Artificial Intelligence In Medicine 96 (2019) 93–106

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

A data-driven approach to referable diabetic retinopathy detection☆

Ramon Pires[a,*], Sandra Avila[a], Jacques Wainer[a], Eduardo Valle[b], Michael D. Abramoff[c,d,e], Anderson Rocha[a]

[a] Institute of Computing, University of Campinas (Unicamp), Campinas 13083-852, Brazil
[b] School of Electrical and Computing Engineering, University of Campinas (Unicamp), Campinas 13083-852, Brazil
[c] Stephen R. Wynn Institute for Vision Research, the Department of Electrical and Computer Engineering, the Department of Biomedical Engineering, the University of Iowa, Iowa City, IA 52242, USA
[d] VA Medical Center, Iowa City, IA 52246, USA
[e] IDx LLC, Iowa City, IA, USA

ARTICLE INFO

Keywords:
Diabetic retinopathy
Referral
Screening
Multi-resolution training
Robust feature-extraction augmentation
Integrated patient-basis analysis

ABSTRACT

Prior art on automated screening of diabetic retinopathy and direct referral decision shows promising performance; yet most methods build upon complex hand-crafted features whose performance often fails to generalize.
Objective: We investigate data-driven approaches that extract powerful abstract representations directly from retinal images to provide a reliable referable diabetic retinopathy detector.
Methods: We gradually build the solution based on convolutional neural networks, adding data augmentation, multi-resolution training, robust feature-extraction augmentation, and a patient-basis analysis, testing the effectiveness of each improvement.
Results: The proposed method achieved an area under the ROC curve of 98.2% (95% CI: 97.4–98.9%) under a strict cross-dataset protocol designed to test the ability to generalize — training on the Kaggle competition dataset and testing using the Messidor-2 dataset. With a 5 × 2-fold cross-validation protocol, similar results are achieved for Messidor-2 and DR2 datasets, reducing the classification error by over 44% when compared to most published studies in existing literature.
Conclusion: Additional boost strategies can improve performance substantially, but it is important to evaluate whether the additional (computation- and implementation-) complexity of each improvement is worth its benefits. We also corroborate that novel families of data-driven methods are the state of the art for diabetic retinopathy screening. Significance: By learning powerful discriminative patterns directly from available training retinal images, it is possible to perform referral diagnostics without detecting individual lesions.

1. Introduction

Every eleventh person in the world suffers from diabetes mellitus, a disorder of sugar metabolism, whose prevalence is expected to reach every tenth person by 2040 [1]. Diabetes sufferers are 25 times more likely to suffer from sight loss resulting from diabetic retinopathy, a major long-term microvascular complication, and the leading cause of blindness in high-income countries [1]. In the U.S. alone, 7.7 million people aged 40+ years have diabetic retinopathy [2], and the prevalence is even larger in developing countries with a shortage of ophthalmologists and optometrists [3].

Detecting early signals of diabetic retinopathy is critical for limiting its progression. The World Health Organization and professional organizations such as the American Academy of Ophthalmology recommend eye examinations at least once a year for diabetic patients. However, poor or isolated communities often cannot afford such frequent consultations with ophthalmologists, frustrating early diagnosis and treatment. Over half of U.S. counties has limited availability of ophthalmologists and optometrists per capita [3], and around 10% of the people with diabetes lives in counties without eye care professionals [4]. Those numbers are much worse in developing countries.

Automated screening addresses such lack of access, helping to decide who should be referred to the ophthalmologist for further examination [5–16]. However, most existing methods focus on identifying

☆ For reproducibility purposes, all of our code is freely available at http://repo.recod.ic.unicamp.br/piresramon/data-driven-referable-diabetic-retinopathy-detection/tree/master.

* Corresponding author.
E-mail addresses: pires.ramon@ic.unicamp.br (R. Pires), sandra@ic.unicamp.br (S. Avila), wainer@ic.unicamp.br (J. Wainer), dovalle@dca.fee.unicamp.br (E. Valle), michael-abramoff@uiowa.edu (M.D. Abramoff), anderson.rocha@ic.unicamp.br (A. Rocha).

https://doi.org/10.1016/j.artmed.2019.03.009
Received 17 October 2018; Received in revised form 23 March 2019; Accepted 26 March 2019

lesions in the retina, isolated or in combination, generally exploiting very specific visual structures — a handcrafted feature engineering involving extensive image pre-processing and ad-hoc decisions. Those features search for specific characteristics of diabetic retinopathy lesions such as brightness, size, and shape. Recent solutions reduce such complexity, with unified approaches that capture collective discriminative patterns of lesions [15,17].

Those hierarchical approaches ("lesion-first, referral-later") assume that preliminary lesion detection is necessary and sufficient for the later decision on referability. Those assumptions are questionable: although conventional machine learning techniques often demand those staged decisions, current art based on deep learning — so called image-based decisions — infer directly from the pixels [18]. Moreover, a preliminary stage of lesion detection discards information that may prove useful for the later stage of referability. Although such image-based systems raise legitimate concerns about spurious associations with confounders in training data, and about their sensitivity to adversarial images [19–21,9], they have become the *de facto* gold standard for visual recognition tasks.

Although the prior research on referral assessment relies on the natural route of identifying DR lesions and gathering individual responses to evaluate referability, with the recent demand for accountable solutions — not only robust and accurate, but also self-explainable — the research has moved toward an opposed direction ("referral first, lesion later") [22,23].

In this work, we showcase data-driven methods for referable diabetic retinopathy detection. Those techniques allow to triage patients who require referral to the ophthalmologist from those who can wait until the next screening. The solution we present is based upon deep Convolutional Neural Networks, one of the most successful image classification models [18]. We propose an original architecture whose design was inspired in two recent networks proposed by groups that ranked second in two outstanding competitions (2015 Kaggle DR detection and ImageNet ILSVRC-2014). Note this was just an inspiration and that we propose a new architecture and solution for the problem. Our similar, but streamlined architecture yields considerable gains in *efficiency*, with comparable and sometimes superior *effectiveness*, specially when we consider the difficult setup of training with images collected with different acquisition conditions than those of testing during system deployment (the cross-dataset validation). Efficiency gain is a factor we concern about since we strive to validate and aggregate the screening proposal into portable and accessible retinal imaging devices. In this regard, our work herein is a first attempt toward developing an efficient solution amenable to be deployed in a portable low-cost device. Our current partnership with a portable retinograph manufacturer, Phelcom Technologies,[1] enables this possibility and emphasizes the need for cross-dataset validation setups. Fig. 1 shows the first MVP (minimum viable product) produced by Phelcom (left) and one resulting retinal image (right) acquired with it.

We organized this paper into five more sections. In Section 2, we overview the prior art. In Section 3, we describe the methodology used to evaluate each refinement of the solution. In Section 4 we present the adopted experimental protocol while, in the Section 5, we present experimental results. Finally, in Section 6, we conclude the paper and discuss possible future work.

2. State of the art

Diabetic-retinopathy referral is currently addressed in different ways [5–9,12,13,15,16]. Solutions vary from custom-tailored handcrafted lesion detectors (sometimes encoding expert-domain knowledge) to mid-level representations (combining a series of low-level descriptors) to data-driven lesion detectors (underpinned by neural network advances). In this section, we overview existing approaches to detect referable diabetic retinopathy.

We start with methods based upon the detection of individual lesions. Then, we discuss some approaches that exploit data mining to diagnose retinal images. Thus, we present some works which integrates telemedicine systems with automated screening systems. In sequence, we describe some recent methods that evaluate referral using deep learning. Afterwards, we show referral solutions that require the lesion detection but alleviate the complexity of traditional approaches employing unified detection techniques. Finally, we present a handcrafted referral solution that does not depend on lesion detection and does not lose critical information.

Many techniques for referral assessment rely upon examining the positioning of lesions on retinal landmarks, are complex and, frequently, tailored to each lesion. Some works rely on the number of microaneurysms [24], although their presence in one quadrant of the retina characterizes only mild nonproliferative diabetic retinopathy, without the need of referral. Naqvi et al. [25] proposed a referral system for hard exudates in diabetic retinopathy using a robust mid-level representation of bag-of-visual-words (BoVW), which is reliable and easily adaptable to other lesions. However, they fall short of the aim of offering reliable referability assessment, ending up providing mostly a screening for exudates. Moreover, in the task of assessing referral, they did not follow any standard grading consensus regarding diabetic retinopathy level classification — necessary for a consistent improvement of communication among specialists [26].

Taking a different path, recent works used data mining to assess referral due to pathologies [5,6,27]. Decencière et al. [5] combined visual information (retinal images) and contextual data from the individuals (e.g., patient age, weight or diabetes history) to detect retinal pathologies and to point out whether or not patients need referral to a specialist. The authors considered image quality metrics, diabetic retinopathy-related lesion information (exudates, microaneurysms, and hemorrhages), demographic and diabetes-related information.

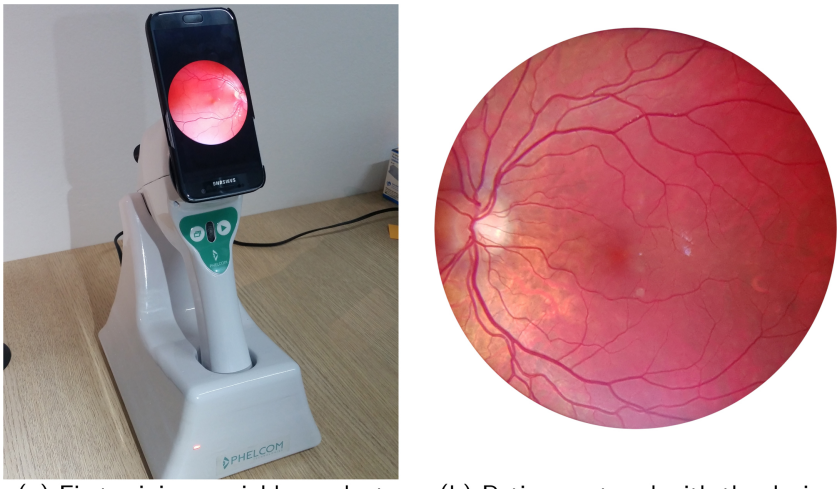
Similarly, Quellec et al. [6] used multiple retina images and contextual information about the patient to detect abnormal retinas. Instead of detecting just one or more lesions related to a particular eye disease, they identify patients who need referral to an eye care provider, regardless of pathology. The method starts by mixing a set of retinal images and building a mosaic for each one. Then, they characterize the images using a BoVW model, extracting multi-granular histograms into a cascade of regions. After characterizing the images, their method extracts diagnosis rules relying on visual word histograms and contextual information previously collected.

