IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 13, 2017, accepted March 5, 2018, date of publication March 15, 2018, date of current version April 23, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2816003

A Weakly-Supervised Framework for Interpretable Diabetic Retinopathy Detection on Retinal Images

PEDRO COSTA[1], ADRIAN GALDRAN[1], ASIM SMAILAGIC[2], (Fellow, IEEE), AND AURÉLIO CAMPILHO[1,3], (Senior Member, IEEE)

[1]Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
[2]Institute for Complex Engineered Systems, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3]Faculdade de Engenharia, Universidade de Porto, 4200-464 Porto, Portugal, Portugal

Corresponding author: Pedro Costa (ei10011@fe.up.pt)

This work was supported in part by the European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation–COMPETE 2020 Programme and in part by the National Funds through the Fundação para a Cîencia e a Tecnologia within under Project CMUPERI/TIC/0028/2014.

ABSTRACT Diabetic retinopathy (DR) detection is a critical retinal image analysis task in the context of early blindness prevention. Unfortunately, in order to train a model to accurately detect DR based on the presence of different retinal lesions, typically a dataset with medical expert's annotations at the pixel level is needed. In this paper, a new methodology based on the multiple instance learning (MIL) framework is developed in order to overcome this necessity by leveraging the implicit information present on annotations made at the image level. Contrary to previous MIL-based DR detection systems, the main contribution of the proposed technique is the joint optimization of the instance encoding and the image classification stages. In this way, more useful mid-level representations of pathological images can be obtained. The explainability of the model decisions is further enhanced by means of a new loss function enforcing appropriate instance and mid-level representations. The proposed technique achieves comparable or better results than other recently proposed methods, with 90% area under the receiver operating characteristic curve (AUC) on Messidor, 93% AUC on DR1, and 96% AUC on DR2, while improving the interpretability of the produced decisions.

INDEX TERMS Multiple instance learning, diabetic retinopathy detection, bag of visual words, retinal image analysis.

I. INTRODUCTION

The retina is a well-known source of biomarkers that enable the early identification of several human disorders, such as hypertension, heart diseases, or Diabetic Retinopathy (DR), among others. DR is known to be the leading cause of preventable blindness, affecting more than 415 million people worldwide [1]. Fortunately, DR can be detected at its early stages by expert ophthalmologists through routine analysis of the eye fundus [2]. Timely DR detection can lead to the administration of preventive treatments and efficient therapies to avoid vision impairment and further consequences.

To provide early disease diagnosis and appropriate eye care, large-scale global screening programs have been implemented by hospitals and local authorities [3], [4] with great success. In the context of such programs, patients are called to clinical settings in order to acquire eye color images with a retinal fundus camera. These images are then submitted to specialists, who look for visual signs of the presence of lesions and perform diagnosis based on them. A sample of these potential signs of disease is shown in Fig. 1.

Unfortunately, more than 83% of undiagnosed DR patients are located in underdeveloped areas, where there is a lack of specialists to attend large masses of population, blocking the appropriate implementation of screening programs [5]. For this reason, Computer-Aided Diagnosis (CAD) systems capable of detecting signs of DR from standard retinal fundus images are becoming highly relevant in recent years [6]. An effective automatic DR detection system can substantially reduce the workload experimented by ophthalmologists in the context of large-scale screening programs, having a large positive impact on population healthcare [7], [8].

However, the design of a CAD system to support expert's decisions in an appropriate manner must take into account several desirable properties. First, the amount of annotated

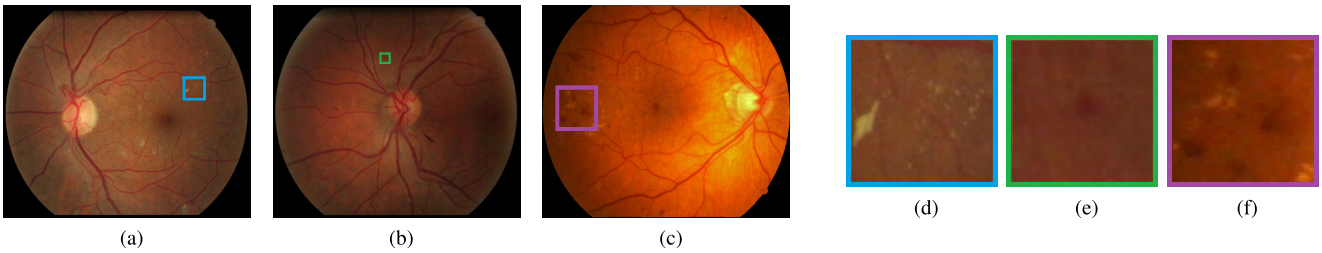


FIGURE 1. (a–c) A Retinal Images showing signs of DR (b–d) Lesions associated to DR (d) Exudates (e) Microaneurysms (f) Hemorrhages.

training data must be medium-to-moderate, since manually labeling each pixel on image regions containing lesions can be a time-consuming and error-prone process. Ideally, a CAD system must be able to learn to detect disease signs from a set of images labeled with a single number indicating the presence or not of DR. Adding a single tag on a retinal image is a much easier and faster task for human experts to accomplish. Moreover, there is already a large quantity of visual data stored at hospitals that has been annotated with this kind of labels, representing an immense source of training data. We refer to this type of annotations as *weakly labeled* data. Second, any CAD system supporting ophthalmologists' clinical decisions must work reliably and in an interpretable manner, in order to fit regular clinical work-flows.

To address both of these challenges, the main contribution of this paper consists of a new technique for DR detection capable of learning from a set of weakly labeled images and performing interpretable diagnose prediction. The proposed technique allows to train a DR detection CAD system at an image level using implicit local information, *e.g.* deciding if the image is healthy or not based on lesions present in certain image regions, even if their location is unknown. This is achieved by means of a novel Multiple Instance Learning (MIL) technique that improves upon previously proposed MIL approaches by jointly learning to encode and classify visual information coming from localized areas of the image. As a second contribution, the interpretability of the proposed model is enhanced by means of a constraint imposed on the learned representations, which forces them to remain sparse in case of healthy images while becoming dense whenever the image contains DR signs. Hence, the proposed system learns from weakly labeled data without requiring strong manual annotations to be trained, but it can still pinpoint the regions on the image that triggered the diagnosis decision, resulting in a highly interpretable CAD system. Comprehensive performance evaluation on several publicly available datasets favor the proposed technique, demonstrating that it competes well with other recent approaches while providing an increased interpretability outcome. The method presented in this paper is a substantial extension of the conference publication [9].

II. RELATED WORK

A. DR DETECTION ON RETINAL IMAGES

DR detection on images of the eye fundus is usually achieved by first locating specific disease signs and lesions in the retinal images. This is typically accomplished based on a conventional machine learning pipeline for detecting objects of interest within images, *i.e.* given a dataset of image regions containing manually delineated lesions:

1) **Lesion Description**: Visual features are extracted to characterize each type of lesion modeling their geometric, textural and color appearance.
2) **Classifier Training**: A classifier is trained to distinguish lesions based on the extracted features.
3) **Lesion Candidate Extraction**: Given a new image, candidate regions are extracted from it, and the retrieved candidates are described with those same features.
4) **Lesion Candidate Classification**: These descriptors are inputted to the classifier, which decides if the candidate is a lesion or not (false positive removal) and/or the most likely type of lesion.

Usually, these techniques are specifically designed to deal with a single type of lesion, *e.g.* microaneurysms [10] or hard exudates detection [11]. More generally, red lesions [12], [13] or bright lesions [14], [15] can be detected. After lesion detection has been performed, DR detection and grading can be realized. Several papers have thus proposed DR detection techniques consisting of combining distinct lesion detection techniques to extract all relevant anomalies, and then merge the results into a single outcome indicating the presence or severity of DR [16], [17].

