

An Ensemble Deep Learning Based Approach for Red Lesion Detection in Fundus Images

José Ignacio Orlando^{a,b,*}, Elena Prokofyeva^{d,e}, Mariana del Fresno^{a,c}, Matthew B. Blaschko^f

^a*Pladema Institute, UNCPBA, Gral. Pinto 399, Tandil, Argentina*

^b*Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina*

^c*Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, CIC-PBA, Buenos Aires, Argentina*

^d*Scientific Institute of Public Health (WIV-ISP), Brussels, Belgium*

^e*Federal Agency for Medicines and Health Products (FAMHP), Brussels, Belgium*

^f*ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium*

Abstract

Background and objectives: Diabetic retinopathy (DR) is one of the leading causes of preventable blindness in the world. Its earliest sign are red lesions, a general term that groups both microaneurysms (MAs) and hemorrhages (HEs). In daily clinical practice, these lesions are manually detected by physicians using fundus photographs. However, this task is tedious and time consuming, and requires an intensive effort due to the small size of the lesions and their lack of contrast. Computer-assisted diagnosis of DR based on red lesion detection is being actively explored due to its improvement effects both in clinicians consistency and accuracy. Moreover, it provides comprehensive feedback that is easy to assess by the physicians. Several methods for detecting red lesions have been proposed in the literature, most of them based on characterizing lesion candidates using hand crafted features, and classifying them into true or false positive detections. Deep learning based approaches, by contrast, are scarce in this domain due to the high expense of annotating the lesions manually.

Methods: In this paper we propose a novel method for red lesion detection based on combining both deep learned and domain knowledge. Features learned by a convolutional neural network (CNN) are augmented by incorporating hand crafted features. Such ensemble vector of descriptors is used afterwards to identify true lesion candidates using a Random Forest classifier.

Results: We empirically observed that combining both sources of information significantly improve results with respect to using each approach separately. Furthermore, our method reported the highest performance on a per-lesion basis on DIARETDB1 and e-ophtha, and for screening and need for referral on MESSIDOR compared to a second human expert.

Conclusions: Results highlight the fact that integrating manually engineered approaches with deep learned features is relevant to improve results when the networks are trained from lesion-level annotated data. An open source implementation of our system is publicly available at <https://github.com/ignaci Orlando/red-lesion-detection>.

Keywords: Fundus images, Diabetic retinopathy, Red lesion detection, Deep learning.

1. Introduction

One of the most common consequences of vascular damage due to diabetes mellitus is Diabetic Retinopathy (DR), which is one of the leading causes of preventable blindness in the world (Prokofyeva and Zrenner, 2012). As the prevalence of diabetes worldwide is expected to

increase from 2.8% to 4.4% from 2000 to 2030, and about 5% of people with Type-2 diabetes have DR, it is expected that the number of patients suffering from this disease will significantly increase in the next years (Abràmoff and Niemeijer, 2015).

One of the earliest signs of DR are microaneurysms (MAs), which are balloon-shaped deformations on the vessel walls, induced by the permeability of the vasculature due to hyperglycemia (Mookiah et al., 2013). While DR progresses, the number of MAs increases,

*Corresponding author
Email address: jiorlando@conicet.gov.ar
(José Ignacio Orlando)



Figure 1: Examples of red lesions observed in fundus photographs from DIARETDB1 (Kauppi et al., 2007).

and some of them can break and produce leakages of blood on the retinal layers, namely hemorrhages (HEs)¹. The most commonly used term to refer to both MAs and small HEs is “red lesions” (Niemeijer et al., 2005, Decencière et al., 2013, Seoud et al., 2016). The accumulation of blood or lipids induce swelling, which can result in retinal damage when it reaches the macula and, potentially, blindness (Abràmoff et al., 2010).

In its early stages, DR might be clinically asymptomatic (Abràmoff et al., 2010). As a consequence, this condition is typically identified when it is more advanced and treatments are significantly less effective (Mookiah et al., 2013). A recent study has shown that 44% of hospitalized patients with diabetes remain undiagnosed (Kovarik et al., 2016). To prevent this, people suffering from diabetes are usually recommended to be regularly examined through fundus images to verify the non-existence of red lesions (Abràmoff et al., 2010). Although fundus photographs are currently the most economical non-invasive imaging technique for this purpose, manual diagnosis requires an intensive effort to screen the images (Mookiah et al., 2013). Red lesions appear as small red dots that might be subtle and too small to be detected at first glance (Figure 1). Large HEs, on the contrary, are more evident and less difficult to visualize.

Automated methods for computer-aided diagnosis are known to significantly reduce the time, cost, and effort of DR screening: their high throughput ensures the more efficient analysis of large populations (Sánchez et al., 2011). They also reduce the intra-expert variability, which is commonly high due to the small size and the irregular shape of the lesions (Abràmoff and Niemeijer, 2015). These systems are usually aided by an automated module for red lesion detection. In general, the problem of red lesion detection is tackled using a two-stage approach, consisting first of detecting a set of potential candidates, and then refining this set with a classifier trained using hand crafted features (Niemeijer et al., 2005, Walter et al., 2007, Niemeijer et al., 2010, Seoud et al., 2016).

¹In some clinical literature, the acronym HE stands Hard Exudates. However, we use it here to refer to hemorrhages, in line with the biomedical computing literature (Seoud et al., 2016).

Convolutional Neural Networks (CNNs) have recently emerged as a powerful framework to solve a large variety of computer vision and medical image analysis problems (Krizhevsky et al., 2012, Zheng et al., 2015, Venkataramani et al., 2016). Such methods are able to learn features automatically from a sufficiently large training set, without requiring the manual design of the filters. CNNs are known to outperform other manually engineered approaches on a large variety of applications (Razavian et al., 2014). Their discrimination ability is usually affected by the amount of available training data: deeper architectures are known to be able to learn more discriminative features, although at the cost of requiring larger data sets to prevent overfitting and ensure a proper generalization error (Goodfellow et al., 2016). Image level annotations of large scale data sets can be obtained in a relatively economical way (Trucco et al., 2013). However, labeling images at a lesion level is costly, tedious and time consuming, as it requires the intervention of experienced experts who must zoom within different areas of the images to identify every single pathological structure, as accurately as possible. This fact significantly influences the performance of deep learning based approaches for red lesion detection, which must be trained using lesion level annotated data.

In this study we propose to take advantage of both deep learned and manual engineered features for red lesion detection in fundus photographs. In particular, we propose to learn a first set of discriminative features using a light CNN architecture, and then augment their original characterization ability by incorporating hand crafted descriptors. These ensemble vectors of features are used to train a Random Forest classifier that is applied at test time to discriminate between true and false lesion candidates. We experimentally observed that the deep learned features are complementary to the manually engineered, and are aided by the incorporation of domain knowledge.

1.1. Related works

Deep learning methods for DR screening have significantly attracted the attention of the research community after the release of the Kaggle competition database,² which provides a large amount of fundus photographs with image-level annotations. Recently, Gulshan et al. (2016) have presented a CNN that achieved impressive performance for detecting patients with moderate/worse and severe/worse DR. The output of such a method is a

²<https://kaggle.com/c/diabetic-retinopathy-detection>

quantitative indicator of the risk of the patient’s being at a moderate or advanced stage of DR. Red lesion detection methods, by contrast, are intended to identify earlier stages of the disease, providing probability maps that indicate the location of its clinical signs. This feature allows physicians to visually assess the correctness of the results, while helping them to achieve a more reliable and accurate early diagnosis.

Red lesion detection in fundus photographs have been extensively explored in the literature, although most of the existing approaches are based on detecting MAs or HEs separately, and not both structures simultaneously (Niemeijer et al., 2010, van Grinsven et al., 2016, Seoud et al., 2016). Moreover, current existing approaches are based exclusively on hand crafted features. This is likely due to the fact that deep learning based methods have to be trained from large data sets with lesion level annotations. This setting has direct implications on why deep learning based models have been ignored for tackling the problem of red lesion detection. One exception is the method for HEs detection by van Grinsven et al. (2016). This approach is focused on detecting HEs at different scales, which are in general more evident than MAs. By contrast, our method is used for detecting both MAs and small HEs simultaneously, which are more difficult to be visually assessed by physicians.

As mentioned above, in this study we present an ensemble approach that improves the features learned by a CNN by incorporating domain knowledge. Only few efforts have been made in the literature to analyze the viability of such an approach. Annunziata and Trucco (2016), for instance, propose to initialize a convolutional sparse coding approach with manually designed filters to accelerate its learning process and improve their original discriminative power. That approach is applied for detecting curvilinear structures such as neurons or retinal vessels, which are easier to manually trace. Venkataramani et al. (2016) have observed that state of the art descriptors significantly improve the performance of transferred CNN features when applied to kidney detection in ultrasound images. The main difference with respect to our approach is that our CNN is trained from scratch from a domain specific data set, while the approach of Venkataramani et al. (2016) is based on fine-tuning a CNN trained from natural images.

From our literature review, we identified two main methods resembling our approach, although with different applications and based on different CNN architectures. Zheng et al. (2015) introduce a method for identifying landmarks in 3D CT scans using the output of a dedicated CNN in combination with Haar fea-

tures to boost the quality of the results. Its deep learning based component is divided into two stages: a first stage, based on a light architecture with only one hidden layer, is used to recover a large set of landmark candidates; the second stage, made up of three hidden layers and trained using sparsity priors, is used to recover a large vector of neural network features, which is combined with Haar features to train a probabilistic boosting-tree classifier. In order to save as much data as possible for training the CNN and the lesion classifier, we avoided performing candidate detection in a supervised way. Instead, a combination of morphological operations and image processing techniques is used to retrieve potential lesions, without using training data. This allows us to train a slightly deeper architecture in the subsequent stage, only dedicated to classifying the lesion candidates, which is able to capture discriminative features from the training patches.