Saleh et al. [27] also explored contextual information (not images) to detect referral patients. The authors proposed ensemble classifiers, based on medical attributes available in the health care record, to diagnose diabetic retinopathy. The medical attributes involve numerical (age, evolution time of diabetes, body mass index, etc.) and categorical (sex, hypertension and medical treatment) information; and the methods are based on two kinds of ensemble classifiers learned from data: fuzzy random forest and dominance-based rough set balanced rule ensemble.

In addition to those solutions, telemedicine systems also improved health care productivity and addressed the lack of access to diabetic retinopathy screening. Besides increasing local access, telemedicine programs provide risk stratification of diabetic patients so that those who require treatment can be scheduled more efficiently [28]. Automated screening systems integrated with telemedicine frameworks make diabetic retinopathy screening more accessible, efficient, and cost-effective [29] and a few such systems were proposed [7–9].

Bhaskaranand et al. [7] integrated a diabetic retinopathy screening solution that assesses severity and referral into a telemedicine system. The screening tool, called EyeArt, analyzes image quality (patients with collected ungradable images are also referred), and enhances images so as to normalize and improve the appearance of the existing lesions.

[1] https://www.phelcom.com.br/english.



(a) First minimum viable product (b) Retina captured with the device

Fig. 1. The first MVP (left) and a retinal image captured with the Phelcom's Retinal Camera. The images captured with the portable devices are currently under process of annotation regarding need of referral, and we intend to use them for validation in future work.

With a set of filterbank descriptors, the method identifies and describes regions with anatomical or pathological structures associated with specific lesions (microaneurysms, exudates, and hemorrhages). A supervised learning ensemble is employed at the very end for referral decision.

As exposed above, bags of visual words (BoVW) became a fundamental approach for image representation and is widely exploited for retinal image analysis. Rocha et al. [30] adopted a class-aware fashion (one codebook per class) especially suitable for diabetic retinopathy-related lesions. The approach was suitable to detect lesions [31–33,25], and assess referability [14–16,6].

Detecting lesions is usually the first stage of traditional referral assessment, and it tends to be specific for each lesion. The BoVW methodology allowed general frameworks adaptable for large classes of lesions [31,32,14,15,33]. Pires et al. [14,15] applied that unified methodology to detect lesions with models based on BoVW mid-level features and SVM classifiers, and gathered the individual scores (per lesion) to referral decision making. Ultimately, the authors extended the initial method by applying better mid-level image representations boosting the detectors capabilities [15].

Abràmoff et al. [8] investigated the potential of the Iowa Detection Program (IDP) to detect referable diabetic retinopathy. The IDP is a framework for quality analysis and diabetic retinopathy lesion detection. In a per-patient setup (two images per patient), the IDP combines analysis of individual lesions, structures, and quality in a simple likelihood that encapsulates the patient's diagnostic about referable diabetic retinopathy.

In a follow-up work, Abràmoff et al. [9] integrated the IDP with a set of convolutional neural networks (CNNs) specialized on detecting hemorrhages, exudates, and neovascularization as well as normal retinal anatomy and image quality. Employing CNNs ranging from the well-known Alexnet [34] to VGG [35], the hybrid system significantly outperforms existing solutions at the task of diabetic retinopathy screening.

As we can see, most of previous work rely on a two-tier method for referral decisions (lesion detection followed by referral decision-making). As critical information might be lost upon adopting such route, Pires et al. [16] proposed a direct referral assessment that replaces lesion classifiers and, instead, relies on the retinal images with all cogent information directly for referral analysis. This method has outperformed several existing lesion-based methodologies.

Unified approaches that capture discriminative patterns of distinct lesions alleviate the complexity of exploiting specific tailored visual characteristics. However, as any handcrafted technique, those approaches are subject to lose critical information that could be evinced in data-driven approaches to provide effective decisions. In this vein, showing the potential of a data-driven system over handcrafted counterparts, Gargeya and Leng [12] customized deep convolutional neural networks for extracting features to classify images into no DR vs. any stage of DR, and no DR vs. mild DR. Those features are combined with retinal image metadata for classification. Quellec et al. [13] trained CNNs to detect referable DR, using a heatmap optimization procedure. To create heatmaps, the authors proposed a training method which involves a third pass on the CNN to propagate second-order derivatives forward. Those CNNs trained for image-level classification are also used to detect lesions related to DR (hard exudates, soft exudates, small red dots, hemorrhages).

Gulshan et al. [11] ensembled 10 CNNs with the Inception-v3 architecture [36], trained with the ImageNet dataset, to make multiple binary decisions such as (1) moderate or worse DR, (2) severe or worse DR, (3) referable diabetic macular edema, or (4) fully gradable. An image fits as referable if it fulfills criterion (1), criterion (3) or both. An important landmark for automated diabetic retinopathy detection was a recent competition promoted at Kaggle[2] by California Health Foundation, with images provided by EyePACS, a platform for retinal screening. The dataset, comprising more than 88,000 samples, was the largest publicly available dataset of retinal images at the time. The aim of the competition was classifying the images into five degrees of severity, ranging from 0 (no sign of diabetic retinopathy) until 5 (proliferative diabetic retinopathy). All winning teams employed variations over deep learning, working directly from the image pixels. Kaggle favor both competition and cooperation among the teams, with an open forum for discussions. Kaggle competitions invite spontaneous contributions, promoting creativity at the expensive of formality: e.g., there are minimal requirements in terms of experimental design, and none in terms of peer-review or statistical validation.

The top-ranking team in the competition was Min-Pooling.[3] They preprocessed the images to compensate different lighting conditions, by subtracting the local average color and removing boundary effects. The authors classified the images with several SparseConvNets CNNs that use 5-class softmax, and selected the three best ones. The output probabilities and additional metadata (probabilities of the other eye size and variance of the original and preprocessed data) are combined

[2] https://www.kaggle.com/c/diabetic-retinopathy-detection.

[3] https://www.kaggle.com/c/diabetic-retinopathy-detection/forums/t/15801/competition-report-min-pooling-and-thank-you.

into a single one using random forest.

The o_O team ranked second.[4] The authors do not apply image preprocessing, and use dynamic resampling to deal with class unbalancing. Two CNNs with similar architectures are trained as a regression problem with mean squared error (MSE) objective, and both applying pre-trainings with smaller architectures. After training the CNNs, one breakthrough of the proposal was the extraction of features from different versions of the data (pseudo random augmentations), that are posteriorly combined into mean and standard deviation. Those features are concatenated with features from the other eye and an additional boolean indicator for right eye, and used as input to a neural network. That process is repeated six times, using the two architectures, and using three sets of parameters to extract augmented features: the final weights, and the weights that generated best score and best kappa in the validation set. The six predictions are combined through averaging and thresholded for the DR decision-making.

Reformed Gamblers ranked third in the competition.[5] They used a variety of 9 CNNs that ranges from VGG [35], followed by models with fractional max poolings, to cyclic poolings. Most models were trained with clipped MSE error function. The authors also combined predictions from both eyes, from the 9 CNNs. Rather than using predefined thresholds, they performed a grid search of possible cutoffs and selected the combination that leads to the best score.

In this paper, we follow-up the competition and showcase data-driven methods for referral decision. We have three aims: (1) designing an effective and efficient data-driven model for the binary task of referral/non-referral decision; (2) reviewing data-driven techniques promoted at the Kaggle DR competition from a rigorous point of view; and (3) introducing and evaluating contributions of our own including an efficient and effective proposal, transfer learning — forbidden in the competition — and cross-dataset evaluations.

3. Methodology

In this section, we present our deep learning-based solution for diabetic retinopathy screening and highlight a series of approaches explored to achieve a robust and effective framework. As explained in Section 2, the Diabetic Retinopathy Detection competition at Kaggle represented a milestone for research in this area. However, here we investigate and propose binary decisions of referability, while the competition aimed the task of severity classification. While from the human point of view there's a relatively direct map between the two tasks, from a Machine Learning point of view, limiting to binary classification has theoretical and practical advantages.

Solutions presented at the Kaggle competition have to improve the target metrics in short time. Rigorous validation of factors leading to performance is less important than quickly improving metrics. Our aim here is opposite: we are less interested in shaving tenths of percents from the classification error, and more interested in evaluating the cost-benefit of each choice and novel contributions. Before presenting the proposed architecture and individual scientific contributions, we provide a brief review of concepts related to deep learning for image classification.

3.1. Convolutional neural networks

For decades, constructing a pattern recognition system with conventional approaches required a considerable domain expertise in order to convert raw data (such as size, shape, color and location of retina lesions) into a proper representation; and a posterior learning stage with algorithms capable of recognizing categorical patterns (e.g., whether or not a patient needs a follow-up).

Deep learning composes non-linear modules that, starting with raw data, transform the representation in one level into a more abstract representation [18]. With enough data, the multiple-level composition enables to learn very complex functions in an end-to-end manner: represent the data and categorize.