The main drawback of the above approach is that it requires an image database that has been previously annotated by a specialist at the lesion level. The lesion borders need to be marked pixel by pixel or with specialized visual tracing tools. This represents a tedious and time-consuming process. Yet another relevant limitation of lesion-detection based DR detection is the necessity of complex and often error-prone pre-processing techniques. For instance, the optic disc typically needs to be located and removed in order to avoid the generation of candidate regions that can be confused as bright lesions [16].

To avoid the need for manually segmented training examples, this work differs from multi-lesion detection approaches by framing the problem within the MIL paradigm. MIL allows to build a weakly-supervised learning system, where only a single indicator is required for a given image but predictions are formulated based implicitly on region-level characteristics. Below we provide a brief theoretical introduction to the MIL framework.

B. MULTIPLE-INSTANCE LEARNING

The MIL framework for binary classification problems considers two main entities, called *bags* and *instances*. In this setting, a bag is composed of an undetermined number of instances. While the goal of a MIL algorithm is to classify bags into positive or negative, it is assumed that instances carry useful information regarding bags containing them. However, the only available ground-truth in MIL is associated to bags. The goal in this approach then becomes to model the implicit relationship between instances and their corresponding bags. A typical example of such a relationship would be: if a bag contains at least a positive instance, regardless of how many negative instances it may contain, it should be declared as positive, whereas a bag should be predicted as negative if it contains only negative instances.

There are two main MIL algorithms categories, namely instance-level techniques (ILT) and bag-level techniques (BLT). In ILT, a classifier is trained to classify instances, and instance-level predictions are aggregated to build a bag-level prediction. Examples of this approach are mi-SVM [18], or MIL-Boost [19]. The way in which instance-level predictions are combined will model the instance/bag relationship. Following the above example, instance-level predictions can be aggregated with a max-rule: the bag-level prediction is given by the top positive instance contained on it. The main disadvantage of this approach is that not always a single instance should condition the bag-level prediction, as the final label may be influenced by a larger set of instances.

BLT differ from ILT in that the classifier is not trained to classify instances, but rather it learns to classify bags directly. The main difference lies in the moment the instance-level information is aggregated. While ILT combine instance-level predictions, in BLT a bag-level representation is built from a combination of instance-level representations, and the classifier is trained on this combined representation [20].

The main problem of MIL-BLT is that the amount of instances within a bag is not known a priori, which gives rise to bag representations of varying dimension. To overcome this obstacle, typically all the instance-level representations of different bags are mapped into a common space. This is achieved with embedding functions followed by pooling operations, and the goal becomes finding a representation space as discriminative as possible.

One particularly interesting MIL-based image classification model is the Bag of Visual Words (BoVW), introduced in [21] for video retrieval. In BoVW, an image (bag) is decomposed into a set of local low-level visual descriptors (instances). These are then mapped onto a common representation, defined by a visual dictionary.

In BoVW techniques, the way the visual dictionary is learned is a critical part of the method. The most popular approaches involve applying unsupervised techniques, such as k-means clustering, on features extracted from a group of images. In this case, the resulting k centroids conform the visual words composing the dictionary. Once the instances associated to an image have been encoded using the visual dictionary, they are combined together via a pooling operation, resulting in a feature vector that is supplied to a standard classifier.

In summary, BoVW is characterized by two separate stages. The first one extracts features from all images and learns a visual dictionary. The second one is composed of four processes:

1) Feature extraction to build instance representations.
2) Encoding of instance representations into a discriminative space.
3) Pooling the encoded representations into a mid-level representation for each image, and
4) Classifier training on these mid-level representations.

An illustration of this process is shown in Fig. 2.

MIL techniques, and in particular BoVW, have been previously proposed with success for medical imaging applications [22]. For instance, MIL was applied in [23] for obstructive pulmonary disease detection on lung CT scans, in [24] for segmentation and diagnosis of histopathology images, or for detecting early signs of dementia on brain MRI in [25]. MIL has also been proposed for retinal image analysis tasks [26], [27]. In this context, the closer techniques to the method proposed in this paper are [22] and [28]. In [22], a MIL-based DR system is proposed, based on a complex pipeline involving multi-scale patch extraction and alternate local-global weight updating to optimize distances between relevant instances in the feature space. In [28], Pires *et al.* introduce a BoVW technique for DR detection based on sparse Speeded Up Robust Features [29] features with a semi-soft encoding scheme and max-pooling.

The approach proposed in this paper is also based on the BoVW framework. However, we depart from the conventional two-stage process which: firstly, learns a visual dictionary, and secondly, trains a binary classifier on mid-level bag representations. This is achieved by simultaneously learning to encode the instance-level feature vectors onto useful representations and learning to classify bags based on them. Thanks to this joint learning process, the learned representations are enforced to be useful for the classification task. This way, the classification performance directly drives the learning process from end to end. Furthermore, the mid-level representations are also constrained via a new loss function that is designed to enhance their interpretability. An overview of the proposed system is illustrated on Fig. 3.

III. BoVW FOR INTERPRETABLE DR DETECTION

To formalize the BoVW approach for MIL in the context of DR detection, let us consider a training dataset $\{(\mathbf{B}_n, l_n)\}_{n=1}^{N}$ of N retinal eye fundus images $\mathbf{B}_n$ with associated labels l_n, indicating whether they contain pathological signs. From each image $\mathbf{B}$, $N(\mathbf{B})$ instances are extracted, consisting of a set of local descriptors from a variable number of image regions.

In this case, each bag $\mathbf{B}$ is modeled as follows:

$$\mathbf{B} \approx \{\mathbf{x}^i, \quad 1 \leq i \leq N(\mathbf{B})\}, \tag{1}$$

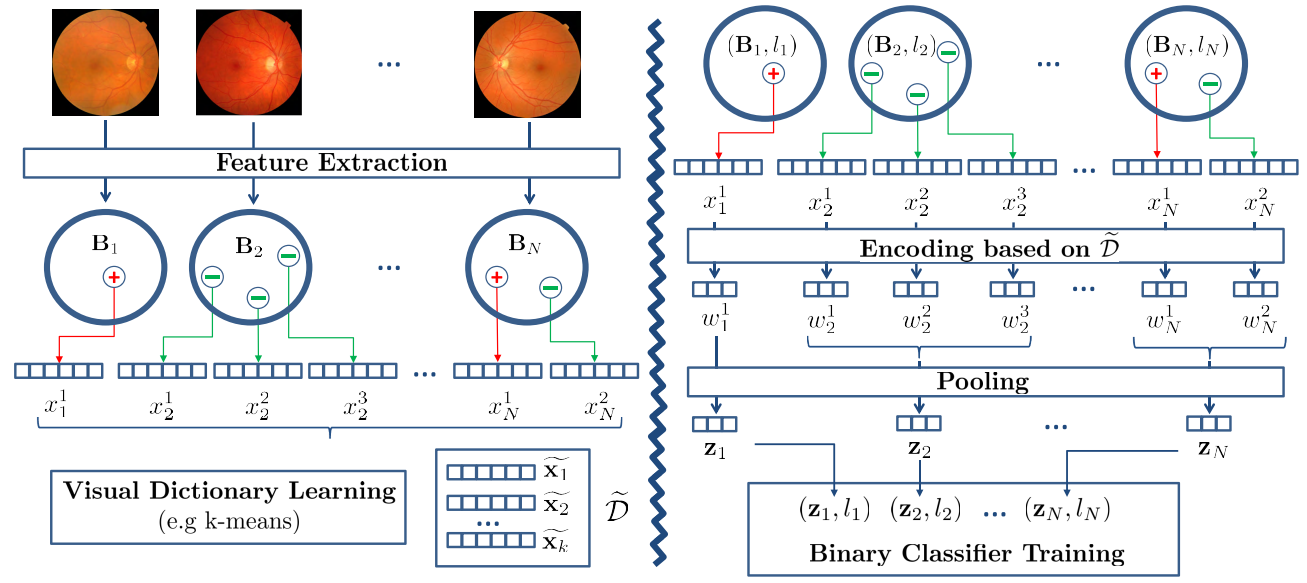


FIGURE 2. Conventional BoVW framework. Note that the visual dictionary is learned in a separate stage. Hence, the learned mid-level representations $\mathbf{z_i}$ are completely independent of the binary classifier's performance.