The method by Wang et al. (2014) for mitosis detection on histopathology images is also similar to ours. It uses candidate detection as well, and a RF classifier trained using hand crafted features is applied to assign a probability of being a true mitosis candidate. In parallel, a CNN with two convolutional layers and one fully connected layer is trained from patches around the candidates to retrieve an additional probability. The final decision is performed via consensus of the predictions of the two classifiers by weighting both probabilities using two manually tuned parameters. We took the alternative approach of using both feature vectors simultaneously to train the RF classifier, as it can take advantage of the interaction between both the deep learned and the hand crafted features.

1.2. Contributions

In this paper we propose to learn discriminative models for red lesion detection by combining both deep learned and hand crafted features. First, an unsupervised, candidate detection approach based on morphological operations is applied to retrieve a set of potential lesions. Next, a CNN is trained from a set of patches around the candidate lesions to learn a first feature vector. These descriptors are augmented with a set of hand crafted features to improve their ability to distinguish the true positive lesions. A Random Forest (RF) classifier is trained using this hybrid feature vector, and is then applied for refining the set of candidates, discriminating between true lesions and false positives. We empirically observed that combining both sources of information improved performance not only when evaluating our method on a per-lesion basis but also when

analyzing its potential for DR screening or need for referral detection on an image-level basis. Our results on benchmark data sets such as e-ophtha (Decencière et al., 2013), DIARETDB1 (Kauppi et al., 2007) and MESSIDOR (Decencière et al., 2014) show that our strategy outperforms other state of the art methods that are not only based on red lesion detection but also in detecting other pathological structures such as exudates or neovascularizations. An extensive analysis of the complementarity of the deep learned features with respect to the hand crafted ones is also provided, with the purpose of assessing their contribution in the discrimination process.

2. Methods

A schematic representation of our method is depicted in Figure 2. Lesion candidates retrieved with morphological operations (Section 2.1) are filtered using a set of hybrid descriptors. Regular patches centered on each candidate connected component are collected to build a training set that is used to train a CNN (Section 2.2). A 63-dimensional vector of hand crafted features (Section 2.3) is also computed per each of the candidates. A Random Forest (RF) classifier (Section 2.4) is afterwards trained on the resulting combination of features, and used to classify the new candidates. Since the presence of red lesions is the first indicator of DR, the maximum over lesion likelihoods is used to assign a DR probability, as done by Seoud et al. (2016) and Antal and Hajdu (2012).

2.1. Candidate detection

Our strategy for candidate detection is illustrated in Figure 3. First, the green band G from the original color image I is taken, since it is the one that allows a better visual discrimination of the red lesions. To avoid artifacts in the borders of the FOV that might hide potential lesions (Figure 4(b)), a wider aperture of $\frac{3}{30}\mathcal{X}$ pixels is simulated (Soares et al., 2006) from G , where \mathcal{X} corresponds to the width in pixels of the field of view (FOV). Since our purpose is to develop a system sufficiently general to be applied at different image resolutions, all the relevant parameters are expressed in terms of \mathcal{X} .

As uneven background illumination might hide potential lesions occurring within the darkest areas of the images, a r -polynomial transformation is applied on

pixel intensities:

$$I_W(i, j) = \begin{cases} \frac{\frac{1}{2}(u_{\max} - u_{\min})}{(\mu_W(i, j) - \min(G))^r}, & G(i, j) \leq \mu_W(i, j) \\ \frac{-\frac{1}{2}(u_{\max} - u_{\min})}{(\mu_W(i, j) - \max(G))^r}, & G(i, j) > \mu_W(i, j) \end{cases} \quad (1)$$

with $r = 2$, $u_{\min} = 0$ and $u_{\max} = 1$, respectively, and where μ_W is the local average intensity on square neighborhoods of length W , computed for each (i, j) pixel (Walter et al., 2007). We observed that using $W = 25$ performed sufficiently well for enhancing images with 536 pixels of horizontal resolution such as those in the DRIVE data set (Niemeijer et al., 2004), so this parameter is automatically adjusted using $W = \frac{25}{536}\mathcal{X}$. Figure 4 illustrates how the expansion of the FOV border and the subsequent intensity transformation improve the contrast of subtle lesions located in the border of the FOV.

A Gaussian filter with $\sigma = 5$ is applied to I_W to reduce noise, resulting in a new image I'_W . Afterwards, different morphological closings are performed on I'_W using linear structuring elements of length $l \in L$ at angles θ spanning from 0 to 180° with increments of 15° . The set of relevant scales L is a fixed parameter that is also automatically adjusted in terms of \mathcal{X} , as explained in Section 3.2. By taking the minimum response over all the considered angles, an image $I_{\text{closed}}^{(l)}$ is obtained in which responses to lesions with sizes smaller than l were reduced, and all the remaining structures are still preserved (Walter et al., 2007). A score map is then obtained by:

$$I_{\text{cand}}^{(l)} = I_{\text{closed}}^{(l)} - I_W. \quad (2)$$

Afterwards, a thresholding operation is applied on $I_{\text{cand}}^{(l)}$, where the threshold is automatically determined in such a way that a maximum of $K = 120$ candidates are retrieved from the score map. In order to achieve this goal, thresholds t_s from $\min(I_{\text{cand}}^{(l)})$ to $\max(I_{\text{cand}}^{(l)})$ with increments of 0.002 are explored until the number of connected components in the resulting binary maps is less than or equal to K . To support the cases in which no lesions are detected or when is not possible to detect less than K candidates, a lower bound t_l and an upper bound t_u are experimentally set such that:

$$t_K = \begin{cases} t_l, & \forall t_s : \text{CC}(I_{\text{cand}}^{(l)} > t_s) < K \\ t_k, & \text{CC}(I_{\text{cand}}^{(l)} > t_s) \leq K \\ t_u, & \forall t_s : \text{CC}(I_{\text{cand}}^{(l)} > t_s) > K \end{cases} \quad (3)$$

where CC is a function than counts the number of connected components in the thresholded score map. Once t_K is fixed, a binary map of candidates is obtained by

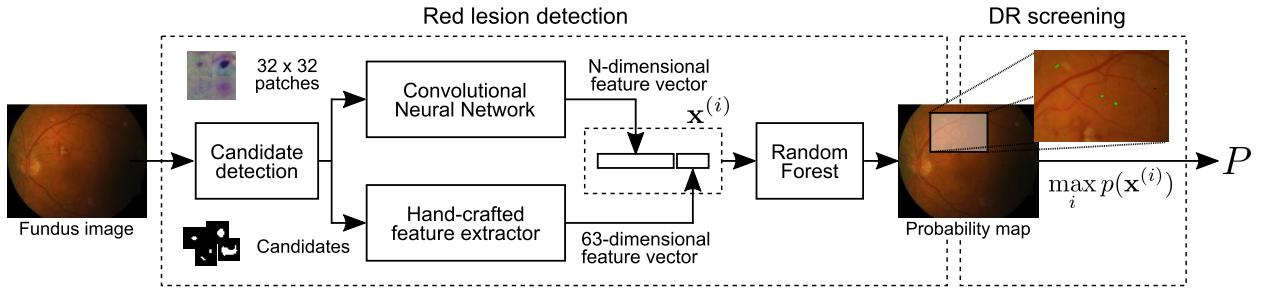


Figure 2: Overview of our method for red lesion detection.

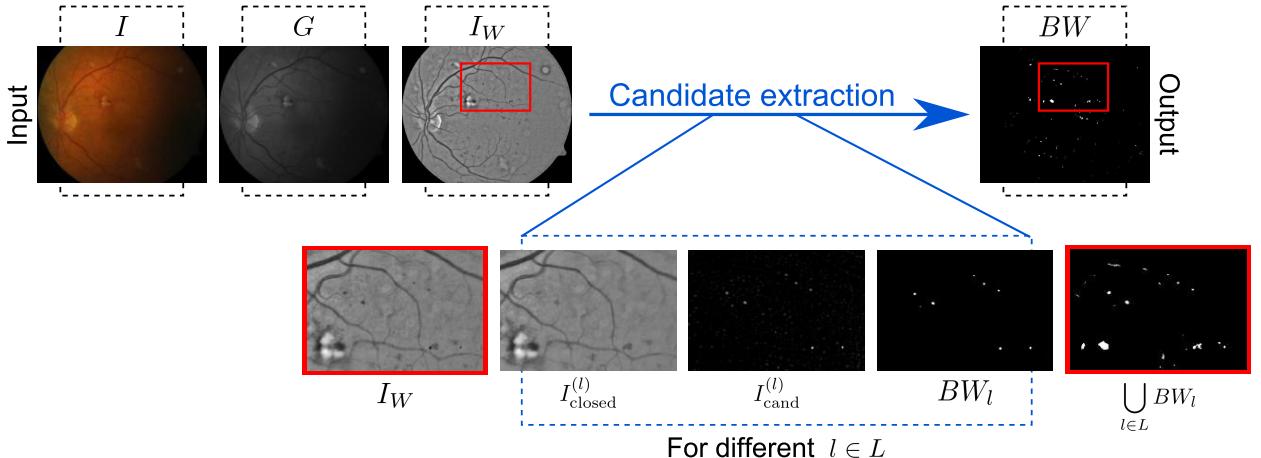


Figure 3: Red lesion candidate detection. See Section 2.1 for a detailed description of the process.

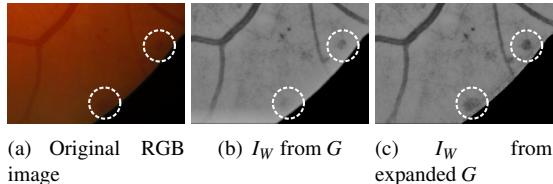


Figure 4: Effect of the FOV expansion on the lesion candidates located closely to the border of the FOV.

thresholding $BW_l = I_{\text{cand}}^{(l)} > t_K$ (Walter et al., 2007). This operation is repeated for different values of $l \in L$ to capture potential lesions at different scales, so the binary map of candidates BW is obtained as $BW = \bigcup_{l \in L} BW_l$. Finally, as BW might include small candidates which usually are not associated to any pathological region but with noise, all connected structures in BW with less than px pixels are discarded. The automated model selection procedure used to set the values of K and px and the scales in L is described in Section 3.2.

