affected areas (hemorrhages, exudates, microaneurysms, neovascularization, etc.).

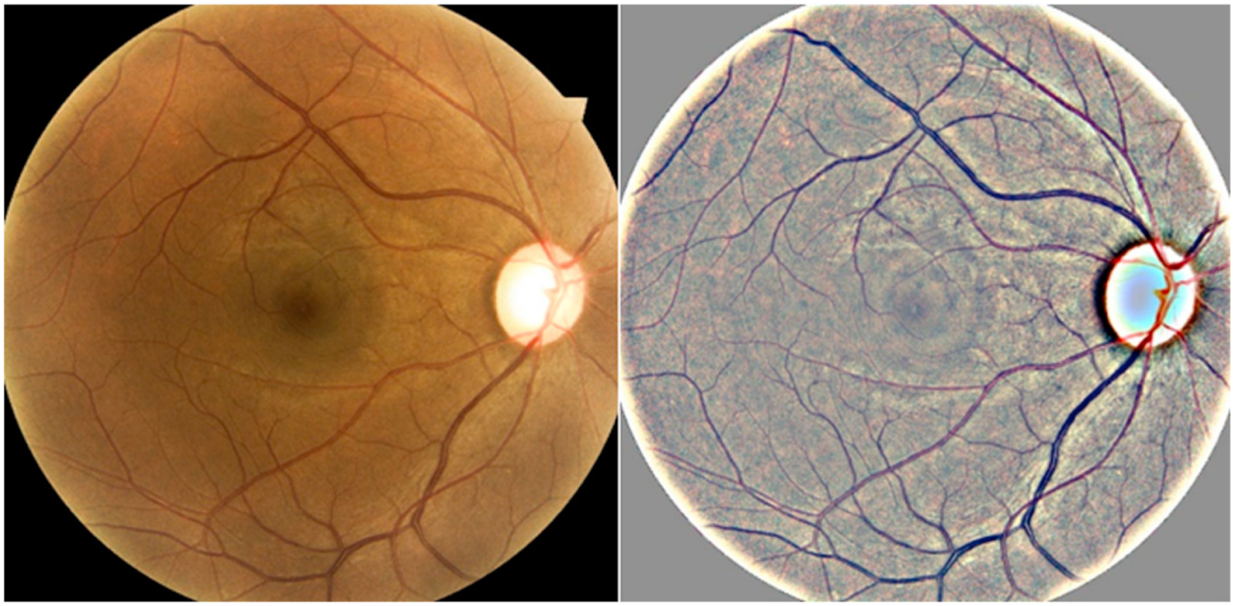

Fig. 3.

Original RGB fundus image (left), processed image (right).

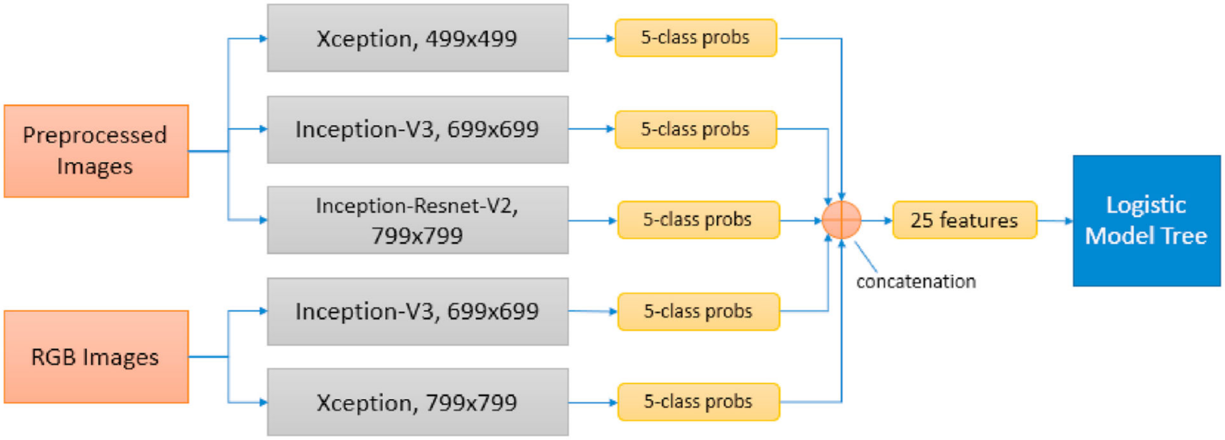

Fig. 4.