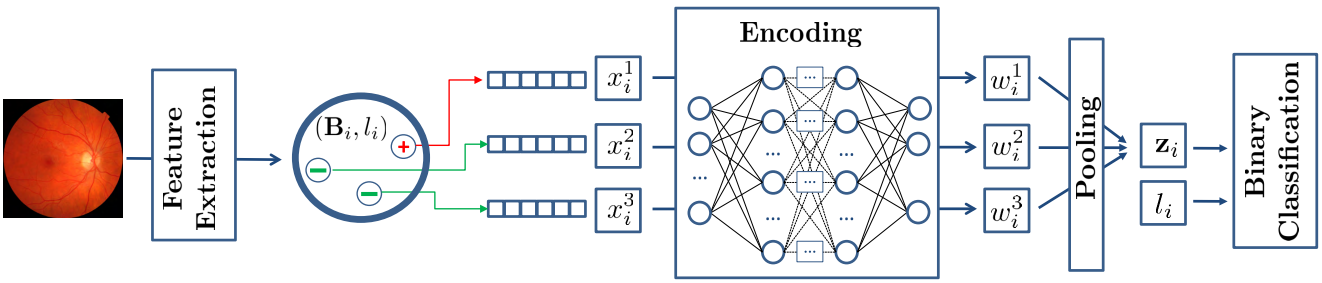


FIGURE 3. Improved BoVW algorithm. The system avoids the explicit creation of a visual dictionary. Furthermore, note that in the proposed version of the BoVW the classifier's performance drives the selection of optimal mid-level representations $\mathbf{z_i}$, as opposed to the conventional approach.

where $\mathbf{x}^i \in \mathbb{R}^d$ is a feature vector describing the i-th instance found in $\mathbf{B}$.

In order to classify a new bag, we need to train a binary classifier. However, since $N(\mathbf{B})$ varies for each image $\mathbf{B}$, the description in eq. (1) is not suitable for this task. The conventional BoVW approach proceeds by extracting the representations for all images in a training set and aggregating them, obtaining a set $\mathcal{D}$ of $(\sum_{i=1}^{N} N(\mathbf{B}_i))$ d–dimensional descriptors. On this set, an unsupervised clustering technique can be applied, *e.g.* k–means. In this case, the set of descriptors is summarized into k centroids, known as visual words, which compose the visual dictionary $\widetilde{\mathcal{D}} = \{\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots \widetilde{\mathbf{x}}_k\}$.

Once a suitable dictionary $\widetilde{\mathcal{D}}$ is learned, then for every training bag $\mathbf{B}$ the method encodes each of its instances $\mathbf{x}^i$ into a set of k-dimensional codes $\mathbf{w}^i$ based on $\widetilde{\mathcal{D}}$ by means of an embedding function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$. All the $N(\mathbf{B})$ codes of different instances extracted from $\mathbf{B}$ are finally pooled in order to obtain a single k-dimensional representation $\mathbf{z}$. This can be achieved for instance with the max-pooling operation $P : \mathbb{R}^{N(\mathbf{B})} \times \mathbb{R}^k \rightarrow \mathbb{R}^k$. In this case, the m–th element in $\mathbf{z}$ is given by:

$$z_m = \max_{1 \leq j \leq N(\mathbf{B})} w_m^j. \tag{2}$$

We refer to this final pooled vector $\mathbf{z}$ as the bag's mid-level representation associated with image $\mathbf{B}$. After processing the training set, the resulting set of mid-level representations, together with the corresponding bag-level labels, are finally supplied to a classifier $\mathcal{C}(\mathbf{B})$, which is trained to discriminate between positive or negative bags.

A. JOINTLY LEARNING TO ENCODE INSTANCES AND CLASSIFY BAGS

The standard BoVW strategy outlined above presents several important disadvantages. A relevant deficiency is that the visual dictionary construction, which determines the resulting mid-level representations, is a process completely isolated from the training of the binary classifier. To compensate for this and build a sufficiently expressive dictionary that can capture the complexity of the feature space, its size is typically large. In some cases the amount of considered

visual words can reach thousands [26]. Existing alternatives comprise for instance considering a hierarchical coarse-to-fine dictionary learning, on which an initial fine-grained dictionary is iteratively refined employing the bag-level labels, until an optimal size is reached [30]. However, this is a cumbersome step requiring itself a separate optimization process.

To overcome these drawbacks, a new strategy to simultaneously learn the encoding and classification steps is proposed in this work. In order to avoid the construction of a visual dictionary, two neural networks are built: the first one, $\mathbf{U}(\mathbf{x}; \theta_{\mathbf{U}})$, learns optimal weights $\theta_{\mathbf{U}}$ to produce useful mid-level representations $\mathbf{z}$ while the second, $\mathbf{D}(\mathbf{z}; \theta_{\mathbf{D}})$ receives those representations and its parameters $\theta_{\mathbf{D}}$ are optimized to perform accurate bag-level classification. The error of $\mathbf{D}$ is back-propagated directly to $\mathbf{U}$, influencing the way in which the mid-level representations are produced by it. In this way, both processes can benefit from each other. An overview of this improved BoVW approach is shown in Fig. 3.

Technically, the neural network $\mathbf{U}(\mathbf{x}; \theta)$ is defined by a series of layers $j \in \{1, \ldots, L\}$ with M_j hidden neurons that perform simple linear operations specified by weights θ^j on their inputs $\mathbf{x}$, followed by a non-linear operation:

$$\mathbf{x} \mapsto \sigma(\theta^j \cdot \mathbf{x}) \tag{3}$$

where the first element of $\mathbf{x}$ is set as $x_0 = 1$ in such a way that θ^j_0 contains the bias term. In the above equation, σ denotes a sigmoid function or any other kind of non-linearity.

For a given bag $\mathbf{B}$, each of its instances $\mathbf{x}^i$ going through $\mathbf{U}$ will be encoded into a M_L-dimensional code vector $\mathbf{w}^i$. Note that the last layer of $\mathbf{U}$ is followed by a softmax activation function, ensuring that $\sum_{j=1}^{k} w^i_j = 1$. The set of codes $\{\mathbf{w}^i,\ 1 \leq i \leq N(\mathbf{B})\}$ computed from every instance in $\mathbf{B}$ are then pooled into the mid-level representation $\mathbf{z}$.

Regarding the pooling stage, several options can be applied, such as average pooling, which averages all codes in $\{\mathbf{w}^i,\ 1 \leq i \leq N(\mathbf{B})\}$. However, this can lead to a smoothing effect due to the contribution of all instances from the bag, even when some of them may be irrelevant.