Figure 5 presents a random sample of the potential lesions retrieved by the method on a randomly selected

image from DIARETDB1 training set. It is possible to see that most of false positive samples correspond to vascular branching or crossing points, vessel segments and beadings, scars due to laser photocoagulation or black spots of dirt in the capture device, as reported by Seoud et al. (2016). This setting underlines the importance of refine the candidates to remove false positives.

2.2. CNN-based features

We train a dedicated CNN to characterize each red lesion candidate. For this purpose, each color band of the original image I is equalized first as proposed by van Grinsven et al. (2016):

$$I_{\text{ce}}(i, j; \sigma) = \alpha \cdot I(i, j) + \tau \cdot \text{Gaussian}(i, j; \sigma) * I(i, j) + \gamma \quad (4)$$

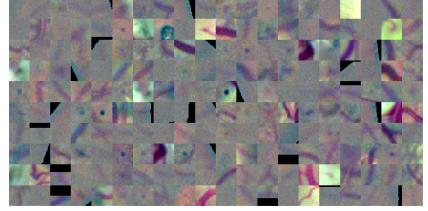
where $*$ is a convolution, the Gaussian filter has a standard deviation $\sigma = \frac{\chi}{30}$, and $\alpha = 4$, $\tau = -4$ and $\gamma = 128$ were set following van Grinsven et al. (2016). We empirically observed that this preprocessing operation not only dramatically diminishes the number of epochs needed for training but also improves the discrimination ability of the CNN. As explained in Section 2.1, a wider

FOV is also simulated for each color band to prevent any undesired effect in the FOV border.

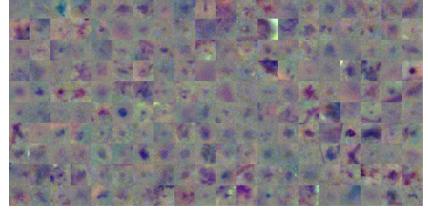
A training set $S = \{(X^{(i)}, y^{(i)})\}, i = 1, \dots, n$ is built for training the CNN, where each sample $X^{(i)}$ is a square patch around the center of each red lesion candidate, as extracted from I_{ce} (Figure 5). The patch size is taken as double the length of the major axis of the candidate, or 32×32 pixels if the major axis of the candidate is smaller than 32 pixels. This setting let us to recover not only the candidate itself but also its surrounding area, which allows the CNN to capture both candidates' internal features and information about its shape, borders and context. Patches larger than 32×32 pixels are down-sized to this resolution to ensure a uniform input size for the CNN. As windows are square by definition, this transformation is isotropic and does not affect the appearance of the lesion candidate. Samples are centered by subtracting the training set mean image. Alternative scaling methods such as ZCA whitening and contrast normalization were also analyzed, although no improvements were observed on the validation set when applying them. The label $y^{(i)} \in \{0, 1\}$ associated to the candidate is assigned according to the ground truth labeling on the data set: if the candidate overlaps with a true labeled region, then the window is assumed to be a true red lesion ($y = 1$); if it does not, then it is assumed to be a false positive ($y = 0$). The CNN is trained from scratch on an 8x augmented version of this training set, which is obtained by rotating each patch by 90° , 180° or 270° , and then flipping the resulting windows horizontally or not. Thus, for each input patch, 8 new patches are generated.

The CNN architecture is depicted in Table 1. It comprises 4 convolutional layers and 1 fully connected layer with 128 units. This layer is used to retrieve the $N = 128$ -dimensional vector of deep learned features. The CNN was designed by using the original LeNet architecture as the initial baseline, and introducing changes by evaluating their contribution on reducing the empirical error on a held-out validation set, which was randomly sampled from the training set. Using deeper architectures such as VGG-S or Inception-V3 was avoided as the increase in the number of parameters would have required a larger training set to reduce overfitting.

Depending on the number of images in the training set with an advanced level of DR and the size of the lesions we focus on detecting, the classification problem is imbalanced to a greater or lesser degree. Hence, the proportion of true positive lesions might be significantly smaller than the number of false positive ones. If this imbalance grows dramatically, it was previously observed



(a) Non-lesions (false positive candidates)



(b) Lesions (true positive candidates)

Figure 5: CNN training set. Random sample of 200 patches for (a) non lesions and (b) true lesions. See Section 2.2 for details of the construction of the training set.

that the typical cross-entropy loss is affected, and, as a consequence, fewer true positives are retrieved (Maninis et al., 2016). Thus, we used a class balanced cross-entropy loss, given by:

$$\begin{aligned} \mathcal{L}_\beta(\mathbf{W}) = & -\beta \sum_{i \in Y_+} \log P(y^{(i)}|X^{(i)}; \mathbf{W}) \\ & - (1 - \beta) \sum_{i \in Y_-} \log P(y^{(i)}|X^{(i)}; \mathbf{W}) \end{aligned} \quad (5)$$

where \mathbf{W} are the weight parameters of the network; P is the probability obtained by applying a sigmoid function to the activation of the fully connected layer; Y_+ and Y_- are the subsets of true and false positive samples, respectively; and $\beta = |Y_-|/(|Y_+| + |Y_-|)$ is the ratio of negative vs. positive samples in S .

The CNN's weights were randomly initialized from a Gaussian distribution with a lower standard deviation for the first layer (0.01) than for the remaining ones (0.05), to prevent vanishing gradients. Standard dropout after all the convolutional layers was analyzed as an alternative to the reported architecture, although it was observed that it did not improve results on the validation set. Moreover, using such an approach increased the training time significantly. We noticed instead that using dropout after the first convolutional layer with a high keep probability $p = 0.99$ slightly improved results. We also used weight decay of 5×10^{-4} for regularization, to penalize large \mathbf{W} values during backpropagation. Batch normalization was also evaluated but no improvements were observed when applying it. The

Table 1: CNN architecture. Convolutional layers (conv) indicate width, height and depth of each learned filter. Pooling layers (pool) include the dimension of the pooling operation and the stride. Dropout is only applied after the first convolutional layer with a low dropout probability.

Block	Layers	Filter size	Output size
1	conv	$5 \times 5 \times 3$	32
	maxpool	3×3 - stride = 2	
	dropout	$p = 0.01$	
2	conv	$5 \times 5 \times 32$	32
	avgpool	3×3 - stride = 2	
3	conv	$5 \times 5 \times 32$	64
	avgpool	3×3 - stride = 2	
4	conv	$4 \times 4 \times 64$	N
5	fully connected	N	N
6	$\mathcal{L}_\beta(\mathbf{W})$	N	2

cost function was optimized using stochastic gradient descent, with a batch size of 100 samples taken from the training set, which is randomly shuffled at the beginning of each epoch. The learning rate was initially fixed in $\eta = 0.05$, and it was divided by a factor of 2 every time that the relative improvement in current $\mathcal{L}_\beta(\mathbf{W})$ value was less than 1% of the average loss over the last 10 epochs. The optimization process was stopped when this relative difference was smaller than 10^{-4} , or a maximum number of 200 epochs was achieved. The CNN was implemented in MATLAB R2015b, using Matconvnet (Vedaldi and Lenc, 2015). To improve performance during training, a NVIDIA Titan X GPU card was used, achieving convergence in 20-30 minutes.

2.3. Hand-crafted feature extraction

As a complementary source of information with respect to the CNN features, a 63 dimensional feature vector of hand-crafted features (HCF) is extracted per each lesion candidate and incorporated to our feature vector. Some of these descriptors were extensively explored in the literature (Niemeijer et al., 2005, 2010, Seoud et al., 2016), while other are introduced here to improve the existing ones. In general, they can be divided into two categories: intensity based and shape based features (Table 2).

Intensity features exploit the visual properties of the candidate areas, and are extracted from different versions of the color image I , obtained by applying different preprocessing strategies. In particular, we extracted descriptors used in the state of the art (Niemeijer et al., 2010, Seoud et al., 2016) but from the following derived images:

- Original red, green and blue color bands (R , G and B , respectively).

- Green band G after illumination correction (I_W , obtained as in Section 2.1).
- Color bands and I_W after CLAHE contrast enhancement (R_c , G_c , B_c , I_{Wc}).
- Color bands after color equalization (R_{ce} , G_{ce} , B_{ce}).
- I_{SC} , which is the difference between the green band G and an estimated background I_{BG} , obtained using a median filter with squared windows of length $\frac{25}{536}\mathcal{X}$.
- I_{match} . This image is obtained by initially computing I_{lesion} , which is a vessel free version of I_{SC} , obtained by inpainting the vasculature as in (Orlando et al., 2017b). The difference between each pixel (i, j) in I_{lesion} and its 11×11 neighborhood is assigned to $I_{match}(i, j)$.
- $I_{cand} = \max_l I_{cand}^{(l)}$, which is the maximum response to the candidate score map described in Section 2.1, taken from I_W , but restricting the size of the structuring elements to the lengths $l \in \{5, 7, \dots, 15\}$.

Shape based features have the ability to characterize the structure of the candidates. Red lesions are expected to be relatively circular, with small area and perimeter, and approximately equal minor and major axis. Such statistics, including compactness, eccentricity and aspect ratio, are also included as part of the domain knowledge feature vector.

We also analyzed the viability of using the segmentation of the retinal vasculature as a potential source of information. As seen in Figure 5(a), most of the false positive detections are located in vessel crossings or beadings. Thus, we compute an initial vessel segmentation using the method reported in (Orlando and Blaschko, 2014, Orlando et al., 2017a), and postprocessing the output by removing every spurious connected component with less than $\frac{100}{536}\mathcal{X}$ pixels Orlando et al. (2017b). A morphological closing with a disk of radius 2 is afterwards applied to fill any gap due to the central reflex in arteries. Then, we measure the ratio of pixels in the candidate region that overlap with the segmentation, divided by the number of pixels in the candidate. Figure 6 illustrates the process of computing this feature. It can be seen that most of the false positive lesions located at the optic disc overlap with the resulting segmentation mask, and can be removed by this descriptor.

Table 2: Summary of the hand crafted features used to complement our CNN. Top: intensity based features. Bottom: shape based features.