The framework of deep learning-based DR screening system. The figure shows the two types of input images (original RGB and preprocessed), which are used to build five deep learning models differing in the type of architecture and input image size. The networks are trained and optimized and the resulting probabilities are then concatenated to form a feature vector, which is used as input to the Logistic model tree (trained separately, not in an end-to-end fashion) that forms the final classifier.

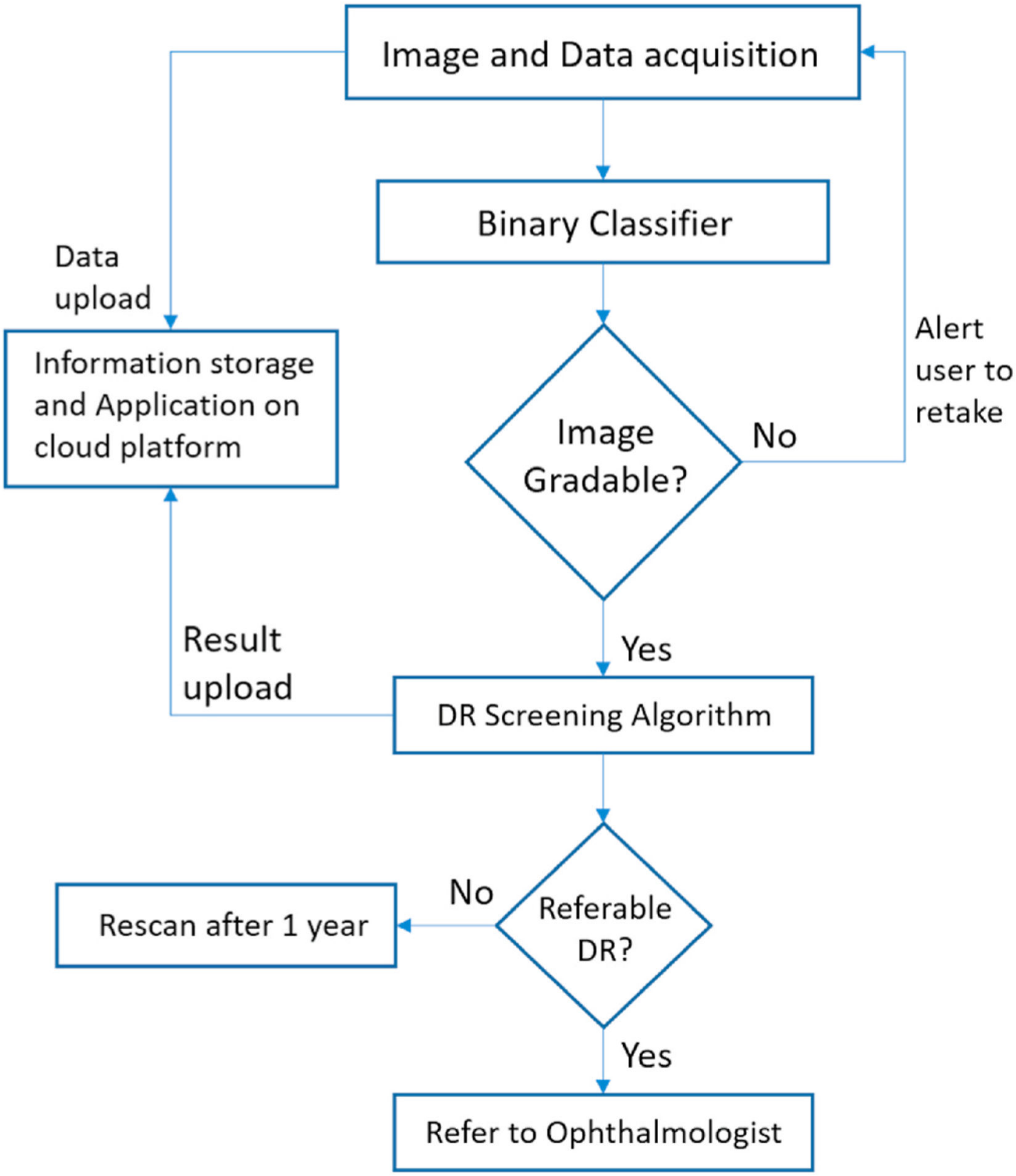

Fig. 5.