An alternative to avoid this effect is to perform a max-pooling operation $\mathbf{P}$, as defined in eq. (2). In addition to sharper mid-level representations, max-pooling matches better the goal of DR detection. If no abnormal instance is found on the mid-level representation associated to an image $\mathbf{B}$, it should be declared as healthy. On the other hand, the presence of a single microaneurysm or any other kind of lesion is enough to classify the image as pathological. In our case, max-pooling is implemented to accept the output codes $\{\mathbf{w}^i,\ 1 \leq i \leq N(\mathbf{B})\}$ of the last hidden layer of $\mathbf{U}$ and pool the results into $\mathbf{z}$.

While training $\mathbf{U}$, the obtained mid-level representations become the input of the second neural network $\mathbf{D}$, defined in the same way as $\mathbf{U}$. The output of the final layer of $\mathbf{D}$ is supplied to a sigmoid activation unit, which produces a single output containing the prediction of the system regarding the presence or not of DR on the image $\mathbf{B}$ from where the instances where extracted. In training time, this prediction is compared with the actual label of the image by means of a cross-entropy loss penalizing inaccurate predictions:

$$\mathcal{L}_{class} = \frac{-1}{N} \sum_{i=1}^{N} l_i \log(\mathbf{D}(\mathbf{z}_i)) + (1 - l_i) \log(1 - \mathbf{D}(\mathbf{z}_i)), \tag{4}$$

where $\mathbf{z}_i$ is the mid-level representation of the training image $\mathbf{B}_i$, and l_i its corresponding ground-truth label. The weights $\theta_{\mathbf{U}}$ and $\theta_{\mathbf{D}}$ of both networks are iteratively updated until convergence by standard back-propagation with mini-batch stochastic gradient descent, in order to minimize the error given by eq. (4).

After jointly training $\mathbf{U}$ and $\mathbf{D}$, given a new image $\mathbf{B}$, the output of the proposed model is a prediction of the probability p of $\mathbf{B}$ being affected by DR, *i.e.*, $p = \mathbf{D}(\mathbf{P}(\mathbf{U}(\mathbf{B})))$.

B. A STRATEGY TO ENFORCE MODEL INTERPRETABILITY

Ideally, the mid-level representations $\mathbf{z}$ obtained from pooling the encoded instances $\{\mathbf{w}^i,\ 1 \leq i \leq N(\mathbf{B})\}$ extracted from an image $\mathbf{B}$ contain visually meaningful content. However, this behavior of the proposed model can be further enforced.

Since we know that healthy images only contain healthy instances, we can act at the instance level on codes $\mathbf{w}^i$ extracted from healthy images. The goal in this case is to force the model to generate sparse mid-level representations. This can be accomplished by requiring that, when the label of an image $\mathbf{B}$ is negative, $\mathbf{U}$ uses few codes to encode all its instances. In this way, after pooling a set of codes from healthy instances, the resulting mid-level representation will necessarily be sparse.

In order to impose this behavior on the model, consider a healthy image $\mathbf{B}$, its set of instances $\{\mathbf{x}^i,\ 1 \leq i \leq N(\mathbf{B})\}$, and their corresponding codes $\{\mathbf{w}^i\ 1 \leq i \leq N(\mathbf{B})\}$. We define the following quantity for healthy bags:

$$\mathcal{L}_h = \frac{1}{N} \sum_{i=1}^{N(\mathbf{B})} -\log(w^i_1), \tag{5}$$

which reaches a minimum whenever the code generated by $\mathbf{U}$ is closer to the unitary vector $(1, 0, \ldots, 0)$, since codes $\mathbf{w}^i$ are normalized to sum up to 1.

On the other hand, if the image $\mathbf{B}$ is pathological, it may contain both pathological and healthy instances. We cannot proceed in the same way and constrain the resulting mid-level representation from instance-level codes $\mathbf{w}^{\mathbf{i}}$. However, we can still act at the bag level, by imposing that the mid-level representation z is dense. In this case, we can define the following quantity for pathological bags:

$$\mathcal{L}_p = \frac{1}{M_L} \sum_{i=1}^{M_L} -\log(z_i), \tag{6}$$

which will be minimized whenever the mid-level representation $\mathbf{z}$ is closer to the vector $(1, 1, \ldots, 1)$. Since $\mathbf{z}$ is the result of pooling codes coming from every instance in $\mathbf{B}$, this can only happen if the model encodes pathological instances with different visual words.

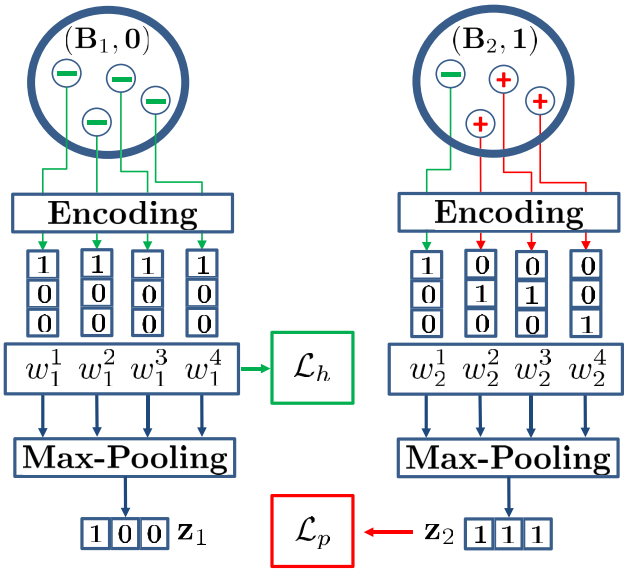


FIGURE 4. A visual explanation of the interpretability-enhancing loss behavior. When the system receives a healthy image, every instance ideally contributes to the same codes in $\mathcal{L}_h$, while disease instances, only present on images showing signs of DR, appear as denser visual words contributing more to $\mathcal{L}_p$.

Finally, given a bag $\mathbf{B}$ with corresponding label l, we define an interpretability-enhancement loss as the combination of both $\mathcal{L}_h$ and $\mathcal{L}_p$:

$$\mathcal{L}_{int} = \alpha(1-l)\mathcal{L}_h + \beta \cdot l \cdot \mathcal{L}_p. \tag{7}$$

Note that $l = 0$ whenever $\mathbf{B}$ is healthy, whereas $l = 1$ for pathological images. In this way, each of the components of $\mathcal{L}_{int}$ becomes active depending of bag-level information. Parameters α and β are positive real-valued hyper-parameters, weighting the contribution of each term. The global loss function that drives the learning of the entire system is simply the addition of the binary classification loss defined in eq. (4) and the interpretability-enhancement loss of eq. (7):

$$\mathcal{L}_{global} = \mathcal{L}_{class} + \mathcal{L}_{int}. \tag{8}$$

A schematic representation illustrating the different situations the model may encounter, and the way in which the interpretability-enhancement loss in eq. (7) reacts to them, is shown in Fig. 4.

C. IMPLEMENTATION DETAILS

In order to apply the proposed method to the DR detection problem, we need to decide on the feature extraction and description methods. In this case, SURF features [29] were employed for both feature description and extraction as they have been shown to perform better than other description methods for DR detection tasks [28]. SURF is a scale and rotation invariant method that detects and describes interest points in an image. To better describe each interest point, 128 dimensional extended descriptors were computed. We used OpenCV's implementation of SURF [31] with default parameters and Theano [32] to implement the two neural networks **U** and **D**.

TABLE 1. DR grading rules for the Messidor Dataset MA = Microaneurysms, HE = Hard Exudates, NV = Neo-Vessels.