Intensity based features (dimensionality)	Extracted from
Average intensity value in the candidate region. (13)	$R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}, I_{top-hat}$
Sum of intensities in the candidate region. (12)	$R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}$
Standard deviation of intensities in the candidate region. (12)	$R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}$
Contrast: Difference between mean intensity in the candidate region and mean intensity of the dilated region (12)	$R, G, B, I_W, R_c, G_c, B_c, I_{Wc}, R_{ce}, G_{ce}, B_{ce}, I_{SC}$
Normalized total intensity: Difference between total and mean intensities of the candidate area in I_{BG} , divided by the candidate's standard deviation in I_{BG} . (3)	G, I_{SC}, I_W
Normalized mean intensity: Difference between mean intensity in I_W and mean intensity of the candidate area in I_{BG} , divided by the standard deviation of the candidate in I_{BG} . (1)	I_W
Minimum intensity in the candidate area. (1)	I_{match}
Shape based features (dimensionality)	Extracted from
Area: Number of pixels of the candidate. (1)	BW
Perimeter: Number of pixels on the border of the candidate. (1)	BW
Aspect ratio: Ratio between the major and minor axis lengths. (1)	BW
Circularity = $4\pi \text{Area}/\text{Perimeter}^2$. (1)	BW
Compactness = $\sqrt{(\sum_{j=1}^n d_j - \bar{d})/n}$, where d_j is the distance from the centroid of the object to its j th boundary pixel and \bar{d} is the mean of all the distances from the centroid to all the edge pixels. n is the number of edge pixels. (1)	BW
Major axis of the ellipse that has the same normalized second central moments as the candidate region. (1)	BW
Minor axis of the ellipse that has the same normalized second central moments as the candidate region. (1)	BW
Eccentricity of the ellipse that has the same second-moments as the candidate region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. (1)	BW
Ratio of the pixels on the candidate region that are also included in the binary segmentation of the retinal vasculature, obtained as in (Orlando et al., 2017b). (1)	Vessel segmentation

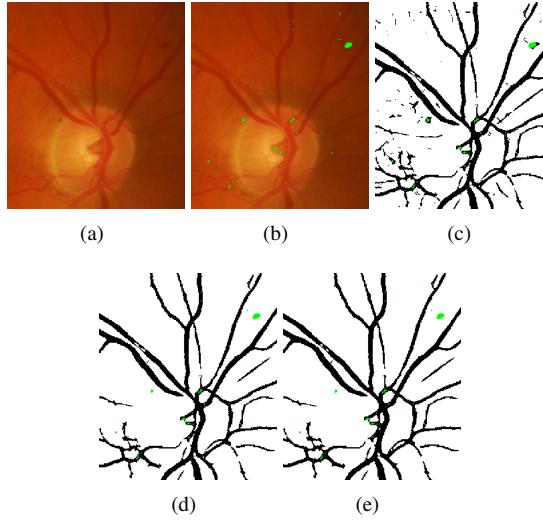


Figure 6: Feature based on vessel segmentation. (a) I . (b) I with candidates superimposed. (c) Vessel segmentation. (d) Vessel segmentation after removing spurious elements. (e) Vessel segmentation after morphological closing.

2.4. Candidate classification with Random Forest

A Random Forest (RF) is an ensemble classifier that is widely used in the literature due to its capability to perform both classification and feature selection simultaneously (Breiman, 2001, Lo Vercio et al., 2017). It is also robust against overfitting, which is relevant when having small training sets, and is suitable to deal with noisy, high dimensional imbalanced data. We trained this classifier for the purpose of refining our set of candidates using our hybrid feature vector. In all our experiments, we standardized the features to zero mean and unit variance.

A RF is a combination of T decision trees. These trees are learned from T examples that are randomly sampled with replacement from our training set S . Each node in a tree corresponds to a split made using the best of a randomly selected subset of $m = \sqrt{d}$ features, with d being the dimensionality of the feature vector. The quality of the split is given by the decrease in the Gini index that the split produces (Breiman, 2001). Given a feature vector $\mathbf{x}^{(j)}$, the RF evaluates the conditional probability $p_i(c|\mathbf{x}^{(j)})$, where $c \in \{-1, 1\}$ is the class—with -1 corresponding to a non lesion and 1 to a true lesion—and i is the index of the tree in the forest. The final probability is then computed by repeating this process for every tree $0 < i \leq T$, and averaging the responses of

Table 3: Distribution of DR grades in the MESSIDOR data set, and diagnostic criterion. MA = microaneurysms, HE = hemorrhages and NV = neovascularizations.

Grade	Criteria	Num. images
R0	$(N_{MA} = 0) \text{ AND } (N_{HE} = 0)$	546
R1	$(0 < N_{MA} \leq 5) \text{ AND } (N_{HE} = 0)$	153
R2	$(5 < N_{MA} < 15) \text{ AND } (0 < N_{HE} < 5) \text{ AND } (N_{NV} = 0)$	247
R3	$(N_{MA} \geq 15) \text{ OR } (N_{HE} \geq 5) \text{ OR } (N_{NV} > 0)$	254

each of them:

$$p(c|\mathbf{x}^{(j)}) = \frac{1}{T} \sum_i^T p_i(c|\mathbf{x}^{(j)}) \quad (6)$$

In order to determine the probability P of the image I corresponding to a DR patient or not, we followed the same procedure used by Seoud et al. (2016):

$$P(I) = \max_j p(c = 1|\mathbf{x}^{(j)}), \quad (7)$$

which means that for a given image I with m lesion candidates, the probability of being DR will be associated with the maximum certainty of the classifier of having observed a true positive lesion ($c = 1$).

3. Experimental setup

3.1. Materials

We conducted experiments using three publicly available data sets: DIARETDB1³ (Kauppi et al., 2007), e-ophtha⁴ (Decencière et al., 2013), and MESSIDOR⁵ (Decencière et al., 2014)

DIARETDB1 and e-ophtha were used to perform a per-lesion evaluation as they provide lesion level annotations. MESSIDOR provides image level annotations indicating the DR grade, assigned using the criterion detailed in Table 3. Thus, this set was used to quantify the performance of our method as a DR screening tool, on a per-image basis. We also used e-ophtha for this purpose, by generating image-level annotations based on the number of red lesions in the ground truth segmentation. Thus, any image with at least one red lesion was labeled as DR. The ROC⁶ (Niemeijer et al., 2010) training set, which comprises 50 fundus photographs taken at different resolutions, was used to augment DIARETDB1 training set for small red lesion detection on

³<http://www.it.lut.fi/project/imageret/diaretbdb1/>

⁴<http://www.adcis.net/en/Download-Third-Party/E-Ophtha.html>

⁵<http://messidor.crihan.fr>

⁶<http://webeye.ophth.uiowa.edu/ROC/>

e-ophtha. Further details about the experimental setup are provided in Table 4.

DIARETDB1 consists of 89 color fundus images taken under varying imaging settings (Kauppi et al., 2007). 84 images contain signs of mild or pre-proliferative DR, and the remaining 5 are considered normal. The entire set is divided into a training set and a test set of 28 and 61 images, respectively. Four different experts have delineated the regions where MA and HE can be found, and a consensus map is provided per each type of lesion. The standard practice is to evaluate MA or HE detection methods at a conservative $\geq 75\%$ agreement (Kauppi et al., 2007). For red lesion detection, however, Seoud et al. (2016) propose to use as ground truth the union of the consensus maps for both MAs and HEs at a $> 25\%$ level of agreement. We followed this latter approach to evaluate our red lesion detection strategy.

e-ophtha (Decencière et al., 2013) is a database generated from a telemedical network for DR screening, and it includes manual annotations of MAs and small HEs. It comprises 148 images with small red lesions, and 233 with no visible sign of DR. In order to obtain per-image labels indicating the presence or absence of DR, images with any red lesion were labeled as DR.

Finally, MESSIDOR (Decencière et al., 2014) comprises 1200 color fundus images acquired by 3 ophthalmic institutions in France. Images were originally captured at different resolutions, and graded into four different DR stages, being R0 the healthy category and R3 the most severe. Two different classification problems are usually derived from MESSIDOR grades: DR screening, which corresponds to distinguishing R0 from the remaining R1, R2 and R3 grades (Antal and Hajdu, 2012, Seoud et al., 2016); and detecting the need for referral, which corresponds to R0 and R1 vs. R2 and R3 grades (Sánchez et al., 2011, Pires et al., 2013). We evaluated our method on a per image basis following these two approaches.

Since these data sets do not include FOV masks, which are necessary for processing the images, we automatically generate them by thresholding the luminosity plane of the CIELab version of the RGB images at 0.15 (for DIARETDB1, e-ophtha and MESSIDOR) and 0.26 (for ROC) (Orlando et al., 2017b). If the resulting binary mask is such that the entire image is estimated as a foreground, an alternative approach is applied where the RGB bands are summed up and the resulting image is thresholded at an empirically tuned value of 150. To smooth borders and reduce noise, all masks are post-processed with a median filter using square windows of side 5, and only its largest connected component is

preserved. In principle, these masks would be available directly from the fundus camera, and the process of replicating this information directly from the images is a necessary but not central task to the present paper. The FOV masks for all the data sets used in this paper are released in the project webpage (Section 6).

3.2. Model selection

Candidate detection relies on three significant parameters: L , which is the set of scales used to retrieve potential candidates; K , the number of candidates retrieved for a given scale; and px , the minimum area in pixels that a candidate must have. In our experiments, these values were experimentally adjusted using the DIARETDB1 training set, resulting in $L = \{3, 6, 9, \dots, 60\}$, $K = 120$ and $px = 5$. The maximum scale from L was adapted on the remaining data sets using a scaling factor of $\frac{X}{1425}$, where 1425 is the average width of the images in DIARETDB1. This allows to recover a set of candidates with a size proportional to the resolution of each image.

The parameters of the CNN (in particular, dropout probability $1 - p$ and the size of the fully connected layer N) were designed according to the performance on a held out validation set, randomly sampled from each training set. The parameters that maximized the area under the precision/recall curve ($N = 128$ and $p = 0.99$) were always used for evaluation on the test set. The number of trees $T \in \{100, 120, \dots, 200\}$ for the RF was fixed to the value that minimized the out-of-bag error on the training set on each experiment (Breiman, 2001). The maximum number of possible trees was fixed to a relatively low value (200) to reduce the computational cost during training and prediction. Nevertheless, experiments adding up to 2000 trees to the model did not show any improvements in reducing the out-of-bag error.