Proposed DR screening through telemedicine. The data flow beginning with image and data acquisition by the healthcare worker that is analyzed by a binary classifier for image quality. The image is further analyzed by a screening algorithm for referable DR cases. The data and results are stored on secure cloud platforms as part of the telemedicine platform.

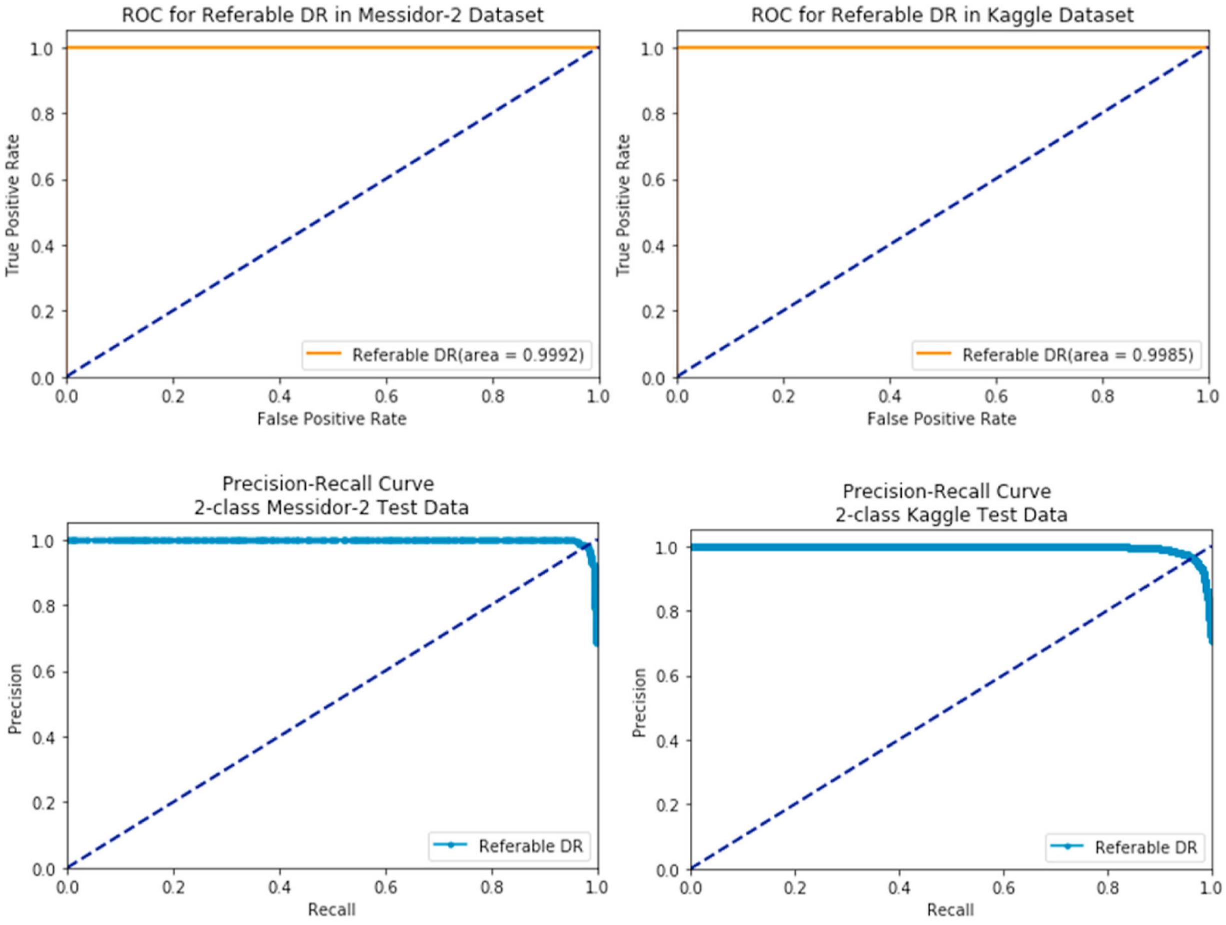

Fig. 6.