CNNs are deep-learning methods designed to deal with data in form of multiple arrays such as images. In CNNs, convolutional layers play the role of non-linear modules. Units within a convolutional layer are connected to local patches of its input (that might correspond to output of a previous layer) through a set of weights, and the weighted-summed local patches pass through a non-linearity. A convolutional layer is composed by a set of filters (sometimes referred to as neurons or kernels), and the local regions on which it slides is called the receptive field. As the filter is sliding or convolving around its input, it is multiplying its values with the original pixel values of the input (element-wise multiplications), and these multiplications are summed up and pass to non-linearization. After sliding the filters of a convolutional layer over all the locations, the layer produces an output called activation map or feature map, representing the activation of neurons. In general, the filters convolve by shifting one unit at a time producing a feature map of same height and width dimensions of the input. Strides are applied to reduce the dimensionality of the output, by shifting more than one unit at a time (e.g., shifting two units produce feature maps with half of the input dimension). Those convolutional layers have the role of detecting local conjunction on features from previous layers [18].

Pooling layers (and its numerous variations) have the role of merging neighbouring features into one. The pooling simply performs downsampling along the spatial dimensionality of the given input, further reducing the number of parameters within that activation. It is expected that those neighbour features are semantically similar.

The dense layers (also referred as fully-connected layers) perform the same services found in standard, shallow neural networks and attempt to produce class scores from the activations, to be used for classification. In general, convolutional, non-linearity and pooling layers are stacked, followed by some dense layers (each unit is fully-connected to the input, followed by a non-linear unit). The weights of convolutional and dense layers are traditionally learned by back-propagating gradients.

For a more detailed introduction of the composition of convolutional neural networks as well as the learning procedure by back-propagation, we recommend introductory works [18] and online courses that present the foundations of deep learning [37].

3.2. Our architecture

Aiming both at effectiveness and efficiency, we present a streamlined architecture for retinal image analysis, which bears some resemblance with two main networks in prior art (namely o_O, a key competitor in the 2015 Kaggle Diabetic Retinopathy Detection Challenge, and VGG, a key competitor for natural image classification in the ImageNet challenge) but with key insights and differences. The solution we propose, which is depicted in Fig. 2, is significantly novel both in practical (CNN architecture) and methodological (scientific procedure leading to the solution) terms. The proposed architecture has 16 weight layers, with about 10 million parameters.

In terms of arrangements of convolutional and pooling layers, our architecture resembles the VGG-16 [35] except that we reduce by half the number of units on the first four out of five convolutional blocks and use strides when appropriate, focusing on efficiency and aiming at working with higher resolution images. We use very small receptive-field (3×3) convolutions. Pooling layers separate sequences of two or three convolutional layers. The first pair of convolutional layers starts with 32 kernels/filters. When pooling layers take place reducing

[4] https://www.kaggle.com/c/diabetic-retinopathy-detection/forums/t/15617/team-o-o-solution-summary.

[5] https://www.kaggle.com/c/diabetic-retinopathy-detection/forums/t/15845/3rd-place-solution-report.

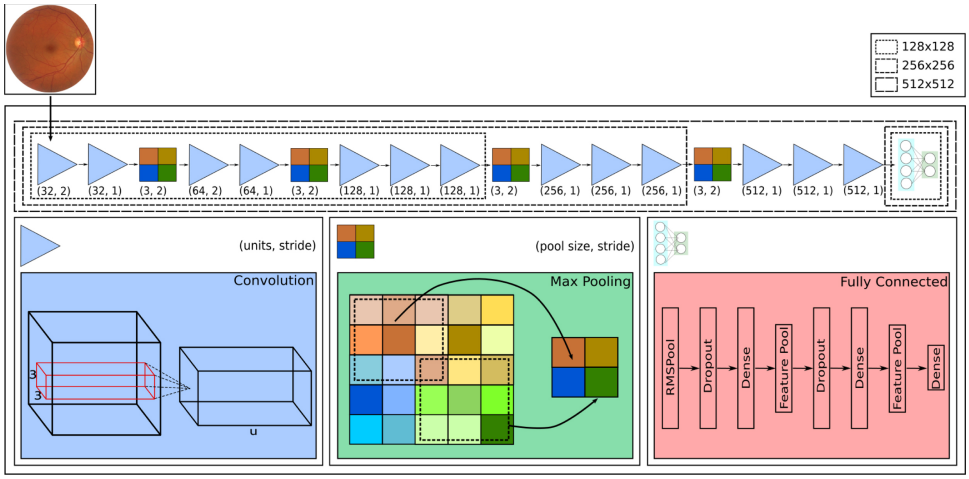


Fig. 2. The proposed solution decides to refer the patient directly from the pixels of the retinal exam, without preliminary feature extraction or lesion detection. From an initial "basic" configuration, we propose and evaluate improvements step-by-step.

drastically the feature map sizes, the convolutional layers double the number of kernels. Additionally, we stride in two the first and third convolutional layers across their respective inputs to reduce more aggressively the initial layers and work with higher dimensional images. In the fully-connected stage, similarly to the o_O team, we apply RMSPool with both pool size and stride of (3, 3), and Feature Pool with pool size of (2, 2). The units of the 1024-unit hidden dense layers employ drop-out with a fixed probability of 0.5. Our final architecture, in this case, is a hybrid one (not presented before), which brings to bear essential ideas for an efficient and effective analysis of retinal images. For instance, in the original VGG-16 design, almost 90% of the parameters reside on the fully-connected layers. In our new arrangement, convolutional layers represent 75% of the parameters while fully-connected layers represent only 25%, a remarkable feat for efficient implementation.

Our idea here is creating a network as similar as possible to VGG-16 [35] in terms of convolutional and pooling arrangements, but also similar to o_O's ones in terms of the fully-connected sequence. Comparing our convolutional/pooling arrangement with one of the o_O's architectures, we essentially moved one layer from the coarse stage to the finer one, resulting in more parameters. Above all else, we tried to preserve the structure that enabled the use of smaller networks for multi-resolution training (see Section 3.4).

After each convolutional and dense layer (except the last one), we use leaky rectifier units (leaky RELU with alpha = 0.01) that applies a small negative slope to address the shortcomings of the simple rectified linear unit ("dying" RELU), while accelerates the convergence of the gradient in comparison with conventional activation functions.

We optimize the CNN with Nesterov momentum [38] that, contrasting with standard stochastic gradient descent, is capable of accelerating convergence in regions of low-curvature. As well as classical momentum, Nesterov momentum accelerates gradient descent by accumulating a velocity vector in directions of persistent reduction in the objective [38]. While conventional momentum computes the gradient update from the current position θ_t, Nesterov momentum first computes $\theta_t + \mu v_t$ (a partial update to θ_t). This allows Nesterov changing velocity in a more responsive and stable way, especially for higher values of μ. Given a objective function f, Nesterov momentum is written as:

$$\begin{aligned} v_{t+1} &= \mu v_t - \varepsilon \nabla f(\theta_t + \mu v_t) \\ \theta_{t+1} &= \theta_t + v_{t+1} \end{aligned} \tag{1}$$

where $\varepsilon > 0$ is the learning rate, $\mu \in [0, 1]$ is the momentum coefficient, and $\nabla f(\theta_t + \mu v_t)$ is the gradient at the partially updated θ_t.

We defined a fixed schedule of 250 epochs for training, starting with a learning rate of 3×10^{-3} in the first 150 epochs, decreasing it to 3×10^{-4} in the following 70 epochs and finally to 3×10^{-5} until the end. Additionally, we applied L2 regularization (weight decay) with factor 0.0005 in all convolutional and dense layers.

We exploited different alternatives for optimization (algorithms and hyper-parameters). Comparing to o_O's proposal, we kept the same optimization schedule although with a different network architecture as that schedule also worked for our problem. Moreover, here we tackle the problem with a classification point of view rather than a regression one and apply cross-entropy as objective function, instead of MSE.

Henceforward, we review and describe, from a rigorous scientific point of view, a set of approaches proposed by the o_O team and new contributions of our own. We use such approaches to improve progressively our "basic" framework. We anticipate that we cannot analyze the following variations (or improvements) independently since it embraces coarser approaches — essential to the convergence of the CNN — that must be kept to analyze finer approaches.

3.3. Resampling and data augmentation

Convolutional Neural Networks thrive with hundreds of thousands, up to several million learning samples for training. However, contrasting to general-purpose object recognition, medical tasks count on relatively small annotated datasets. In the rare situations where we can count on relatively large medical image datasets, the data tends to be very unbalanced, with the overwhelming portion of the images corresponding to the control group (healthy patients).

Data augmentation can improve the learning process [34,39]. The augmentation can consist of image perturbations by geometric (e.g., zoom, translations, rotations, cropping) or photometric (e.g., histogram equalizations, contrast enhancements) transformations.

In the technique evaluated here, we propose a data augmentation method to simultaneously address the small sample size and the class unbalancing problems. The aim of data augmentation here is not just inflating the training set for each epoch, but also dealing with the under-sampled classes by dynamically applying random transformations to their samples, while keeping the classes balanced.

The set of operations (or perturbations) we consider comprises zoom, flipping, rotations, translations, stretching, and color augmentation. The perturbations are dynamically performed right before submitting an image to the network, bypassing the need for saving numerous versions of each image. We exploit zooms, rotations, and translations by randomly choosing a variable into a predefined interval (for instance, we apply rotations between 0 and 360 degrees, zooms between 1/1.15 and 1.15, and translations between -40 and 40).

The color augmentation, proposed by Krizhevsky et al. [34], consists on changing pixel intensities of RGB channels by adding multiples of principal components (PCA) found in the training set, with magnitudes proportional to the corresponding eigenvalues times a random variable drawn from a Gaussian with mean zero and standard deviation 0.5. The resampling ensures that all classes will be represented equally. The number of randomly perturbed versions (data augmentation) for each class depends on the balance weights, that is inversely proportional to the number of images for each class.