3.3. Evaluation metrics

Free-response ROC (FROC) curves were used to evaluate the performance of our red lesion detection method on a per lesion basis. These plots, which are extensively used in the literature to estimate the overall performance on this task, represent the per lesion sensitivity against the average number of false positive detections per image (FPI) obtained on the data set for different thresholds applied to the candidate probabilities. Thus, FROC curves provide a graphical representation of how the model is able to deal with the detection of true lesions in all the images of the data set. We also computed the Competition Metric (CPM) as proposed in the Retinopathy Online Challenge (Niemeijer

Table 4: Experimental setup. β is the value for the balanced cross-entropy loss (Equation (5)).

Exp. ID	Detection	Training set	GT labels	True lesions	Non lesions	β	Per lesion evaluation	Per image evaluation
1	Red lesions with multiple sizes	DIARETDB1 training set (28 images)	MA > 25% \cup HE > 25 %	1059 (27%)	2905 (73%)	$\beta = 0.5$	DIARETDB1 test set	MESSIDOR
2	Small red lesions	DIARETDB1 & ROC training sets (78 images)	MA > 75% from DIARETDB1 & ROC MA labels	407 (4%)	10282 (96%)	$\beta \sim 0.96$	e-ophtha	e-ophtha

et al., 2010), which is the average per lesion sensitivity at the reference FPI values $\in \{1/8, 1/4, 1/2, 1, 2, 4, 8\}$. The protocol used by Seoud et al. (2016) was followed when evaluating in DIARETDB1, as indicated in Section 3.1.

When evaluating on a per image basis, we used standard ROC curves, where both the sensitivity ($Se = \frac{TP}{TP+FN}$) and 1 – specificity ($Sp = \frac{TN}{FP+TN}$) are depicted within the same plot for different DR probability values, obtained as indicated in Equation 7. Additionally, we studied the Se at $Sp = 50\%$, which is a standard comparison metric for screening systems (Sánchez et al., 2011).

4. Results

4.1. Per lesion evaluation

Two different experiments were conducted for per lesion evaluation, as detailed in Table 4. FROC curves are used for comparison, and Wilcoxon signed rank tests were performed to estimate the statistical significance of the differences in the per lesion sensitivity values. These tests were conducted using 100 sensitivity values retrieved for logarithmically spaced FPI values in the interval $[\frac{1}{8}, \dots, 8]$, which corresponds to a more dense version of the reference FPI values used for computing the CPM (Niemeijer et al., 2010).

Experiment 1 evaluates the model ability to deal with both MAs and HEs simultaneously at multiple scales, following the same protocol as Seoud et al. (2016) (Figure 7). Results obtained by Seoud et al. (2016) were provided by the authors and obtained using the same training and test configuration, and are included for comparison purposes. Hypothesis tests show a statistically significant improvement in the per lesion sensitivity values when using the combined approach compared to using each representation separately ($p < 2 \times 10^{-18}$ and $p < 4 \times 10^{-17}$ for the CNN probabilities and the hand crafted features, respectively). Moreover, the hybrid method reported better results compared to Seoud et al. ($p < 2 \times 10^{-18}$).

As DIARETDB1 includes labels for both MAs and HEs, it is possible to quantitatively assess the accuracy

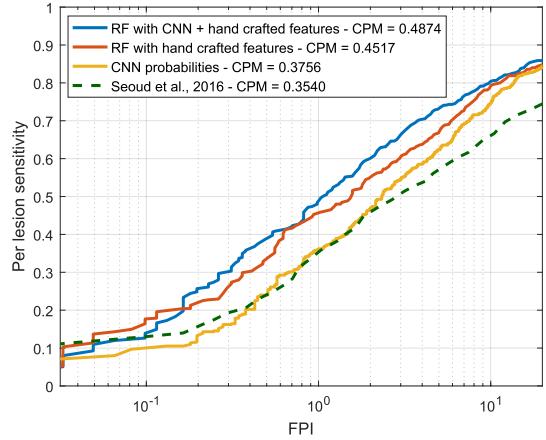


Figure 7: Per lesion evaluation in Experiment 1. FROC curve and CPM values obtained on the DIARETDB1 test set.

of the method to detect each type of lesion. Figure 8 illustrates the FROC curves and the CPM values obtained by the models learned in Experiment 1, when analyzing MAs and HEs separately. For MA detection, the combined approach achieves higher per lesion sensitivity values than using each approach separately ($p < 2 \times 10^{-18}$ and $p < 3 \times 10^{-17}$ for the hand crafted features and the CNN, respectively), with a noticeable improvement at the clinically relevant FPI=1 value (0.2885 versus 0.202 and 0.2 for combined, CNN, and hand crafted, respectively). Moreover, the differences between the manually tuned approach and the CNN probabilities are not statistically significant. When evaluating the ability of the system to detect HEs on the DIARETDB1 test set, it is possible to see that the per lesion sensitivities are higher than those reported for MA detection. Furthermore, the hand crafted features are able to achieve better per lesion sensitivity values than the combined approach ($p < 5 \times 10^{-5}$) for this specific task. At the clinically relevant FPI value of 1, however, the combined approach reports a slightly higher per lesion sensitivity compared to the manually engineered descriptors (0.4907 versus 0.4724).

Experiment 2 was carried out on e-ophtha to estimate the ability of our method to segment MAs and smaller

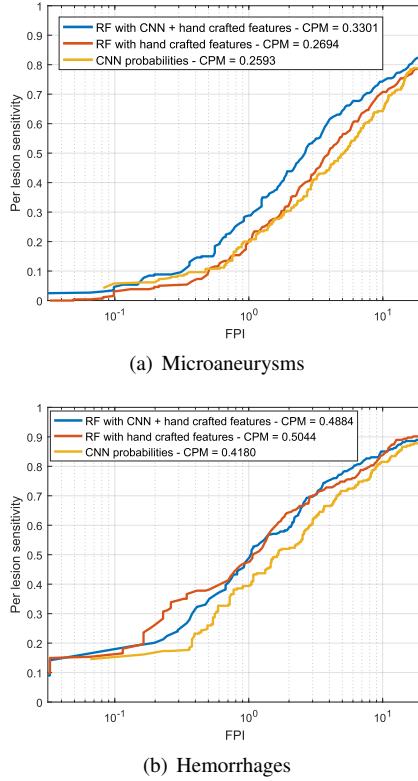


Figure 8: Per lesion evaluation for each lesion type in the DIARETDB1 test set: (a) Microaneurysms, (b) Hemorrhages.

HEs simultaneously. In this case, a combination of both the DIARETDB1 (MA labels with a level of agreement $\geq 75\%$) and ROC training sets was used for learning, as we observed that few MAs (only 182 for the entire DIARETDB1 set) are retrieved at $\geq 75\%$ agreement. To the best of our knowledge, the only method evaluated on e-ophtha is by Wu et al. (2017), although their analysis is performed on a subsample of 74 images with lesions instead of the full data set. By contrast, we used a more challenging evaluation comprising the entire e-ophtha set, including also the 233 images with no visible sign of DR. Figure 9 presents the FROC curves obtained using each approach. As in the previous experiment, the Wilcoxon signed rank tests showed a statistical significant improvement in the per lesion sensitivity values using the hybrid vector of both deep learned features and domain knowledge with respect to the CNN probabilities and the hand crafted features ($p < 2 \times 10^{-18}$ and $p < 2 \times 10^{-9}$, respectively).

Table 5 summarizes the CPM values obtained for each experiment and each feature combination, and also using each of the two recently published state-of-the-art methods. Per lesion sensitivities at FPI= 1, which is

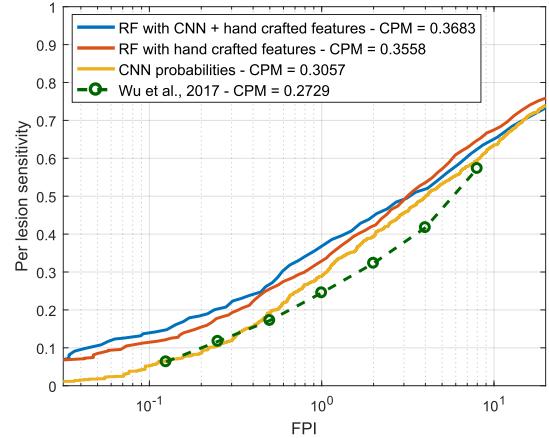


Figure 9: Per lesion evaluation in Experiment 2. FROC curve and CPM values obtained on e-ophtha.

Table 5: CPM values and per lesion sensitivities at FPI= 1 for Experiments 1 (red lesions with multiple sizes) and 2 (small red lesions) (Table 4).

Method	Experiment 1		Experiment 2	
	CPM	Se	CPM	Se
Seoud et al. (2016)	0.3540	0.3462	-	-
Wu et al. (2017)	-	-	0.2729	0.2450
CNN probabilities	0.3756	0.3621	0.3057	0.2894
RF with HCF	0.4517	0.4601	0.3558	0.3291
RF with CNN + HCF	0.4874	0.4883	0.3683	0.3680

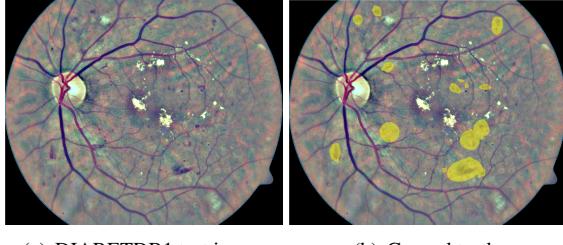
considered a clinically relevant number of false positives (Niemeijer et al., 2010) are also provided.