ROC Curves and Precision-Recall Curves for detecting referable DR in the two datasets (MD-set and KG-set). The corresponding AUCs are shown in the bottom right portion of the curve. The curves (orange lines) are not obvious because of very high AUCs, where they graze the y-axis making the orange line blend with the axis. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1

Number of images in referable and non-referable DR categories in MD-set, and KG-set.

Referable/non-referable	MD-set	PCP-clinic	KG-set test & validation	KG-set Training
Non-referable (%)	1368 (78.26%)	251 (95.07%)	8850 (81.15%)	50198 (80.45%)
Referable (%)	380 (21.74%)	13 (4.93%)	2056 (18.85%)	12198 (19.55%)
Total (%, 100 by definition)	1748 (100%)	264 (100%)	10906 (100%)	62396 (100%)

Table 2

Number of Images in KG-set in each DR category (No DR, Mild, Moderate, Severe, and Proliferative DR) as graded in the Kaggle competition dataset and subsequent division into training, validation, and test datasets.

Classes	Overall	Training	Validation	Test
0 - No DR	65343 (73.66%)	45813 (73.42%)	11400 (74.02%)	8130 (74.55%)
1 - Mild DR	6205 (7%)	4385 (7.02%)	1100 (7.14%)	720 (6.6%)
2 - Moderate DR	13153 (14.8%)	9374 (15.02%)	2200 (14.29%)	1579 (14.47%)
3 - Severe DR	2087 (2.35%)	1500 (2.4%)	350 (2.27%)	237 (2.17%)
4 – Proliferative DR	1914 (2.15%)	1324 (2.12%)	350 (2.27%)	240 (2.2%)
Total	88702 (100%)	62396 (100%)	15400 (100%)	10906 (100%)

Table 3

Performance of the proposed referable DR system on two public datasets (MD-set and KG-set). Metrics – Sensitivity, Specificity, Accuracy, and AUC – are used to measure the performance on the said datasets. Two different operating points –high sensitivity and high specificity – were evaluated for KG-set (which has a high number of images), and the results were reported.

Metric	**Messidor-2 (MD-set)**	**PCP-clinic set**	**KG-set test**	
Operating Point	N/A	N/A	High Sensitivity	High Specificity
Sensitivity (95% CI)	0.9763 (0.9555–0.9891)	0.9231 (0.6397–0.9981)	0.9921 (0.9874–0.9955)	0.9387 (0.9275–0.9487)
Specificity (95% Q)	0.9949 (0.9895–0.9979)	0.9283 (0.8890–0.9569)	0.9759 (0.9725–0.9790)	0.9922 (0.9901–0.9939)
Accuracy (95% CI)	0.9908 (0.9852–0.9948)	0.9280 (0.8899–0.9561)	0.9794 (0.9761–0.9816)	0.9821 (0.9795–0.9845)
AUC	0.9992 (0.9981–1.0)	0.8950 (0.879–0.926)	0.9985 (0.9979–0.9991)	
Positive Predictive Value	0.9474 (0.92–0.965)	0.905 (0.88–0.9136)	0.9195 (0.9055–0.9345)	

Table 4

Comparison of Specificity, Sensitivity, and AUC of our proposed Referable DR model with those of iDX DR model on the Messidor-2 dataset.

Metric	iHealthScreen	iDX
Specificity	0.994	0.870
Sensitivity	0.976	0.968
AUC	0.99	0.98

Table 5

The comparison of the baseline model (Inception-V3), which is a single model, and the final ensemble model compared side by side on the Messidor-2 dataset.

Metric	Baseline (single architecture - Inception-V3)	Ensemble Method
Sensitivity	0.934	0.994
Specificity	0.951	0.976
AUC	0.96	0.99