3.4. Multi-resolution training

Poor initialization of network weights leads to poor local minima and, consequently, to an ineffective solution. Additionally, training large CNNs from scratch requires a very large dataset. To address those shortcomings, we propose a multi-resolution training strategy that consists on training simplified variants of the CNN — requiring less training samples — and using the learned parameters as starting point for next stages. Those variants have less layers but preserve the number of units of each layer of the original network.

We train reduced versions of the entire network (fewer convolution layers) using smaller images, and preinitialize larger networks with the learned parameters. Simonyan and Zisserman [35] employed a similar approach, training the network with the shallowest configuration and after initializing the first four convolutional and the last three fully-connected layers of deeper configurations with the learned parameters.

Fig. 2 shows in dashed lines the composition of the original and the two smallest networks. Initially, using the same images resized to 128×128 pixels and the same ground-truth, we train a small variant of the network that does not have the two last pooling layers and convolution trios (sequences of three convolution layers). Note that this is a simplified version that comprises just seven convolution layers with the same number of parameters of the corresponding layers in the original network. Thereafter, using images resized to 256×256 pixels, we produce another simplified network based on the complete architecture, but without the last pooling and the last convolution trio (with ten convolution layers), and use the parameters learned with the previously trained network with 128×128-pixel images to initialize the weights before training. Subsequently, we initialize the first ten convolution layers with the learned parameters and train the architecture using 512×512 pixels images, our final image dimensions of interest.

Fig. 3 depicts the sequence of training multiple networks, each one using images of different resolutions. The multi-resolution training accelerates the optimization of deeper networks by using parameters pre-trained with smaller networks.

The convolutional and fully-connected layers of the first network configuration (input as 128×128 pixels) as well as the additional layers of the following networks are initialized with random orthogonal matrices.

3.5. Robust feature-extraction augmentation

Deep learning-based methods provide us with an end-to-end learning process: the learning models receive raw images as inputs and produce probabilities as outcomes (resulting classes), after an extensive and strongly abstract learning highway. In this case, image representation and pattern recognition are performed together, enabling us to extract features in any layer before the one responsible for the final decision, and use that information in a posterior decision process. This feature extraction procedure is highly flexible as it allows us to use different machine learning algorithms.

To exploit that flexibility (training a posterior classifier or testing the images), we extract features in a different pathway. Following the o_O team's proposal, we apply pseudo-random data augmentation and create n versions for each image, both from training and test set. Pseudo-random augmentation ensures that the same sets of perturbations are always applied to all images. The final feature vector for each image is achieved by concatenating mean and standard deviation of the n individual feature vectors from the respective n image versions:

$$\mathbf{x_i} = [\mu_i, \alpha_i], \tag{2}$$

where $\mathbf{x_i}$ represents the image i. We created 20 image versions ($n = 20$) in our experiments and extracted features in the last pooling layer of the CNN. The whole process is data-driven in the sense that all features are extracted directly from the data with no human intervention.

3.6. Per patient analysis

As our ultimate goal is checking whether or not a patient needs to see a doctor within 12 months, and not merely pinpointing the presence of lesions within his/her retinas, whenever photographs of the two retinas are available, we leverage this additional imagery to make the final referral decision.

To provide an outcome for each retina, we concatenate features of both of them and include a binary indicator variable that refers to left or right, before feeding the additional classifier for the final decision. The feature vector for a retina is created as follows:

$$\mathbf{x_{retina}} = [\mu_{\text{retina}}, \mu_{\text{retina}'}, \alpha_{\text{retina}}, \alpha_{\text{retina}'}, \delta_{\text{right}}] \tag{3}$$

where *retina′* represents the complementary retina.

In addition to combining information and diagnosing retinas individually, we go beyond and assign to the patient the response of the retina that presents the highest risk (highest probability of needing referral).

3.7. Transfer learning

Transfer learning aims to transfer knowledge learned in one or more source tasks to improve the learning process in a target task. Normally, the transfer is sought when there are not enough training samples in the target domain or when there is already a reasonable solution for a related (source) problem and it would be natural to leverage such knowledge while solving the target problem. Most of the existing works that exploit transfer of skills/knowledge implicitly assume that the source and target domains are related to each other [40].

Transfer learning normally appears in two distinct scenarios: feature extraction and fine-tuning. In the former case, we freeze the network layers and correspondent parameters and use them just to extract features to the target task. In practice, this corresponds to remove the last dense layer and treat the CNN as feature extractor that will be followed by a linear classifier (e.g. linear support vector machine) for the new dataset.

In turn, the latter case involves continuing the backpropagation, updating of the network internal parameters to better adjust them to the target domain. In this situation, it is possible and sometimes recommended to keep the earlier layers frozen, since they contain generic features that recognize edges and color blobs. Assuming the domains are similar, smaller learning rates might be used since the pre-learned weight can be distorted too quickly and too much.

In this work, the domain adaptation corresponds to transfer knowledge from a domain also related to retinal image analysis, in which the source is severity of diabetic retinopathy (a 5-class decision problem), and the target is referability of diabetic retinopathy (a 2-class decision problem). We explore transfer learning in its two scenarios — feature extraction and fine-tuning — since we intend to compare data-driven approaches with previous work that employed handcraft methods with relatively small datasets under cross-validation protocol.

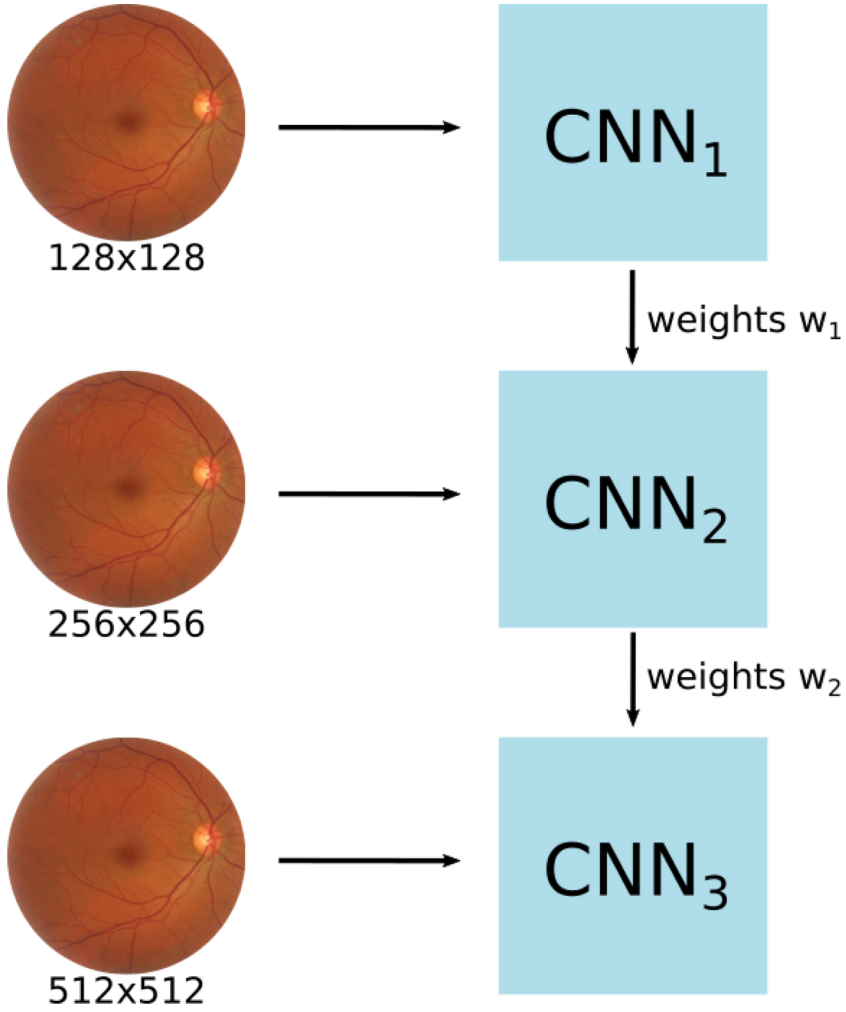


Fig. 3. The multi-resolution training consists on using weights previously learned with smallest versions of the network in order to accelerate the optimization and boost the effectiveness. The weights learned with CNN_1 are used to initialize the CNN_2 weights (the ones they have in common). The same process is applied to initialize the CNN_3 weights.

4. Experimental protocol

In this section, we describe the datasets and validation protocols adopted in this work.

4.1. Datasets

The previously-mentioned Kaggle competition provided participants with 88,702 images collected and annotated by EyePACS, from which 35,126 (see Table 1) are used for training and 53,576 for testing. In this work, we use the same training and test sets as in the challenge. The images, whose sizes range from 320 × 211 pixels to 5184 × 3456 pixels, were taken under a variety of imaging conditions. The dataset comprises images of left and right eyes, all graded by severity level. We convert the labels (originally for disease stages, source domain) to referral necessity (target domain), following the International Clinical Diabetic Retinopathy recommendations (ICDR) [26]: tagging as non-referable only those patients with no diabetic retinopathy signal or mild non-proliferative diabetic retinopathy (NPDR). Patients with moderate, severe, or proliferative DR must be referred (note that the images are not labeled for macular edema). The conversion of labels is not algorithmic, but it is done manually before training the classifiers. Nevertheless, the amount of retina images was not enough to develop a robust referral decision based on our architecture as the number of parameters is still much higher than the number of available training images. Thus, we adopted this architecture with the discussed data augmentation policy to train the CNN.

Table 1
Profile of Kaggle dataset according to original labels of diabetic retinopathy severity.