Finally, qualitative results for a randomly selected image in the DIARETDB1 test set are depicted in Figure 10. Green circles are detected lesions according to the ground truth labeling provided in the data set, while yellow circles correspond to lesions detected by our method but that are not labeled in the ground truth. Finally, red circles surround the lesions that were manually annotated as true lesions but were ignored by the method. Qualitatively, many of the yellow circles appear to be microaneurysms or hemorrhages that were not detected during manual labeling due to their subtle appearance in the original RGB image.

4.2. Per image evaluation

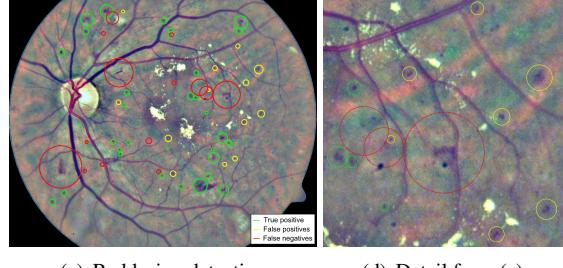
Two different experiments were conducted on MESSIDOR to estimate the performance of our method on a per image basis, one focused on detecting patients with DR, and a second based on detecting those need for immediate referral to a specialist. In both cases, we used the model learned from Experiment 1.

Figure 11(a) illustrates the ROC curves for DR screening on MESSIDOR, obtained using our hybrid



(a) DIARETDB1 test image

(b) Ground truth.



(c) Red lesion detection

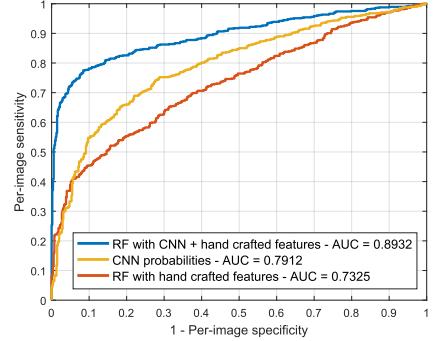
(d) Detail from (c)

Figure 10: Qualitative results. (a) image015 from the DIARETDB1 test set. (b) Ground truth labeling at a $> 25\%$ level agreement. (c) Red lesion detections obtained by thresholding the probabilities at 0.644, which corresponds to an average FPI value of 1. (d) Detail from (c) showing lesions unlabeled on the ground truth but identified by our method.

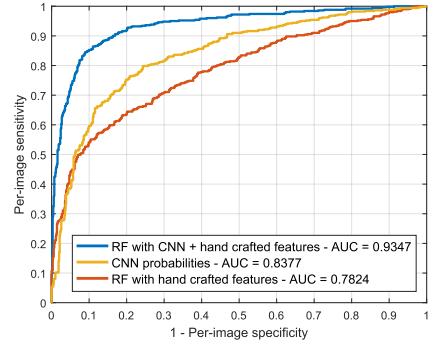
representation and each of the approaches separately. CNN results were obtained using the network as a classifier. A series of Mann-Whitney U tests ($\alpha = 0.05$) were performed to study the statistical significance of the differences in the AUC values (Hanley and McNeil, 1982). CNN features ($AUC = 0.7912$) perform significantly better ($p < 1 \times 10^{-3}$) than hand crafted features ($AUC = 0.7325$) for this specific task, and the combination of both sources of information results in a substantially higher AUC value of 0.8932 ($p < 1 \times 10^{-6}$). Figure 11(b) shows analogous behavior for detecting patients that need referral, with the CNN performing better than the hand crafted features ($p < 2 \times 10^{-3}$), and the combined approach outperforms both individual techniques ($p < 1 \times 10^{-6}$).

Our combined approach shows an analogous behavior when evaluating on e-ophtha for DR screening, as illustrated in Figure 12. Our combined approach retrieved a significantly higher AUC value (0.9031) than the one reported by the CNN ($AUC = 0.8374$, $p < 5 \times 10^{-3}$) and the RF classifier trained with hand crafted features ($AUC = 0.8812$). Hand crafted features perform better than the CNN for screening in this data set, although the difference is not statistically significant according to the Mann-Whitney U test.

A comparison with respect to other state of the art



(a) Performance of DR screening



(b) Performance of detecting patients that need referral

Figure 11: Per image evaluation. ROC curves for (a) DR screening (R1 vs. R2, R3 and R4) and (b) need for referral (R1 and R2 vs. R3 and R4) on the MESSIDOR data set.

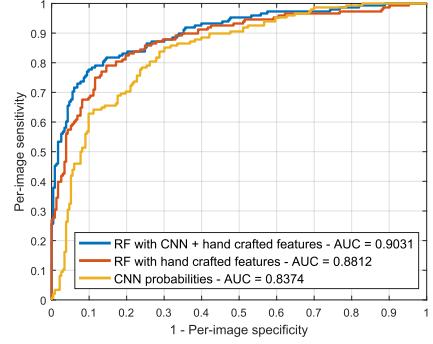


Figure 12: Per image evaluation on e-ophtha. ROC curve for DR screening.

strategies is presented in Table 6. The performance obtained by two human experts, as reported by Sánchez et al. (2011), is also included in the table. The results of the baseline method by Seoud et al. (2016) were obtained using DIARETDB1 as a training set. The other methods included are either based only on red lesion detection or complemented by other features such as im-

Table 6: Comparison of DR screening and need of referral performance on the MESSIDOR data set. Se values correspond to those obtained at a $S_p = 50\%$.

Method	Screening		Need for referral	
	AUC	Se	AUC	Se
Expert A (Sánchez et al., 2011)	0.9220	0.9450	0.9400	0.9820
Expert B (Sánchez et al., 2011)	0.8650	0.9120	0.9200	0.9760
Antal and Hajdu (2012)	0.8750	-	-	-
Costa et al. (2016)	0.8700	-	-	-
Giancardo et al. (2013)	0.8540	-	-	-
Nandy et al. (2016)	-	-	0.9210	-
Pires et al. (2015)	-	-	0.8630	-
Sánchez et al. (2011)	0.8760	0.9220	0.9100	0.9440
Seoud et al. (2016) (DIARETDB1)	0.844	-	-	-
Vo and Verma (2016) (I)	0.8620	-	0.8910	-
Vo and Verma (2016) (II)	0.8700	-	0.8870	-
HCF	0.7325	0.7645	0.7824	0.8283
CNN	0.7912	0.8471	0.8377	0.9102
HCF + CNN	0.8932	0.9109	0.9347	0.9721

age quality assessment or the detection of exudates and neovascularizations.

4.3. Feature assessment

In order to assess the visual appearance of the deep learned features, a graphical representation of the 32 filters of size $5 \times 5 \times 3$ learned on the first layer of the CNN is presented in Figure 13. These representations allow to verify which types of high level characteristics are detected by the first layer of the network (Zeiler and Fergus, 2014). Thus, they are suitable to confirm if the network was trained for long enough, as well-trained CNNs usually display smooth filters without noisy patterns, as in this case. From Figure 13(a) it is possible to see that filters learned in Experiment 1 are mostly descriptors of the color properties of the lesions. This setting is in line with the fact that the training set used in this case contains not only small MAs but also medium size HEs, which can be more easily described in terms of their internal color homogeneity rather than their edges, which significantly varies from one to another. Other filters are able to capture purple, ellipsoidal structures corresponding to true lesions like those illustrated in Figure 5(b). This last type of filter is more common in the first layer of the CNN learned in Experiment 2 (Figure 13(b)), which might be associated with the smaller true positive structures observed in the training set built with ROC and DIARETDB1 MAs.

CNNs can be interpreted as models that transform the input images into a feature representation in which classes can be separated by the linear model in the last layer. The topology of such a space will depend on the ability of the deep learning features to characterize the inputs. Hence, if features are sufficiently good to differentiate each type of input, at least two well separated

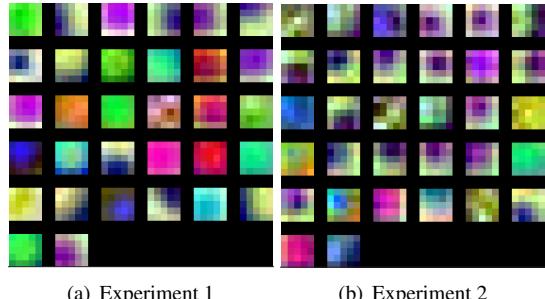


Figure 13: Learned filters on the first layer of our CNN, as obtained for each experiment in Table 4:(a) Experiment 1 (training on DIARETDB1 training set), (b) Experiment 2 (training on DIARETDB1 and ROC training sets).

regions would be visually identified. Due to the high dimensionality of the feature space, a method is needed to embed multidimensional vectors in a 2D space, while preserving the pairwise distances of the points. The t -distributed stochastic neighbor embedding (t -SNE) was recently introduced for this purpose (Van Der Maaten, 2014). We followed this approach to study the complementarity of each characterization method, and to qualitatively assess how their integration contribute to improve their original discrimination ability. Figure 14 presents the t -SNE mappings of the DIARETDB1 test samples for each characterization approach and for our combined feature vector. The CNN descriptors corresponds to those learned in Experiment 1. The figure also includes a visual representation of the organization of the patches in the embedding space. In general, it is possible to see that the ensemble approach groups the majority of the true positive candidates within a single neighboring area. By contrast, the individual characterization strategies are not able to achieve a single cluster but rather obtain two—in the case of the deep learned features—or more—using the hand crafted features.

Detailed regions of the embeddings are depicted in Figure 15. This allows better visualization of particular scenarios such as the patches around the true red lesions, the false positive candidates located in the vascular structures, the artifacts due to speckles of dirt in the lens—which are typical of the images in DIARETDB1—and the false detections within the optic disc. In general, it is possible to observe that CNN features are able to better characterize the orientation and the visual appearance of the true lesion candidates, while the hand crafted features can detect the less obvious lesions under low contrast conditions. The ability of the CNN features to discriminate orientations are more evident when dealing with vascular structures. The hand crafted ap-

proach, by contrast, is only able to capture the overall size of the vessels and their intensity properties. When combining both strategies, the main advantages of each of them are maintained. The robustness against artifacts is evident for both the deep learning based and the hand crafted features, as these false positive candidates are grouped together into separate clusters from the true lesions. A similar behavior is observed when dealing with false candidates within the optic disc area.