DR gr.	Description	Images
0	$N_{MA} = 0$ and $N_{HE} = 0$	546
1	$0 < N_{MA} \leq 5$ and $N_{HE} = 0$	153
2	$5 < N_{MA} < 15$ and $0 < N_{HE} \leq 5$ and $N_{NV} = 0$	247
3	$N_{MA} \geq 15$ or $N_{HE} \geq 5$ or $N_{NV} > 0$	254

IV. EXPERIMENTAL EVALUATION

In this section, we provide experimental assessment of the performance of the proposed method when compared with other recent approaches. Performance is reported in terms of Area Under the receiver operating characteristic Curve (AUC) for the task of DR detection and DR referral, and we finally verify the enhanced interpretability of the proposed model, illustrating that it can effectively reveal the regions contributing to detect pathological images.

A. DR DETECTION PERFORMANCE EVALUATION

We first evaluate the proposed DR detection technique on the publicly available Messidor dataset [33]. Messidor contains 1200 color retinal fundus images acquired on three different French hospitals between 2005 and 2006. Images were obtained with TRC-NW6 non-mydriatic retinographs (Topcon, Tokyo) with a 45° field of view, at a varying resolution of 1440×960, 2240×1488 and 2304×1536. No image pre-processing was applied before extracting and describing the instances within these images.

The clinical information associated to each image on Messidor consists of two labels indicating the grade of DR and the risk of macular edema, based on the presence and number of different types of lesions, see Table 1. In order to build a DR presence label for each image, the provided values are merged in such a way that any image associated to a DR severity greater or equal than one is labeled as pathological, meaning that it contains early signs of DR, while only grade 0 images are considered as healthy. This resulted in a dataset on which 546 images were labeled as normal and 654 as pathological.

The Messidor dataset has been widely employed in the literature to assess the performance of DR detection and grading techniques. Some methods approach the problem by designing a separate detector for each of possible lesions, and then applying it to Messidor images. The output of these lesion detectors is then combined with the set of rules in Table 1 in order to produce a DR presence/grade decision. However, it is important to stress that this approach requires the availability of an independent database containing pixel-wise ground-truth at the lesion level. This is precisely the challenge that our technique and other MIL-based methods try to overcome.

TABLE 2. Performance comparison of DR detection methods tested on the Messidor dataset.

Method	AUC	Observations
Red+Bright Lesion Detection [34]	88%	Shape, color, contrast features + kNN, combines [35] & [14]. **Requires lesion annotations to be trained**.
Ensemble [36]	88%	Ensembling of lesion detectors. **Requires lesion annotations to be trained**.
Red-Lesion Detection [13]	90%	Dynamic shape features + Random Forest classification. **Requires lesion annotations to be trained**.
DREAM [16]	90%	Ensembling of lesion detectors. **Requires lesion annotations to be trained**.
Inception V3 [37]	68%	Fine-tuned Deep Convolutional Neural Network.
MIL Benchmark [38]	81%	Benchmarking of a set of 11 MIL techinques. Best result reported here.
Multiscale AM/FM [39]	84%	Frequency Analysis for Feature Extraction + Mahalanobis dist. for Classification
AM/FM - SVM [40]	86%	Frequency Analysis for Feature Extraction + SVM for Classification
MIL for DR detection [22]	88%	BoVW with separate Encoding+Classification training.
Data-Mined Context [30]	89%	GFTT+SIFT features. Mines contextual data from patient's record, including text.
Ours	**90%**	**Proposed Approach: MIL with joint Encoding+Classification training.**

A standard evaluation procedure was followed: 20% of the dataset was held-out for testing, while 65% was employed for training and 15% for validation. Hyper-parameters were found using random search [41], selecting the best values in terms of AUC on the validation set. Performance results of the proposed technique are shown in Table 2 in terms of AUC, together with the performance obtained in the same dataset by different state-of-the-art techniques. We include both methods trained on independent datasets for the task of lesion detection and methods that learn directly from the image-level ground-truth in order to predict DR detection.

The results on Table 2 lead to several conclusions. First, the proposed method achieves a superior performance comparing to the other techniques that have been trained without access to pixel-level lesion annotations. It is particularly interesting to note that other MIL-based techniques such as [22], or the best of the DR detection techniques reported in [38], obtain a lower AUC. The main difference between all these methods and the technique introduced in this paper is the propagation of the bag-level labels until the encoding process, which directly benefits from this information in order to produce more useful mid-level representations, leading to a better detection performance. Second, performance of methods trained with lesion-level ground-truth is comparable but not superior to the introduced technique. This means that the proposed MIL-based approach for DR detection can effectively make use of local information on the image to the same extent as these techniques, but without having explicit access to it.

It should be noted that other recent approaches based on Deep Convolutional Neural Networks (CNN) have been tested with great success on the Messidor dataset [1], [6], achieving even larger AUC values without the need of lesion-level information. However, these studies propose models trained on external large dataset of retinal images, containing several dozens of thousands of training images. Moreover, the output of this kind of CNN-based models typically lacks interpretability, which may hinder the predisposition of doctors towards its acceptance in a regular clinical workflow. The method introduced in this paper addresses both issues by leveraging as much information as possible from a moderate-size dataset, while enforcing the interpretable behavior of the model, as illustrated in section IV-C. However, in order to develop a more comprehensive assessment of deep-learning based systems on medium-sized datasets, we include for comparison the performance of a fine-tuned deep convolutional neural network [37], after pre-training on the ImageNet database [42]. It has been proven in [43] that fine-tuning a CNN that has been pre-trained on ImageNet is the most meaningful approach to implement deep neural networks on moderate-sized datasets.

TABLE 3. Performance comparison of DR detection on the DR1 dataset.

Method	AUC	Observations
Multi-Lesion fusion [45]	84%	Lesion detection + meta-SVM.
Inception V3 [37]	97%	Fine-tuned Deep CNN.
Ours	**93%**	**Proposed Approach.**

In order to test if the proposed method generalizes to different datasets, we also tried to detect DR as the presence of any single lesion in the DR1 dataset, introduced in [44]. This dataset contains 1077 retinal images captured with a TRC-50X (Topcon Inc., Tokyo, Japan) mydriatic camera with a 45° field of view and an average resolution of 640 × 1077 pixels. From all the images, 595 were classified as containing no sign of DR and 482 as showing pathological signs. In this case, we are only aware of a work addressing the task of DR detection on this dataset [45]. Since DR1 contains ground-truth regarding the presence of different lesions within each pathological image, DR detection is achieved by training separate detectors for each of them and then fusing the results with a meta-classifier. Performance comparison with the results in [45] is shown in Table 3. In this case, the proposed technique clearly outperforms the method proposed in [45], which confirms the generality of our approach.

TABLE 4. Performance comparison of DR referral methods tested on the DR2 dataset.

Method	AUC	Observations
Inception V3 [37]	91%	Fine-tuned Deep Convolutional Neural Network.
Pires *et al.* 2013 [44]	93%	Separate Lesion Detectors + Meta-Classifier. **Requires weak lesion information to be trained**.
Pires *et al.* 2014 [28]	94%	Separate Lesion Detectors + Meta-Classifier. **Requires weak lesion information to be trained**.
Pires *et al.* 2017 [26]	96%	Bypasses Lesion Detection. BossaNova and Fisher Vector features + BoVW.
Ours	**96%**	**Proposed Approach.**

B. DR REFERRAL PERFORMANCE EVALUATION

It has been argued in [44] that the presence of a given lesion may not be enough to make a decision on the need to refer a patient for further examination. Table 1 contains a set of rules designed in order to make such a decision, but it may not cover all the signs an expert ophthalmologist takes into account when recommending further examination of a patient. For instance, in [46], referral is defined as having a DR grade above mild non-proliferative (R1) and/or macular edema. Lesion location may also impact the clinical decision regarding referral.