#images	Label
25,810	0 — no diabetic retinopathy
2443	1 — mild NPDR
5292	2 — moderate NPDR
873	3 — severe NPDR
708	4 — proliferative diabetic retinopathy

In addition to the Kaggle dataset, we also consider the DR2 dataset[6] from the Department of Ophthalmology, Federal University of São Paulo, that comprises 520 images captured using a TRC-NW8 (Topcon Inc., Tokyo, Japan) nonmydriatic retinal camera with a Nikon D90 camera. DR2 provides referral annotations for 435 images, manually categorized by two independent specialists whose mean intergrader κ is 0.77. Of these, 98 images were graded by at least one expert as requiring referral (56 images graded as positive by both experts), while

[6] Publicly available under accession number 10.6084 and URL https://doi.org/10.6084/m9.figshare.953671.

337 images were annotated by both experts as not requiring referral within one year.

Finally, the third adopted dataset is Messidor-2 [41,42] an extension of the Messidor dataset that is a collection of diabetic retinopathy examinations, each consisting of two macula-centered eye fundus images (one per eye). Also captured with a Topcon TRC NW6 non-mydriatic fundus camera with a 45 degree field of view, the Messidor-2 contains 874 examinations (1748 retina images). As reported in [8,9], the images from Messidor-2 were independently graded by three board certified retinal specialists from all subjects according to the ICDR severity scale and a modified definition of macular edema (ME).[7] The mean κ value among the three experts is 0.822. For details about ICDR severity scale, please consult the reference [26].

4.2. Validation protocol

We use three validation protocols: training and testing from the same dataset (without intersection), the 5 × 2-fold cross-validation and the cross-dataset validation protocols.

In the first protocol, we perform training and testing operations with different parts of the dataset in a hold-out fashion. The idea here is finding the best configuration of our method in a dataset that already provides a clear division of training and testing (Kaggle).

The 5 × 2-fold cross-validation protocol consists of repeating by five times the process of two-fold cross validation [43] in which we randomly separate the samples into two groups balanced by class, and use one of them for training and the other for testing. We perform two experiments per step, with the groups switching roles. We use this protocol to compare to previous work, mainly evaluations with the DR2 dataset.

Finally, the cross-dataset protocol is the strictest and therefore closer to real-world operational conditions, in which we train and test the classifiers on different datasets collected in very different environments with different cameras, at least one year apart and in different hospitals. This protocol plays an important role in the design, since in clinical practice, rarely the analyzed images will have the same image specification (camera, resolution, operator, FOV) as the images used for training the classification method. We use this protocol to show results when training the classifier with Kaggle data and testing on different datasets (e.g., Messidor-2).

5. Results

In this section, we present results for the data-driven approach to referable diabetic retinopathy detection. We divide the section into three parts: in part 1, we describe and then refine the approach; in part 2, we validate the solution on different datasets and evaluate how efficient and effective it is; finally, in part 3, we investigate the capability of transfer learning in the context of diabetic retinopathy screening.

5.1. Initial solution and further refinements

In order to produce the final data-driven solution, starting from scratch and refining according to the performance for referral assessment, we present the first model which we will refer to as baseline, and investigate some hypothesis aiming at progressively improving the approach before advancing to the following steps. We use the protocol of training and testing with the Kaggle dataset following the splits proposed therein.

In this first version of the solution, we still do not employ any data augmentation technique nor use the robust feature-extraction augmentation policy discussed in Section 3.5. We perform a naïve data balancing through sample removal for the most favored classes, repeating the process in each epoch. Such initial baseline method leads to an AUC of 71.6%. Starting with this baseline, we now pose a series of research questions in order to evaluate possible improvements and design decisions. For reference, questions Q1–Q4 refer to results depicted in Fig. 4.

Q1: *Is data augmentation essential to train the proposed CNN?* To investigate the first question, we applied geometric image perturbations and color augmentation, always aiming at balancing the classes, as detailed in Section 3.3. After augmenting the training set, the CNN reached an AUC of 93.1%. The data augmentation remarkably improved the initial results, showing that it is critical to choose a good policy of data augmentation/balancing.

Q2: *Is the multi-resolution training important to train with larger images?* To investigate this question, we start the data-driven process by training reduced versions of the architecture, adapted to images of lower resolutions, as explained in Section 3.4. The multi-resolution training slightly boosted the AUC to 93.9%, and showed that it was essential for the convergence of the CNN since deeper networks require larger datasets, and is a satisfactory option to provide an effective solution.

Q3: *Is the robust feature-extraction augmentation satisfactory?* Here we extracted features from augmented versions of each image and create a final feature vector by concatenating mean and standard deviation of the image versions, as exposed in Section 3.5. With the augmented features, which feed an extra neural network of two hidden layers for referral decision, we achieved an AUC of 94.6%, a satisfactory improvement.

Q4: *Is the per-patient analysis important to provide more robustness?* Since the Kaggle dataset contains images from the left and right eyes for each patient, we contrasted per-image decision (each image analyzed independently) to per-patient decision (aggregating features of both eyes). In the system we advance, the patient need of referral is the highest resulting classification probability between the two eyes (see Section 3.6). The per-patient analysis improved results considerably, leading to a 95.5% AUC (95% CI: 95.1–95.8%).

After testing each progressive enhancement, the final solution comprises a (1) CNN trained with data augmentation; (2) employing multi-resolution training with parameters from smaller networks; (3) using robust feature-extraction augmentation and training a decision classifier (Neural Network) on top of these features. In addition, we also consider (4) diagnosing patients with images from both eyes. We compare our method with the original o_O's proposal. Using an ensemble of **six** networks (two physical networks with three distinct sets of parameters) and a per-patient analysis, the o_O solution yields an AUC of 95.8% AUC (95% CI: 95.5–96.1%), while our method, which only relies on one network instead of six, yields and AUC of 95.5%.

In Section 5.2, we validate the solution with distinct datasets in a challenging cross-dataset validation, provide more comparisons regarding efficiency and effectiveness and a deeper cost-benefit analysis of our claims.

5.2. Cross-dataset validation — training on Kaggle and testing on Messidor-2 and DR2

After refining the automated solution for referable diabetic retinopathy screening from scratch, we investigate the performance over distinct datasets. In this section, we investigate the performance of the proposed method when training with the Kaggle dataset and testing it with Messidor-2 and DR2, which have very different acquisition conditions. Basically, we use the CNN trained with data augmentation and multi-resolution training (Section 5.1) to extract features for the test sets (robust feature-extraction augmentation), and test the features with the classification that provides referral decisions. We emphasize again that one of the most valuable advantages of extracting features (in this case, robust feature-extraction augmentation) is that it provides

[7] The reference standard for referable diabetic retinopathy is available for researchers at http://www.medicine.uiowa.edu/eye/abramoff/.

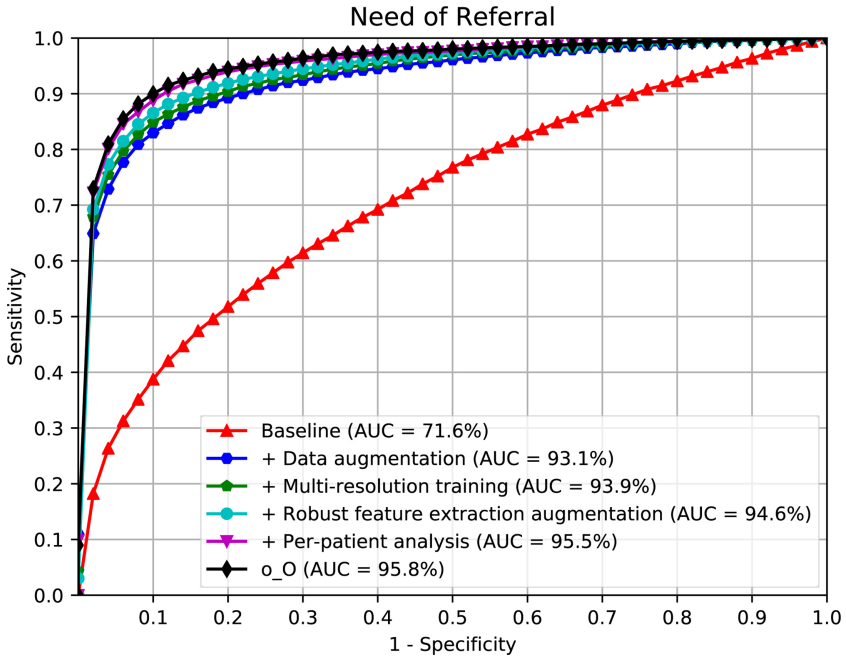


Fig. 4. ROC for referral assessment on the Kaggle dataset (with the official competition splits). The baseline CNN results are compared to the progressive proposed improvements, showing that those are advantageous for the task. We also compare with the original o_O's solution (ensemble of six classifiers).

flexibility to choose different machine learning algorithms. Henceforward, we use two algorithms: Neural Network and Random Forest. Therefore, for research question Q5 (below), we follow a challenging cross-dataset validation protocol with the best solution we found in the previous section that was trained with Kaggle data.

We highlight that the solution incorporates the proposed data augmentation for training, multi-resolution training, and robust feature-extraction augmentation steps. We also exploit the per-patient information whenever we have access to images of both eyes.

We configure the Neural Network as a shallow three-layer network: the first two with 32 units and ReLU as non-linear activation function, while the last one has 2 units. The layers are intercalated by feature pool layers, and initialized with orthogonal matrices. We trained the network for 100 epochs with Adam optimizer, with learning rate starting at 0.0005 and reducing 0.1 in the following epochs: 60, 80 and 90. For Random Forest we firstly perform an extensive grid-search to select the hyper-parameters that provide best AUC, varying the number of estimators from 50 to 300, and criterion as gini or entropy. All the experiments used entropy criterion, with 200 or 300 estimators.

Q5: *Can we diagnose retinal images collected under different acquisition conditions?* In this section, we assess the possibility of training a retinal image diagnosis system with one set of images and test it using images collected under very different acquisition conditions. Here, we use Kaggle dataset images to train the expert CNN and then use it as a feature extractor for DR2 and Messidor-2. For DR2, the per-patient analysis is not feasible as the dataset does not have two images per individual.