5. Discussion

In general, the integration of both the deep learned and the hand crafted features significantly improved results compared to using either approach separately. In a per lesion evaluation, the combined approach achieved a consistently higher CPM value both in the e-ophtha and DIARETDB1 test sets, and also a higher per lesion sensitivity for FPI=1, which corresponds to a clinically relevant number of false positives (Niemeijer et al., 2010). These values are also higher than those obtained by two recently published baseline methods that were evaluated on the same data set. A similar behavior is observed when evaluating the method on a per image basis. The combined approach improved the performance obtained by each characterization approach separately, meaning that the integration of both sources of information obtains a better characterization of the lesion candidates and, consequently, a more accurate detection of the individual lesions. This is supported by the extensive analysis presented in Section 4.3. Despite the fact that sufficiently deep CNNs are known to be able to learn any function of arbitrary complexity, the lack of data with lesion-level annotations does not allow our network to identify the same properties that the hand crafted features do. Nevertheless, in the analysis of the t-SNE mapping presented for each method (Figures 14 and 15) it is possible to see that the CNN has the ability to characterize fine-grained details such as the orientation of the lesion that are ignored by the manually selected descriptors. On the other hand, the hand crafted features have the ability to discriminate other low contrast lesions (Figure 15), specially hemorrhages (Figure 8(b)). As a result, the ensemble approach is able to outperform each individual alternative, improving performance for detecting both MA and HE simultaneously. Due to the high cost of accurately annotating small lesions, we hypothesize that this observation will continue to stand in the near future.

Results on the per image evaluation also showed that the proposed strategy is able to achieve higher AUC values than other approaches for DR screening and need-

for-referral detection. Moreover, the methods included in Table 6 are based not only on red lesion detector (Antal and Hajdu, 2012, Giancardo et al., 2013, Seoud et al., 2016) but also on additional features such as the assessment of the image quality (Sánchez et al., 2011) and/or the presence of other pathological structures such as exudates and neovascularizations (Sánchez et al., 2011, Pires et al., 2015, Costa et al., 2016). Compared with respect to all these approaches, our method achieved a higher AUC value. Furthermore, it performed better than the DR grading method by Vo and Verma (2016), which uses fine tuned CNNs trained on a data set with 50.000 images with image level annotations. An almost equal performance was obtained for DR screening compared with the recently published method by Quellec et al. (2016), which reported an AUC= 0.893 in the MESSIDOR data set. However, such an approach uses multiple images per patient, contextual information and clinical records to learn diagnostic rules from a data set with 12.000 examinations. Our method is able to achieve a slightly higher AUC value without including any additional clinical information. Furthermore, a competitive *Se* value was obtained in comparison with Expert B (Sánchez et al., 2011), indicating that this approach can match the ability of a human observer for DR screening and detecting patients that need referral. Thus, our automated red lesion detection system could be integrated in a more general DR screening platform to improve the ability to detect DR patients. In particular, methods such as the one proposed by Gulshan et al. (2016), which is able to identify moderate/worse and severe DR cases, can be aided by the incorporation of a red lesion detection so that the early stages of the disease can also be determined. Moreover, a reliable DR likelihood can be complemented by an indication of the abnormal areas, allowing physicians to better identify the clinical signs of the disease and to have more comprehensive feedback from the system. Furthermore, incorporating other modules for detecting other pathological structures can eventually improve the reported performance.

It is also important to underline that all the stages in the proposed method have parameters that are automatically adjusted to each image resolution. Their values, which are reported in Section 3.2, were empirically selected using different data than that used for evaluation, and were proportionally scaled in the subsequent experiments to compensate resolution changes. This simple approach provides an approximate scale invariance that is valuable to facilitate the adaptability of the method to be applied on images obtained using different fundus cameras.

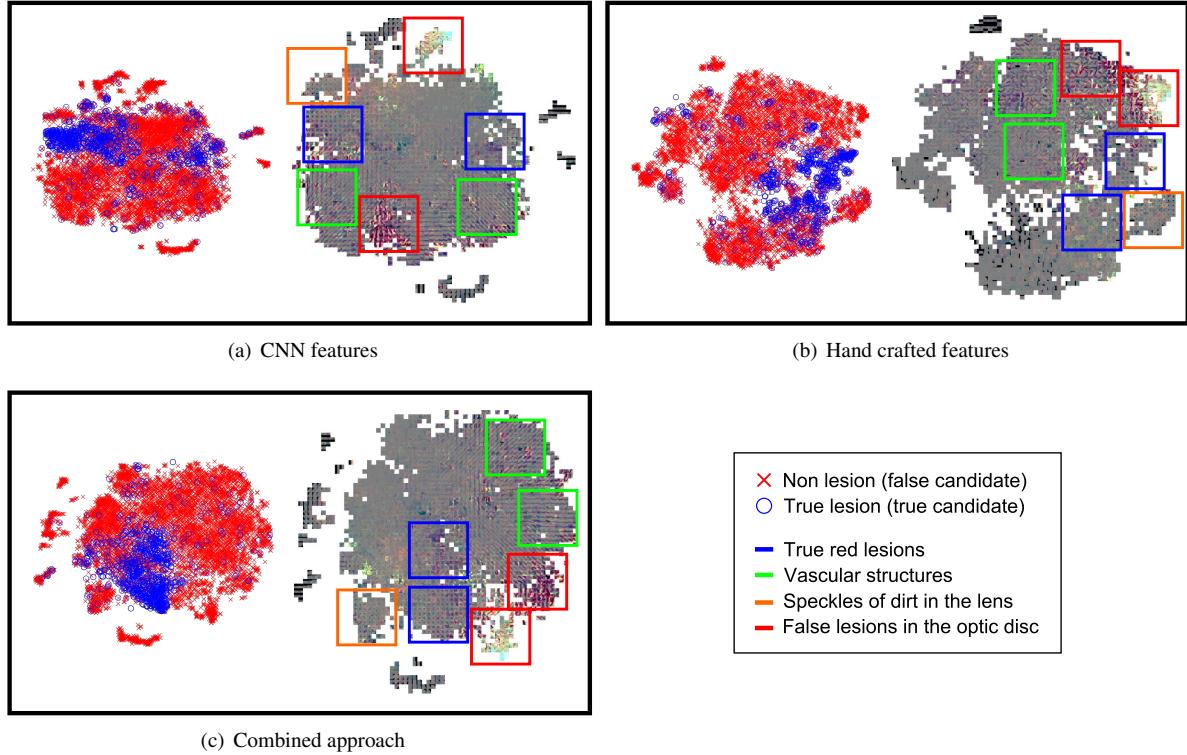


Figure 14: The t -SNE visualization of the patches from DIARETDB1 test set as mapped using (a) the deep learned features, (b) the hand crafted features and (c) our hybrid feature vector. Left side: color coded labels for each test sample. Right side: patches around the candidates, as visualized using the t -SNE mappings. Details for different types of lesion candidates are shown in Figure 15.

When analyzing each individual characterization approach, it is possible to see in Experiment 1 that both the RF trained with hand crafted features and the CNN achieved higher per lesion sensitivities than the method by Seoud et al. (2016) ($p < 2 \times 10^{-18}$ and $p < 2 \times 10^{-4}$, respectively). This is likely due to the fact that our method for extracting candidates differs from the one used by the alternative approach. Moreover, Seoud et al. (2016) eliminate the lesion candidates occurring within an estimated area around the optic disc center, which is determined using an automated approach. As a consequence, if the diameter of the optic disc is accidentally overestimated by such a method, candidates within valid regions will be suppressed and it will not be possible to recover them afterwards during the classification stage. As seen in Figures 14 and 15, our combined approach is able to discriminate the candidates within the optic disc area and the vascular structures. Hence, instead of using a rigid elimination step based on optic disc segmentation, we let the classifier to decide whether a candidate is actually a true positive or a false positive occurring on an anatomical region. This

approach increases the maximum achievable per lesion sensitivity on each image, allowing to train our classifier with a larger amount of false positive lesions and to get a higher sensitivity in test time. A similar observation can be made from the results of Experiment 2, in which the hand crafted features and the deep learning based approach reported higher per lesion sensitivities than those reported by Wu et al. (2017). It must be underlined, also, that the Wu et al. (2017) method was trained on the first half of the images with pathologies on e-ophtha and evaluated on the second half, rather than trained on a separate data set and evaluated on the complete set, as in our case. Moreover, it is worth noting that the images of the healthy patients were also included during evaluation to get a more accurate estimation of its actual performance on a real, clinical scenario.

On a per image basis, it is possible to see that the individual approaches trained in Experiment 1 are not able to achieve AUC values higher than those reported by Seoud et al. (2016) (Table 6). This is likely due to the fact that, as indicated by the authors, their method is more accurate for detecting blot HEs and MAs than

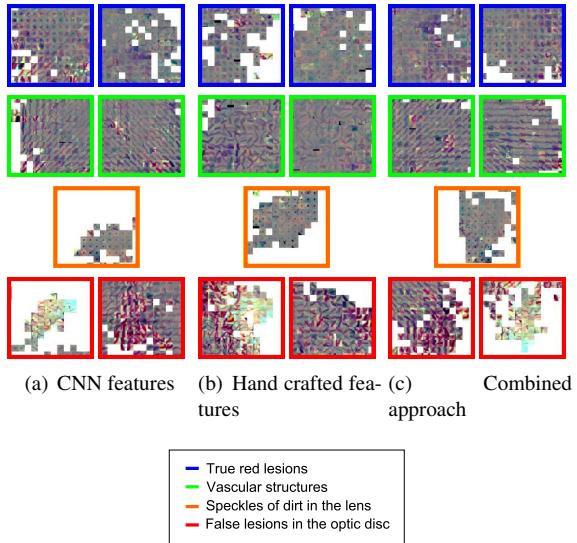


Figure 15: Details from the t-SNE visualization in Figure 14 for different types of red lesion candidates (true lesions, vascular structures, speckles of dirt in the lens and false detections in vessel curves in the optic disc): (a) Deep learned features, (b) Hand crafted features, (c) Combined approach.