In order to assess the performance of the proposed method in terms of DR referral prediction, we employ the publicly available DR2 dataset. This dataset was introduced in [44], and it is composed of 520 retinal fundus images, from which 337 images were categorized by two independent ophthalmologists as not requiring referral, and 98 were deemed to require referral within one year by a specialist. It is important to note that while labeling the images, the experts were required to categorize them ignoring specific lesions and considering only if the image should lead to referral. The medical specialists based their decision on any reason they considered to be clinically relevant, not only on the presence of particular lesions. DR2 images were acquired with a TRC-NW8 (Topcon Inc., Tokyo, Japan) nonmydriatic retinal camera, and they all have 867×575 pixel resolution with a 45° field of view.

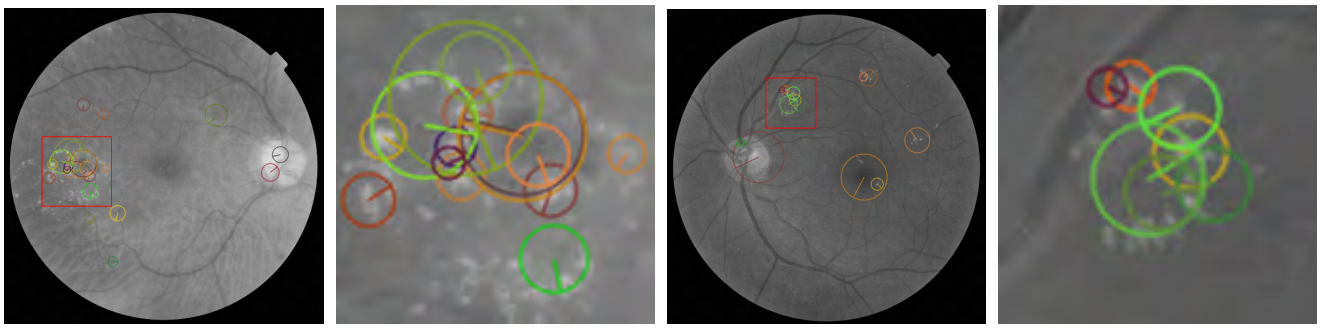
Several methods have been proposed in the past for DR referral prediction [40], [46], [47]. Unfortunately, there is not a standard definition of referral, which results in different problems of varying difficulty being solved. In [47] images containing signs of macular edema were considered as referrable, while in [40] the adopted definition was the presence of signs of high DR grade or high risk of macular edema. In order to be able to properly compare the proposed approach in a fair manner, we select those studies that were tested on DR2, since it contains direct referral opinion from medical experts. In this case, [44] proposed a solution consisting of training individual lesion detectors, and employing the resulting decision scores in order to train a meta-classifier to predict referral. The individual detectors consist on a variant of BoVW with more advanced pooling and encoding operations. In a later work [28], this scheme was improved by means of a semi-soft encoding strategy, with results outperforming those of [44].

It is worth noting that these methods, even if being MIL-based approaches, still need to be trained with weak lesion-level information. In this case, information regarding which kind of lesion is present in an image, but not the exact location and its delineation, is used. In order to train these techniques, the DR1 dataset described in the previous section was used. In contrast, the technique proposed in this paper was trained only on DR2, with no other information than the need for referral of each image. This characteristic is shared by another recent technique introduced in [26]. In that work, it was effectively shown that DR referral could be predicted without the need for explicit lesion detection. However, the proposed method still presents a separate visual dictionary construction and classifier training.

Performance results for DR referrable predictions in terms of AUC is presented in table 4. It can be observed that the proposed technique improves or matches the performance of previously reported methods also in the task of DR referral. The arguments suggested in [26] about the possibility of training a referral prediction system without the need of explicitly building separate lesion detectors are confirmed by these results. We can conclude that both the technique proposed in this work and the one introduced in [26] obtain superior performance than lesion-detection based techniques, confirming the validity of this approach. It is important to notice, however, that the method from [26] only addresses DR referral, while the technique presented here is tested both in DR detection and referral. Moreover, the results obtained by our technique are better interpretable than those produced by [26]. In the next section we analyze this aspect of our model.

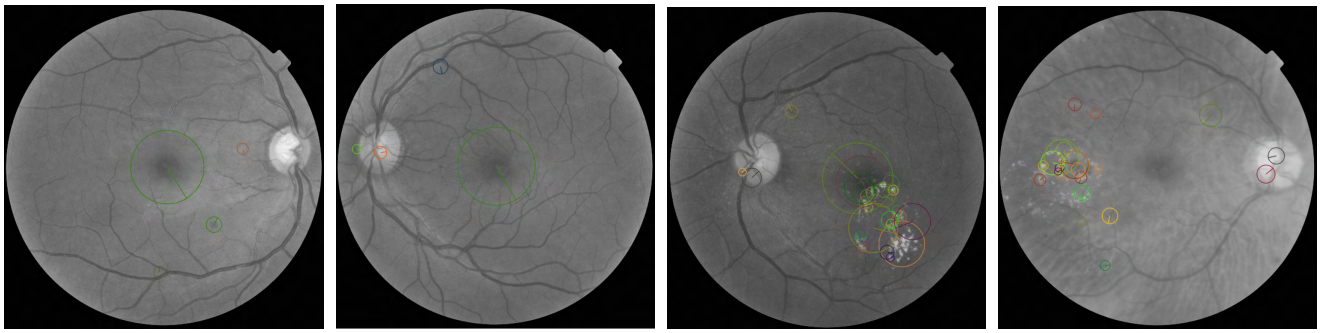
C. INTERPRETABILITY OF THE MODEL

One of the most relevant features of the DR detection system proposed in this paper is its enhanced interpretability, allowed by the joint minimization of the two loss functions in eqs. (4) and (7). This enables us to explain which instances within the images most likely caused the model to reach the produced decision. To experimentally demonstrate this aspect of the model, a first example of this behavior on pathological images from the DR2 dataset is shown in Fig. 5. In Figs. (5a) and (5c), SURF keypoints contributing to the resulting mid-level representations of these images are depicted. We can clearly see how most of the selected instances correspond to keypoints extracted from bright lesions, while only few keypoints are related to instances that are typically present on both normal and pathological retinal images, such as the macula or the optic disc. Zoomed-in



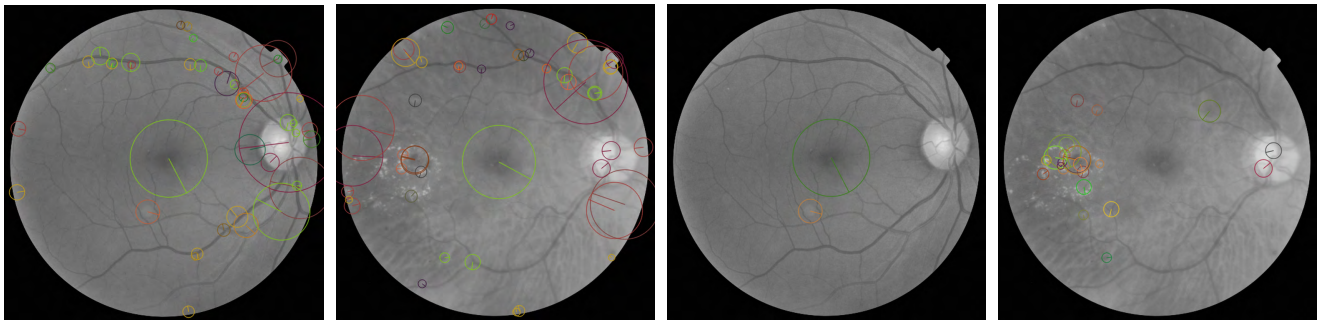
(a) (b) (c) (d)

FIGURE 5. Instances on two pathological images that contributed to the decision produced by the proposed system. In this case, it can be appreciated that most of the SURF keypoints are located on top of bright lesions. Best viewed in color.