Fig. 5 depicts the ROC curves achieved with the two considered classifiers in the cross-dataset validation protocol with the DR2 or Messidor-2 datasets for testing. The Neural Network-based classifier yields the best result for testing with DR2 dataset, with an AUC of 96.3% (95% CI: 93.8–98.1%). With Random Forest, in turn, the AUC is 96.1% (95% CI: 93.1–98.0%). The results show that the models learned with the Kaggle dataset images (with higher variance) produced relevant results with a very different dataset (DR2).

We now turn to the case in which we extract features from the Messidor-2 dataset using network and parameters learned with the Kaggle dataset and test the method with the individual decision models (Neural Network and Random Forest) on Messidor-2 dataset. As the Messidor-2 dataset also provides pairs of images from left and right eyes for each patient, we employ the per-patient analysis herein. The Neural Network classifier reached the best result, resulting in an AUC of 98.2% (95% CI: 97.4–98.9%) in a per-patient analysis. The Random Forest algorithm achieved an AUC of 97.9% (95% CI: 97.0–98.6%). These results show the solution has a remarkable performance also with Messidor-2 dataset, even considering the challenging cross-dataset validation protocol.

These results corroborate the hypothesis that it is possible to train a robust data-driven solution to precisely pinpoint diabetic retinopathy referral needs, independently of operators and camera settings of the training set of images.

5.2.1. Comparison with previous work on Messidor-2 dataset

Just for the sake of completeness, we now compare our work (solution explained in Section 5.1) with Abràmoff et al. [8,9]. Note, however, that this comparison is not totally direct as both methods use different training sets. Considering the Messidor-2 dataset on a per-patient basis, Abràmoff et al. [8] reported an AUC of 93.7%, and after replacing most of the feature detectors with CNNs trained to detect features, further improved to 98.0% (95% CI: 96.8–99.2%) [9].

In turn, the method described here and trained with Kaggle data produces equally remarkable results for the Messidor-2 dataset with an AUC of 98.2% (95% CI: 97.4–98.9%), reinforcing the fact that detecting diabetic retinopathy lesions is not essential for a reliable and effective diabetic retinopathy screening. Note here that, for the method in this paper, the CNN was trained and optimized using only Kaggle data, and Messidor-2 data was never used for optimizing nor training the CNN, showing the robustness of the method.

5.2.2. Comparison with o_O's solution

For the sake of comparison, we adapted the o_O solution for a two-class classification problem and evaluated its performance for referable DR detection in terms of efficiency and effectiveness. To do so, we replaced the last one-neuron fully-connected layer by another with two neurons (one per class), and tackled the problem with a classification point of view rather than regression (cross-entropy as objective function, instead of MSE). We recall that the o_O solution consists on a ensemble of six different models, trained with features extracted from

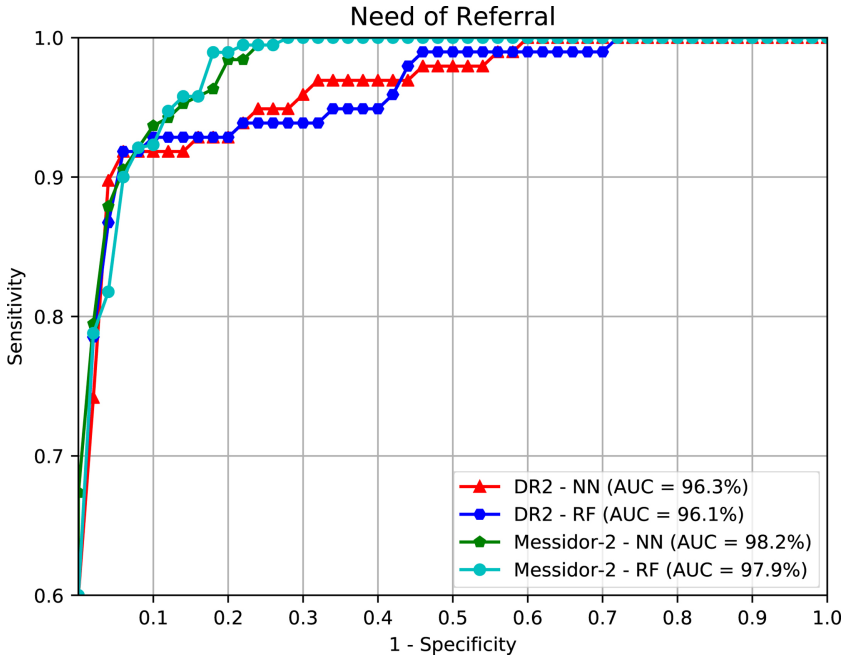


Fig. 5. ROC results for referral assessment in a cross-dataset protocol (training with Kaggle; and testing with either DR2 or Messidor-2). The cross-dataset protocol is the most strict and realistic one, as the dataset present different acquisition characteristics (operators, equipment, population of patients, etc.)

Table 2
Efficiency: time and memory comparisons.

Work	Time (s)	Memory (MB)
o_O	295	285
Ours	**60**	**56**
Improvement	4.91×	5.08×

two different CNNs.

Table 2 reports the time and memory required for inference. We performed the tests using one GeForce GTX TITAN X. We simulated a real-time diagnostic environment in which 50 patients are screened for referral after capturing their fundus images (left and right), totaling one hundred images. In terms of time consumption, we report the "real" time, that encompasses loading all required libraries, loading parameters of the CNNs, pseudo-augmenting and describing the images and, finally, the inference part with higher probability among the eyes. For the memory footprint, we consider the disk space required to keep the CNN parameters and of the two-hidden-layer neural networks in memory.

Fig. 6 depicts the results achieved with the current work and o_O proposal, both using the cross-dataset validation protocol over DR2 and Messidor-2 datasets. The DR2 results correspond to diagnosing one image at a time while Messidor-2 are for patients analysis. For DR2, the o_O's ensemble reached an AUC of 96.1% (sens. = 86.7%, spec. = 95.5%), while we achieved an AUC of 96.3% (sens. = 90.8%, spec. = 95.5%). For Messidor-2, o_O reached 97.9% of AUC (sens. = 98.4%, spec. = 79.5%), and we reached 98.2% (sens. = 95.8%, spec. = 83.3%).

The proposed solution yields an improvement close to 5× in terms of efficiency and memory footprint when compared to o_O's method. Furthermore, we have a slightly superior performance in terms of effectiveness, especially considering the difficult cross-dataset validation setup.

5.3. Transfer learning explorations

After testing the proposed screening solution under different conditions, we now investigate the performance of the transfer learning concept to improve decisions about the need of referral. The transfer learning field is conducted as a domain adaptation, which generally arises when the goal is learning an effective model from a source task on a different, but related, target task. In this case, we use the same dataset, but adapt the model from a different but related problem: from severity to referability analyses of diabetic retinopathy.

For this experiment, we do not use the CNN trained in Section 5.1 to explore the transfer learning concept. Rather, as we want to evaluate the potential of using transfer learning from a source problem to a target problem, we initially train, from scratch, a CNN with the same architecture (except that the decision layer has five outputs) to assess severity of diabetic retinopathy incorporating the improvements — data augmentation and multi-resolution training — we discussed in Section 5.1. We also apply robust feature-extraction augmentation, but just for referral assessment when transfer learning comes into play. Note that the robust feature-extraction augmentation is not necessary for the source problem, since diagnosing severity is beyond the scope of the current work. Again, we use the per-patient protocol just whenever it is possible.

We explore transfer learning in its two setups — feature extraction and fine-tuning — explained in Section 3.7. We start the experiments with the DR2 dataset in a strict 5 × 2-fold cross-validation protocol using the same dataset splits reported in [14–16]. Fig. 7 depicts the results obtained with the two transfer learning setups and DR2 dataset. Looking at the classification algorithms individually, we note that Random Forests considerably outperforms the Neural Network classifier. Additionally, the fine-tuning setup excels considerably the feature extraction, reaching an AUC of 98.0%.

We also evaluate the effectiveness of transfer learning with the Messidor-2 dataset. Figs. hyperlinkfig:5x2_MESSIDOR2_transfer_learning_perimage8 and 9 show the ROC curves achieved with Messidor-2

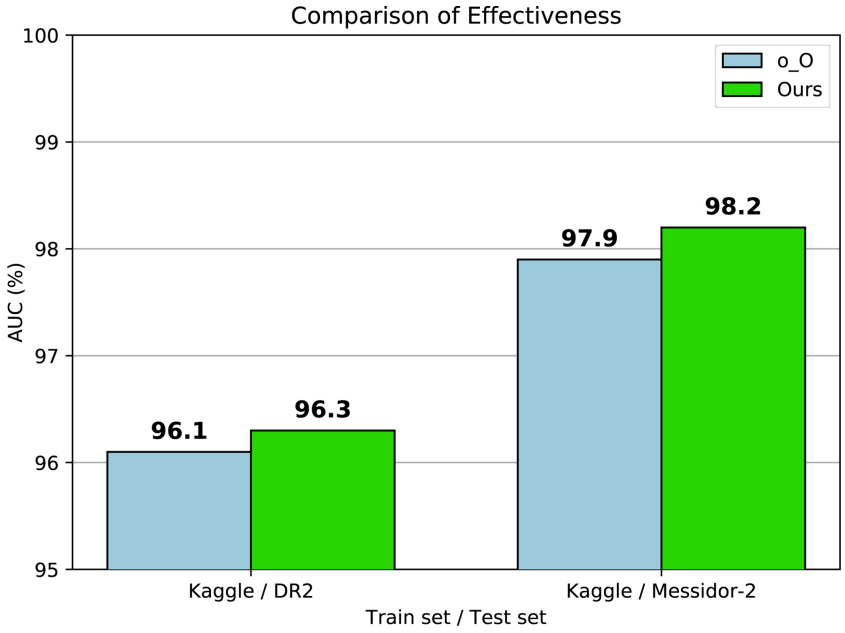


Fig. 6. Comparison of our solution with the o_O's method on a cross-dataset validation protocol in terms of effectiveness (quality of referral assessments).

dataset with transfer learning on per-image analyses and per-patient analysis. Observing the results, we note that Random Forest is superior in the two transfer learning scenarios on a per image or integrated per-patient diagnostic.