HEs with other shapes. The images in MESSIDOR were originally graded as R0 and R1 taking into account the number of MAs (Table 3) (Decencière et al., 2014). Hence, being more accurate in the detection of MAs will result in a better ability to distinguish much earlier stages. When individually using the hand crafted features or the CNN, both methods are less precise for detecting MAs but better for discriminating other HEs. This argument is supported by results presented in Figure 8, in which it is possible to see that the per lesion sensitivity values obtained for MA detection are lower than those reported for HEs. Moreover, it was observed that the CNN performed equally or better than the RF trained with manually engineered features on the low FPI regime for MA detection. This explains the behavior observed in Figure 11, where the CNN probabilities achieved a higher AUC value for DR screening and need for referral detection. Nevertheless, the combination of both approaches with the RF classifier consistently improved their individual performance, achieving a much better characterization of the MAs (as observed in the improvements reported in Figure 8(a)) and, consequently, a better discrimination of the DR patients.

6. Conclusions

We have proposed a novel method for red lesion detection in fundus images based on a hybrid vector of

both CNN-based and hand crafted features. A CNN is trained using patches around lesion candidates to learn features automatically, and those descriptors are complemented using domain knowledge to improve their discrimination ability. Results on benchmark data sets empirically demonstrated that the resulting system achieves a new state-of-the-art in this domain, and that combining both sources of information provides statistically significant improvements compared to using each of them separately. A similar behavior is observed when evaluating our screening system both for DR and need-for-referral detection, reporting higher AUC values than those obtained by other existing approaches based not only on red lesion detection but also on analyzing other pathologies such as bright lesions or neovascularizations, or even learning classifiers using additional clinical information. Considering the high cost of manually labeling fundus photographs at a lesion level, our method represents a robust alternative to improve performance of other deep learning based approaches. An open source implementation and the detection masks are made publicly available at <https://github.com/ignaciotorlano/red-lesion-detection>.

Acknowledgements

This work is supported by Internal Funds KU Leuven, FP7-MC-CIG 334380, an Nvidia Hardware Grant and ANPCyT PICT 2014-1730, PICT Start-up 2015-0006 and PICT 2016-0116. J.I.O. is funded by a doctoral scholarship granted by CONICET. We would also like to thank Dr. Seoud for her assistance.

Conflicts of interest

The authors declare that there are no conflicts of interest in this work.

References

References

- Michael D Abràmoff, Mona K Garvin, and Milan Sonka. Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering*, 3:169–208, 2010.

Michael David Abràmoff and Meindert Niemeijer. Mass screening of diabetic retinopathy using automated methods. In *Teleophthalmology in Preventive Medicine*, pages 41–50. Springer, 2015.

Roberto Annunziata and Emanuele Trucco. Accelerating convolutional sparse coding for curvilinear structures segmentation by refining SCIRD-TS filter banks. *IEEE Transactions on Medical Imaging*, 35(11):2381–2392, 2016.

- Balint Antal and Andras Hajdu. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE Transactions on Biomedical Engineering*, 59(6):1720–1726, 2012.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- João Costa, Inês Sousa, and Filipe Soares. Smartphone-based decision support system for elimination of pathology-free images in diabetic retinopathy screening. In *International Conference on IoT Technologies for HealthCare*, pages 83–88. Springer, 2016.
- Etienne Decencière, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, et al. Teleophtha: Machine learning and image processing methods for teleophthalmology. *IRBM*, 34(2):196–203, 2013.
- Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Laÿ, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the MESSIDOR database. *Image Analysis and Stereology*, pages 231–234, 2014.
- Luca Giancardo, Thomas P Karnowski, Kenneth W Tobin, Fabrice Meriaudeau, and Edward Chaum. Validation of microaneurysm-based diabetic retinopathy screening across retina fundus datasets. In *IEEE 26th international symposium on Computer-based medical systems (CBMS)*, pages 125–130. IEEE, 2013.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.
- Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.
- James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.
- Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raija Voutilainen, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. The DIARETDB1 diabetic retinopathy database and evaluation protocol. In *11th Conference on Medical Image Understanding and Analysis*, 2007.
- Jessica J Kovarik, Andrew W Eller, Lauren A Willard, Jiaxi Ding, Jann M Johnston, and Evan L Waxman. Prevalence of undiagnosed diabetic retinopathy among inpatients with diabetes: the diabetic retinopathy inpatient study (DRIPS). *BMJ Open Diabetes Research & Care*, 4(1):e000164, 2016.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- Lucas Lo Vercio, Mariana del Fresno, and Ignacio Larrañide. Detection of morphological structures for vessel wall segmentation in ivus using random forests. In *12th International Symposium on Medical Information Processing and Analysis*, page 1016012. International Society for Optics and Photonics, 2017.
- Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Deep retinal image understanding. In *Medical Image Computing and Computer-Assisted Intervention*, pages 140–148. Springer, 2016.
- Muthu Rama Krishnan Mookiah et al. Computer-aided diagnosis of diabetic retinopathy: A review. *Computers in Biology and Medicine*, 43(12):2136–2155, 2013.
- Jay Nandy, Wynne Hsu, and Mong Li Lee. An incremental feature extraction framework for referable diabetic retinopathy detection. In *IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 908–912. IEEE, 2016.
- Meindert Niemeijer, Joes Staal, Bram van Ginneken, Marco Loog, and Michael D Abramoff. Comparative study of retinal vessel segmentation methods on a new publicly available database. In *Medical Imaging*, pages 648–656. International Society for Optics and Photonics, 2004.
- Meindert Niemeijer, Bram Van Ginneken, Joes Staal, Maria SA Suttorp-Schulten, and Michael D Abràmoff. Automatic detection of red lesions in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 24(5):584–592, 2005.
- Meindert Niemeijer et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 29(1):185–195, 2010.
- José Ignacio Orlando and Matthew B. Blaschko. Learning fully-connected CRFs for blood vessel segmentation in retinal images. In *Medical Image Computing and Computer-Assisted Intervention*, volume 8673 of *Lecture Notes in Computer Science*, pages 634–641. Springer, 2014.
- José Ignacio Orlando, Elena Prokofyeva, and Matthew B. Blaschko. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Transactions on Biomedical Engineering*, 64(1):16–27, Jan 2017a.
- José Ignacio Orlando, Elena Prokofyeva, Mariana del Fresno, and Matthew Blaschko. Convolutional neural network transfer for automated glaucoma identification. In *12th International Symposium on Medical Information Processing and Analysis*, volume 10160 of *Proc. SPIE*, page 101600U. International Society for Optics and Photonics, 2017b.
- Ramon Pires, Herbert F Jelinek, Jacques Wainer, Siome Goldenstein, Eduardo Valle, and Anderson Rocha. Assessing the need for referral in automatic diabetic retinopathy detection. *IEEE Transactions on Biomedical Engineering*, 60(12):3391–3398, 2013.
- Ramon Pires, Sandra Avila, Herbert Jelinek, Jacques Wainer, Eduardo Valle, and Anderson Rocha. Beyond lesion-based diabetic retinopathy: a direct approach for referral. *IEEE Journal of Biomedical and Health Informatics*, 2015.
- Elena Prokofyeva and Eberhart Zrenner. Epidemiology of major eye diseases leading to blindness in Europe: A literature review. *Ophthalmic Research*, 47(4):171–188, 2012.
- Gwenolé Quellec, Mathieu Lamard, Ali Erginay, Agnès Chabouis, Pascale Massin, Béatrice Cochener, and Guy Cazuguel. Automatic detection of referral patients due to retinal pathologies through data mining. *Medical image analysis*, 29:47–64, 2016.
- Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519. IEEE, 2014.
- Clara I Sánchez, Meindert Niemeijer, Alina V Dumitrescu, Maria SA Suttorp-Schulten, Michael D Abràmoff, and Bram van Ginneken. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Investigative Ophthalmology & Visual Science*, 52(7):4866–4871, 2011.
- Lama Seoud, Thomas Hurtut, Jihed Chelbi, Farida Cheriet, and JM Pierre Langlois. Red lesion detection using dynamic shape features for diabetic retinopathy screening. *IEEE Transactions on Medical Imaging*, 35(4):1116–1126, 2016.
- Joao VB Soares et al. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, 25(9), 2006.
- Emanuele Trucco, Alfredo Ruggeri, Thomas Karnowski, Luca Giancardo, Edward Chaum, Jean Pierre Hubschman, Bashir Al-Diri, Carol Y Cheung, Damon Wong, Michael Abramoff, et al. Validating retinal fundus image analysis algorithms: Issues and a proposal. *Investigative Ophthalmology & Visual Science*, 54(5):3546–3559, 2013.
- Laurens Van Der Maaten. Accelerating t-SNE using tree-based algo-

- rithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.
- Mark JJP van Grinsven, Bram van Ginneken, Carel B Hoyng, Thomas Theelen, and Clara I Sánchez. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, 2016.
- Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.
- Rahul Venkataramani, Sheshadri Thiruvekadam, Pavan Annangi, Narayanan Babu, and Vivek Vaidya. Understanding the mechanisms of deep transfer learning for medical images. In *LABELS and DLMIA Workshops, at Medical Image Computing and Computer Assisted Intervention*, volume 10008, page 188. Springer, 2016.
- Holly H Vo and Abhishek Verma. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 209–215. IEEE, 2016.
- Thomas Walter, Pascale Massin, Ali Erginay, Richard Ordonez, Clotilde Jeulin, and Jean-Claude Klein. Automatic detection of microaneurysms in color fundus images. *Medical Image Analysis*, 11(6):555–566, 2007.
- Haibo Wang, Angel Cruz-Roa, Ajay Basavanhally, Hannah Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003–034003, 2014.
- Bo Wu, Weifang Zhu, Fei Shi, Shuxia Zhu, and Xinjian Chen. Automatic detection of microaneurysms in retinal fundus images. *Computerized Medical Imaging and Graphics*, 55:106–112, 2017.
- Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- Yefeng Zheng et al. 3D deep learning for efficient and robust landmark detection in volumetric data. In *Medical Image Computing and Computer Assisted Intervention*, pages 565–572, 2015.