(a) (b) (c) (d)

FIGURE 6. SURF keypoints associated to image instances considered by the proposed model in order to produce a decision on DR presence. (a) and (b) depict healthy images, on which fewer instances were taken into account. On the contrary, (c) and (d) show pathological images, on which a greater amount of instances are considered to reach a decision. Best viewed in color.



(a) (b) (c) (d)

FIGURE 7. Comparison between results produced by a model trained without the interpretability-enhacnement loss - (a), (b) - and after adding it - (c), (d). In this case, (a) and (c) show a healthy image, while (b) and (d) show a pathological example. Best viewed in color.

details are also shown in Figs. (5b) and (5d) to better present this observation.

A second experiment was run to better illustrate that the model trained with the interpretability-enhancement loss behaves as desired. The loss function introduced in eq. (7) aims at promoting a sparse mid-level representation for healthy images, on which few instances are ideally considered, while in the case of pathological images, the produced mid-level representations are expected to be denser. This should translate into more SURF keypoints appearing when considering pathological examples. This is visually verified in Fig. 6. There it can be observed that fewer keypoints were taken into account when reaching a decision regarding a healthy image, see Figs. (6a) and (6b), than when a pathological image was considered, as shown in Figs. (6c) and (6d).

To further verify that the loss term in eq. (7) effectively contributes to the explainability of the model's decision, we trained a separate classifier by minimizing only the loss function of eq. (4), without including the interpretability-enhancement loss term. Both results are visually compared in Fig. 7. It can be seen how the extra loss term leads to more interpretable results by a better identification of the pathological instances. Only a fraction of the input instances are used by the model to produce a decision, while

irrelevant instances are filtered out. It is worth noting that when the interpretability-enhancement loss term was not included, the model considered roughly the same number of keypoints on normal and pathological images, as shown in Figs. (7a) and (7b). However, when the global loss in eq. (7) is minimized, the resulting model considers substantially less keypoints in a healthy example than in a pathological image in order to make a decision, as can be observed in Figs. (7c) and (7d).

V. DISCUSSION

From the above experiments, it can be seen that the results obtained by the proposed approach outperformed other DR detection techniques in most cases. It is worth noting, however, that in the case of the DR1 dataset, a state-of-the-art Convolutional Neural Network achieved higher AUC. CNNs are powerful classification models, but they are known to provide results that are hard to interpret by a user. In addition, the same model was tested on the Messidor and the DR2 datasets, obtaining a lower performance. This indicates that the model generalizes poorly to different data sources, in particular when there is few training data available (as in the case of the DR2 dataset). However, it is important to mention that, for a fair comparison, neither for Inception V3 nor for any other technique (including the proposed approach), we performed artificial data augmentation. This could be expected to lead to some accuracy improvement on CNN models.

When compared with the remaining techniques, it can be observed that the proposed technique brought substantial performance increases in every considered dataset. When evaluated for DR detection in the Messidor dataset, the weakly-supervised technique introduced in this paper achieved the largest performance among every method that does not require pixel-wise lesion annotations to be trained. From the four considered techniques that employ lesion annotations, two of them achieved a similar performance, while the other two obtained lower DR detection AUCs. This is highly relevant, since producing this kind of annotations is a costly and time-consuming process, and the ability to bypass it is a great advantage.

The results for the task of DR referral on the DR2 dataset confirmed the good performance of the proposed approach: from the four other considered techniques, only one achieved similar performance, indicating that our technique competes well or outperforms other previous methods. Moreover, the proposed interpretability-enhancement mechanism was qualitatively shown to offer highly interpretable results, which has great importance for the potential adoption of an automatic computer-aided diagnosis in a real clinical work-flow.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, a new weakly-supervised Diabetic Retinopathy (DR) detection system has been presented, based on the Multiple-Instance Learning framework. The method can learn from weak information regarding only the presence or absence of disease to formulate predictions on new images based on implicit local information. The main novelty of the proposed model with respect to previously existing MIL-based DR detection systems is a joint-learning scheme in which the encoding and the classification stages are connected. Thanks to this approach, the mid-level representations generated by the model are optimized to improve DR detection accuracy. Furthermore, a novel strategy to enforce the interpretability of the resulting predictions has been introduced, resulting in a better understanding of the output of the model. Performance comparisons against other recent DR detection and DR referral techniques give advantage to the proposed technique, confirming previous observations stating that weak expert labels (at the image level only) can be leveraged to produce accurate predictions without the need of pixel-level information related to the different lesions indicating the presence of DR.

The developed technique achieves good performance, but further improvements can be achieved. Speeded-Up Robust Features (SURF) were employed in this work to locate and describe instances within retinal images. Even if the proposed technique jointly optimizes the encoding and the classification stages of the model, instance location and description may be included in the same global optimization process. This could be achieved with an end-to-end system in which the most appropriate image representation for the task of DR detection is also learned by using a Deep Convolutional Neural Network. Further work will involve exploring this direction of research, in order to obtain higher performance in terms of DR detection, as well as extending the approach to predicting different levels of DR severity.

REFERENCES

[1] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.
[2] D. M. Squirrell and J. F. Talbot, "Screening for diabetic retinopathy," *J. Roy. Soc. Med.*, vol. 96, no. 6, pp. 273–276, Jun. 2003.
[3] P. H. Scanlon, "The english national screening programme for diabetic retinopathy 2003–2016," *Acta Diabetol.*, vol. 54, no. 6, pp. 515–525, 2017.
[4] L. P. Daskivich, C. Vasquez, C. Martinez, C.-H. Tseng, and C. M. Mangione, "Implementation and evaluation of a large-scale teleretinal diabetic retinopathy screening program in the los angeles county department of health services," *JAMA Internal Med.*, vol. 177, no. 5, pp. 642–649, 2017.
[5] J. Beagley, L. Guariguata, C. Weil, and A. A. Motala, "Global estimates of undiagnosed diabetes in adults," *Diabetes Res. Clin. Pract.*, vol. 103, no. 2, pp. 150–160, 2014.
[6] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.
[7] D. Fleming, S. Philip, K. A. Goatman, G. J. Prescott, P. F. Sharp, and J. A. Olson, "The evidence for automated grading in diabetic retinopathy screening," *Current Diabetes Rev.*, vol. 7, no. 4, pp. 246–252, 2011.
[8] E. Soto-Pedre *et al.*, "Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload," *Acta Ophthalmol.*, vol. 93, no. 1, p. e52–56, 2015.
[9] P. Costa and A. Campilho, "Convolutional bag of words for diabetic retinopathy detection from eye fundus images," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl.*, 2017, pp. 165–168.
[10] M. Niemeijer *et al.*, "Retinopathy Online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.

[11] C. I. Sánchez, M. Niemeijer, M. S. A. S. Schulten, M. Abràmoff, and B. van Ginneken, "Improving hard exudate detection in retinal images through a combination of local and contextual information," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro*, Apr. 2010, pp. 5–8.

[12] R. Srivastava, D. W. K. Wong, L. Duan, J. Liu, and T. Y. Wong, "Red lesion detection in retinal fundus images using Frangi-based filters," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 5663–5666.