When evaluating each eye individually (Fig. 8), we achieve an AUC of 95.3% when we use the frozen CNN to extract features, and improve to 96.0% when we tune the parameters to the target problem. Diagnosing the patient instead of giving a score for each eye is promising, as Fig. 9 shows. The performance using fine-tuning with Random Forest is slightly higher than just using feature extraction: 98.3% over 98.2%.

The results reported herein make it clear that transfer learning with feature extraction is promising for referable diabetic retinopathy detection, and fine-tuning has the potential to enhance considerably the effectiveness of the solution. The result also confirms that a patient-basis diagnostic decision is more effective than just a single-image based decision.

5.3.1. *Comparison with related methods on DR2 dataset*

In this section, we compare our solution with prior work [14–16] that employed the 5 × 2-fold cross-validation protocol over the same DR2 dataset.

Those researches have proposed general frameworks adaptable to large classes of lesions, and recently bypassed the lesion detection and evaluated directly the referability of diabetic retinopathy. Initially, the authors provided referral decision with an AUC of 93.4% [14], further improving it to 94.2% by enhancing the lesion detectors with better mid-level image features [15]. Finally, bypassing lesion detection and directly training custom-tailored referral classifiers, they achieved an

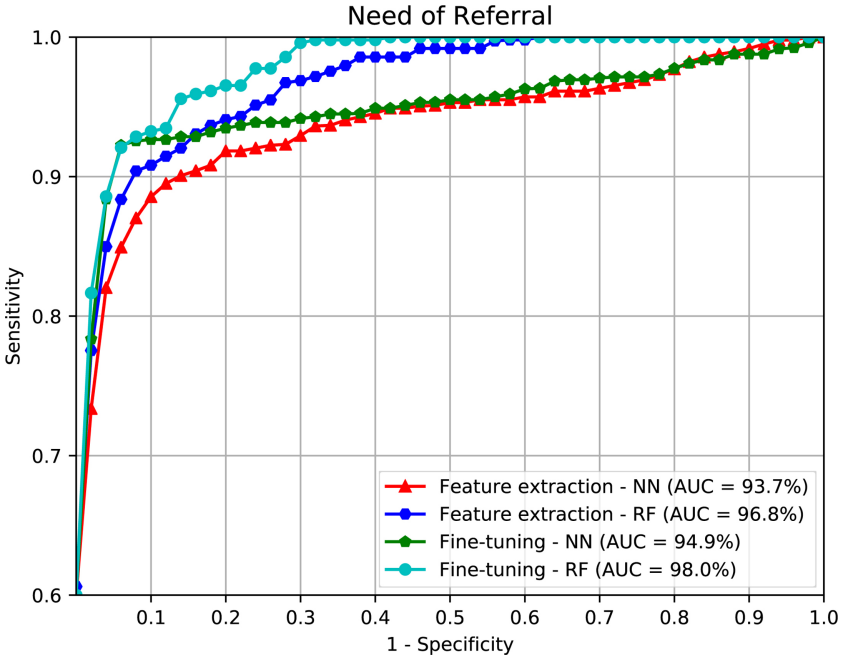


Fig. 7. ROC results for referral assessment evaluating different transfer learning schemes over DR2 dataset. Two classifiers (Neural Network and Random Forest) are used to make the final decision, with fine-tuning and without it (using the CNN for "feature extraction"). The best technique employs Random Forests with fine-tuning.

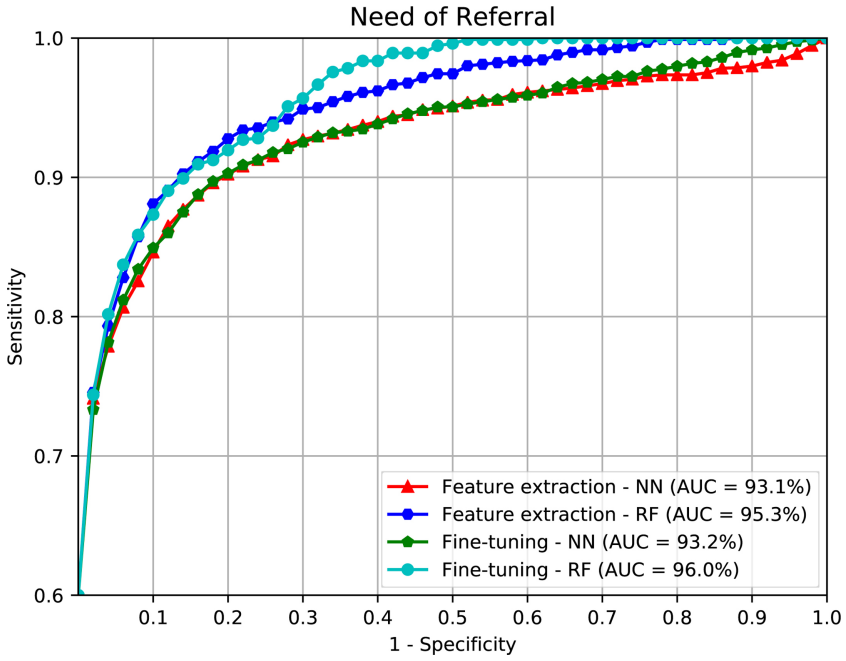


Fig. 8. ROC results for referral assessment using transfer learning over Messidor-2 dataset, on per-image analysis. The analysis provide independent decisions for each eye.

AUC of 96.4% [16]. Putting in context, the data-driven method described here outperforms all previous solutions with an AUC of 98.0% when using the concept of transfer learning by fine-tuning discussed in Section 3. We reduce the classification error by over 44% over the current state of the art [16], 65% over the solution that applied enhanced lesion detectors [15], and 70% over the first referral proposal with DR2 that depends on explicit information of lesions [14].

6. Conclusion

In this paper, we presented a data-driven solution for referable diabetic retinopathy detection inspired on approaches proposed by the second-place teams in two recent and landmark competitions: o_O (Kaggle competition for diabetic retinopathy detection, 2015) and VGG (image classification task of ImageNet ILSVRC-2014). In contrast to the Kaggle competition, here we take into account not only the aim of improving classification accuracies, but we also take into consideration two important issues for real-world deployment: the (computational and implementation) complexity of the solution, and its ability to generalize under the stricter cross-dataset protocol.

In addition to the solution itself, we also offer a novel research methodology regarding a procedural investigation approach. We emphasize that our method is based upon investigating, in a rigorous scientific point of view, the advantages and disadvantages of applying

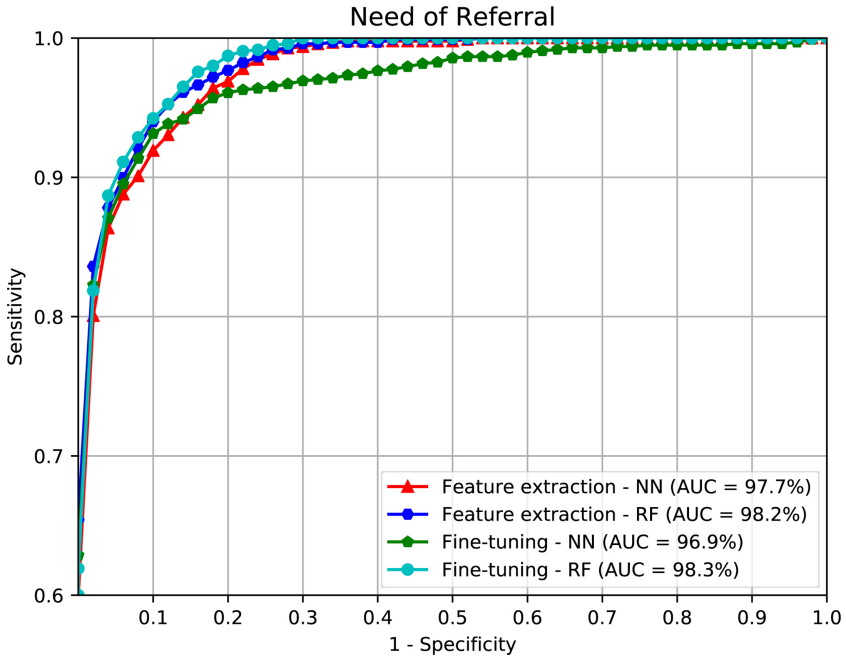


Fig. 9. ROC results for referral assessment using transfer learning over Messidor-2 dataset on per-patient analysis. The results show that combining both eyes systematically outperforms the per-image analysis.

novel and also consolidated techniques, and measuring how much it improves the solution. Our evaluation shows CNNs performance can be boosted by a set of directives. First, good data augmentation is essential for robust decision. Also, a robust feature-extraction augmentation improves performance considerably, while allowing for a diverse choice of machine learning algorithms at the final decision layer. The experiments also show that it is possible (and advantageous) to train high-resolution networks using the weights of low-resolution ones as initialization.

As expected, more information about the patients translates to better-informed decisions for referral. Thus, per patient diagnosis is more effective than per image, even under a challenging cross-dataset protocol. By training with the publicly available Kaggle dataset images and testing on Messidor-2 — both comprising left- and right-eye images for each patient — we achieved an AUC of 98.2%, which is statistically the same performance as the previously published top-performing algorithm, and is within the maximum achievable AUC interval [44] for this dataset. The current best result for referral assessment is 99.0%, reached with an ensemble of 10 CNNs [11]. However, focusing on efficiency and simplicity, in this work we relied upon a single network. This result is also important if we think of deploying a solution such as this one in mobile and low-power devices to reach low-income and often unattended/remote areas.