[13] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. M. P. Langlois, "Red lesion detection using dynamic shape features for diabetic retinopathy screening," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 1116–1126, Apr. 2016.

[14] M. Niemeijer, B. van Ginneken, S. R. Russell, M. S. A. Suttorp-Schulten, and M. D. Abràmoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," *Invest. Ophthalmol. Vis. Sci.*, vol. 48, no. 5, pp. 2260–2267, 2007.

[15] K. S. Deepak, A. Chakravarty, and J. Sivaswamy, "Visual saliency based bright lesion detection and discrimination in retinal images," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, Apr. 2013, pp. 1436–1439.

[16] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1717–1728, Sep. 2014.

[17] I. N. Figueiredo, S. Kumar, C. M. Oliveira, J. D. Ramos, and B. Engquist, "Automated lesion detectors in retinal fundus images," *Comput. Biol. Med.*, vol. 66, pp. 47–65, Nov. 2015.

[18] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. 15th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 2002, pp. 577–584.

[19] P. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 2005, pp. 1417–1424.

[20] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[21] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.

[22] G. Quellec *et al.*, "A multiple-instance learning framework for diabetic retinopathy screening," *Med. Image Anal.*, vol. 16, no. 6, pp. 1228–1240, 2012.

[23] V. Cheplygina, L. Sørensen, D. M. J. Tax, J. H. Pedersen, M. Loog, and M. de Bruijne, "Classification of COPD with multiple instance learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1508–1513.

[24] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 964–971.

[25] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, and D. Rueckert, "Multiple instance learning for classification of dementia in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, 2014.

[26] R. Pires, S. Avila, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Beyond lesion-based diabetic retinopathy: A direct approach for referral," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 193–200, Jan. 2017.

[27] S. Manivannan, C. Cobb, S. Burgess, and E. Trucco, "Subcategory classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification," *IEEE Trans. Med. Imag.*, vol. 36, no. 5, pp. 1140–1150, May 2017.

[28] R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Advancing bag-of-visual-words representations for lesion classification in retinal images," *PLOS ONE*, vol. 9, no. 6, p. e96814, 2014.

[29] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[30] G. Quellec *et al.*, "Automatic detection of referral patients due to retinal pathologies through data mining," *Med. Image Anal.*, vol. 29, pp. 47–64, Apr. 2016.

[31] OPENCV, (2015). *Open Source Computer Vision Library*. [Online]. Available: https://github.com/itseez/opencv

[32] The Theano Development Tea *et al.*.(May 2016)"Theano: A python framework for fast computation of mathematical expressions." [Online]. Available: https://arxiv.org/abs/1605.02688

[33] E. Decencière *et al.*, "Feedback on a publicly distributed image database: The messidor database," *Image Anal. Stereology*, vol. 33, no. 3, pp. 231–234, 2014.

[34] C. I. Sánchez, M. Niemeijer, A. V. Dumitrescu, M. S. A. Suttorp-Schulten, M. D. Abràmoff, and B. van Ginneken, "Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data," *Invest. Ophthalmol. Vis. Sci.*, vol. 52, no. 7, pp. 4866–4871, Jun. 2011.

[35] M. Niemeijer, B. V. Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abramoff, "Automatic detection of red lesions in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 24, no. 5, pp. 584–592, May 2005.

[36] B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1720–1726, Jun. 2012.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[38] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Comput. Med. Imag. Graph.*, vol. 42, pp. 44–50, Jun. 2015.

[39] C. Agurto *et al.*, "Multiscale AM-FM methods for diabetic retinopathy lesion detection," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 502–512, Feb. 2010.

[40] E. S. Barriga *et al.*, "Automatic system for diabetic retinopathy screening based on AM-FM, partial least squares, and support vector machines," in *Proc. IEEE Int. Symp. Biomed. Imag., Nano Macro*, Apr. 2010, pp. 1349–1352.

[41] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[42] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015. [Online]. Available: https://link.springer.com/article/10.1007/s11263-015-0816-y

[43] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[44] R. Pires, H. F. Jelinek, J. Wainer, S. Goldenstein, E. Valle, and A. Rocha, "Assessing the need for referral in automatic diabetic retinopathy detection," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3391–3398, Dec. 2013.

[45] H. F. Jelinek *et al.*, "Data fusion for multi-lesion diabetic retinopathy detection," in *Proc. 25th IEEE Int. Symp. Comput. Based Med. Syst. (CBMS)*, Jun. 2012, pp. 1–4.

[46] M. D. Abràmoff *et al.*, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, 2013.

[47] K. S. Deepak and J. Sivaswamy, "Automatic assessment of macular edema from color retinal images," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 766–776, Mar. 2012.

PEDRO COSTA is currently the M.Sc. Researcher with the Center for Biomedical Engineering Research, Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal. His research interests include medical image processing using machine learning techniques.

ADRIAN GALDRAN is currently a Post-Doctoral Fellow with the Center for Biomedical Engineering Research, Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal. His research interests range from low-level image processing to machine learning for computer vision applications, with special emphasis in medical imaging applications.

ASIM SMAILAGIC (F'10) has been a Research Professor with the Department of Electrical and Computer Engineering, College of Engineering, Carnegie Mellon University (CMU), where he has been a faculty since 1992. He is currently the Director of the Laboratory for Interactive and Wearable Computer Systems, and the Leader of the Virtual Coaches Research Thrust, NSF Engineering Research Center on Quality of Life Technology, CMU, combining machine learning, sensors, and image analysis. He was a recipient of the Fulbright post-doctoral award in computer science, CMU, in 1988, the 2000 Allen Newell Award for Research Excellence from CMU's School of Computer Science, the 2003 Carnegie Science Center Award for Excellence in Information Technology, the 2003 Steve Fenves Systems Research Award from the CMU College of Engineering and other prestigious awards. He was three times Program Chairman of the Quality of Life Technology Symposium, sponsored by the NSF and CMU. He was a Program Chairman of the IEEE conferences over 10 times and a Chair of the IEEE Technical Committee on Wearable Information Systems. He was a Co-Editor, an Associate Editor, and a Guest Editor in leading technical journals, such as the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON VLSI SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the *Journal on VLSI Signal Processing*, and the *Journal on Pervasive Computing*.

AURÉLIO CAMPILHO (SM'14) was an INEB President and a Research Coordinator from 1994 to 2000. He was an Adjunct Professor with the Department of Electrical and Computer Engineering and also with the Department of Systems Design, Faculty of Engineering, University of Waterloo, Canada, from 2002 to 2008. For several years, he served as a President of the Portuguese Association for Pattern Recognition, which is a member of the IAPR. In 2014, he was the Coordinator of the Center for Biomedical Engineering Research, Institute for Systems and Computer Engineering, Technology and Science. He is currently a Full Professor with the Department of Electrical and Computer Engineering, Faculty of Engineering, University of Porto, Portugal. His current research interests include the areas of biomedical engineering, medical image analysis, image processing and computer vision, particularly in computer-aided diagnosis applied in several imaging modalities, including ophthalmic images, carotid ultrasound imaging and computed tomography of the lung. He is currently coordinating Portugal Entrepreneurial Research Initiative SCREEN-DR-Image Analysis and Machine Learning Platform for Innovation in Diabetic Retinopathy Screening at Carnegie Mellon University. He has authored one book (with two editions), co-edited 20 books and published over 200 papers in journals and conferences. He served as an organizer of several special issues and conferences. He is the General Chair of the series of International Conferences on Image Analysis and Recognition. He served as an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and the *Machine Vision Applications Journal*.