Another novel aspect of our work is the investigation of the capability of transfer learning in the context of diabetic retinopathy screening, in order to compare data-driven approaches with previous work that employed handcraft methods with relatively small datasets (DR2 and Messidor-2) under cross-validation protocol. The model trained (fine-tuned) with DR2 (AUC of 98.0%) clearly outperforms a recent work for referral assessment over the same dataset [16], reducing the classification error by over 44% (from 3.6% to 2.0%). This is in agreement with recent studies in the literature: a novel family of data-driven methods is the state of the art for diabetic retinopathy screening.

All results point out that it is possible to perform diabetic retinopathy referral diagnostics without relying on preliminary lesion detection — contrarily to what is usually proposed in existing art. But, as the CNN thus trained does not have explicit detectors, currently there is no insight into the potential for seemingly spurious associations possibly arising from confounders in the training data, or the sensitivity to adversarial images in such image-based systems [19–21,9]. For example, sites with high prevalence of a particular manifestation of the disease, may, by chance have operators that acquire patients' images in a particular way, creating spurious associations between the two phenomena. Larger and more diverse training/test datasets — and in particular, the use of cross-dataset protocols — alleviate the problem. Still, strategies of debiasing by analysing confounders, or the most prominent places in the actual retina associated with DR disease, might be associated to data-driven solutions in order to improve performance.

The described data-driven approach to diabetic retinopathy screening has reached high performance in different setups, thus yielding a reliable decision about who needs closer follow-up for diabetic retinopathy. In future work, we want to better understand how the network reaches its decision. Using techniques of activation visualization, among others, we intend to make the decision procedure more interpretable, in order to provide an accountable screening procedure to clinicians. In future work, we also intend to validate the methods with fundus images captured via mobile devices, and intensify investigations of models that efficiently trade off between effectiveness (accuracy) and time consumption/memory footprint, in order to bypass cloud services and embed into portable retinal cameras.

Acknowledgment

We thank the medical team from the Department of Ophthalmology, Federal University of São Paulo, for the data collection and annotation and Dr. Alexandre Ferreira for suggestions on earlier drafts of this work. We thank José A. Stuchi, Flavio P. Vieira and Diego Lencione, co-founders of Phelcom Technologies for encouraging us in investigation of efficient solutions. We thank Kaggle, the California Healthcare Foundation, and EyePACs for sponsoring a diabetic retinopathy competition and for providing the community with a useful and wide dataset. The Messidor-2 dataset was kindly provided by the LaTIM laboratory (see http://latim.univ-brest.fr/) and the Messidor program partners (see http://messidor.crihan.fr/). We also thank the University of Iowa for providing the reference standard of Messidor-2 dataset. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research. Finally, this work was supported in part by Microsoft Research, São Paulo Research Foundation (Fapesp) under the grant #2017/12646-3, the National Council for Scientific Research (CNPq) under grants #311486/2014-2, and 304472/2015-8, Amazon Web Services, CAPES DeepEyes project, CAPES/PNPD, and Google Research. MDA is Director and shareholder of IDx, LLC, Iowa City, and has patents and patent applications, assigned to the University of Iowa, that may compete with the technology that is the subject of this study. IDx, LLC is not associated with the present study and has no interest in the presented methods.

References

[1] International Diabetes Federation. IDF diabetes atlas. 7th ed. 2015 (accessed: 26.6.2017). http://www.idf.org/diabetesatlas.
[2] Vision problems in the U.S., Prevalence of adult vision impairment and age-related eye disease in America. 2015 URL http://www.visionproblemsus.org/(accessed: 26.6.2017).
[3] Gibson DM. The geographic distribution of eye care providers in the united states: implications for a national strategy to improve vision health. Prev Med 2015;73:30–6.
[4] Chou C-F, Zhang X, Crews JE, Barker LE, Lee PP, Saaddine JB. Impact of geographic density of eye care professionals on eye care among adults with diabetes. Ophthalmic Epidemiol 2012;19(6):340–9.
[5] Decencière E, Cazuguel G, Zhang X, Thibault G, Klein J-C, Meyer F, et al. Teleophta: machine learning and image processing methods for teleophthalmology. Ingénierie et Recherche Biomédicale 2013;34(2):196–203.
[6] Quellec G, Lamard M, Erginay A, Chabouis A, Massin P, Cochener B, et al. Automatic detection of referral patients due to retinal pathologies through data mining. Med Image Anal 2016;29:47–64.
[7] Bhaskaranand M, Cuadros J, Ramachandra C, Bhat S, Nittala M, Sadda S, et al. EyeArt + EyePACS: automated retinal image analysis for diabetic retinopathy screening in a telemedicine system. Ophthal Med Image Anal Sec Int Workshop 2015:105–12.
[8] 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 Ophthalmol 2013;131(3):351–7.
[9] Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk J, 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.
[10] Colas E, Besse A, Orgogozo A, Schmauch B, Meric N, Besse E. Deep learning approach for diabetic retinopathy screening. Acta Ophthalmol (Copenh) 2016;94(S256). https://doi.org/10.1111/j.1755-3768.2016.0635.
[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.
[12] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124(7):962–9.
[13] Quellec G, Charrière K, Boudib Y, Cochenerc B, Lamardc M. Deep image mining for diabetic retinopathy screening. Med Image Anal 2017;39:178–93.
[14] Pires R, Jelinek H, Wainer J, Goldenstein S, Valle E, Rocha A. Assessing the need for referral in automatic diabetic retinopathy detection. IEEE Trans Biomed Eng 2013;60(12):3391–8.
[15] Pires R, Jelinek HF, Wainer J, Valle E, Rocha A. Advancing bag-of-visual-words representations for lesion classification in retinal images. PLoS ONE 2014;9(6):e96814. https://doi.org/10.1371/journal.pone.0096814.
[16] Pires R, Avila S, Jelinek H, Wainer J, Valle E, Rocha A. Beyond lesion-based diabetic retinopathy: a direct approach for referral. IEEE J Biomed Health Inform 2017;21(1):193–200.
[17] Quellec G, Russell SR, Abràmoff MD. Optimal filter framework for automated, instantaneous detection of lesions in retinal images, IEEE Trans. Med Imag 2011;30(2):523–33.
[18] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[19] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in Neural Information Processing Systems 2014:2672–80.
[20] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. Int Conf Learn Representations 2014.
[21] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high

confidence predictions for unrecognizable images. IEEE Int Conf Comput Vis Pattern Recog 2015:427–36.
[22] de la Torre J, Valls A, Puig D. A deep learning interpretable classifier for diabetic retinopathy disease grading. arXiv Preprint 2017. [arXiv:1712.08107].
[23] de la Torre J, Valls A, Puig D, Romero-Aroca P. Identification and visualization of the underlying independent causes of the diagnostic of diabetic retinopathy made by a deep learning classifier. arXiv Preprint 2018. [arXiv:1809.08567].
[24] Soto-Pedre E, Navea A, Millan S, Hernaez-Ortega MC, Morales J, Desco MC, et al. Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload. Acta Ophthalmol (Copenh) 2014;93(1):e52–6.
[25] Naqvi S, Zafar M, ul Haq I. Referral system for hard exudates in eye fundus. Comput Biol Med 2015;64:217–35.
[26] Wilkinson C, Ferris III F, Klein R, Lee P, Agardh C, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110(9):1677–82.
[27] Saleh E, Błaszczyński J, Moreno A, Valls A, Romero-Aroca P, de la Riva-Fernández S, et al. Learning ensemble classifiers for diabetic retinopathy assessment. Artif Intell Med 2018;85:50–63.
[28] Tozer K, Woodward MA, Newman-Casey PA. Telemedicine and diabetic retinopathy: review of published screening programs. J Endocrinol Diabetes 2015;2(4):1–10.
[29] Sim DA, Keane PA, Tufail A, Egan CA, Aiello LP, Silva PS. Automated retinal image analysis for diabetic retinopathy in telemedicine. Curr Diabetes Rep 2015;15(3):1–9.
[30] Rocha A, Carvalho T, Jelinek H, Goldenstein S, Wainer J. Points of interest and visual dictionaries for automatic retinal lesion detection. IEEE Trans Biomed Eng 2012;59(8):2244–53.
[31] Jelinek H, Pires R, Padilha R, Goldenstein S, Wainer J, Rocha A. Data fusion for multi-lesion diabetic retinopathy detection. Proc IEEE Comput-Based Med 2012:1–4.
[32] Sidibé D, Sadek I, Mériaudeau F. Discrimination of retinal images containing bright lesions using sparse coded features and SVM. Comput Biol Med 2015;62:175–84.
[33] Pires R, Carvalho T, Spurling G, Goldenstein S, Wainer J, Luckie A, et al. Automated multi-lesion detection for referable diabetic retinopathy in indigenous health care. PLoS ONE 2015;10(6):e0127664. https://doi.org/10.1371/journal.pone.0127664.
[34] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012:1097–105.
[35] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations 2015.
[36] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016:2818–26.
[37] Ng A. Deep learning specialization. 2018. URL https://www.coursera.org/specializations/deep-learning.
[38] Sutskever I, Martens J, Dahl GE, Hinton GE. On the importance of initialization and momentum in deep learning. Intl Conf Machine Learn 2013:1139–47.
[39] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proc Eur Conf Comput Vis 2014:818–33.
[40] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59.
[41] Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. Optimal wavelet transform for the detection of microaneurysms in retina photographs. IEEE Trans Med Imaging 2008;27(9):1230–41.
[42] Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, et al. Feedback on a publicly distributed image database: the Messidor database. Image Anal Stereol 2014;33(3):231–4.
[43] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923.
[44] Quellec G, Abràmoff MD. Estimating maximal measurable performance for automated decision systems from the characteristics of the reference standard. application to diabetic retinopathy screening. IEEE Intl Conf Eng Med Biol Soc. 2014. p. 154–7.