Sub-Category Classifiers for Multiple-Instance Learning and its Application to Retinal Nerve Fiber Layer Visibility Classification

Siyamalan Manivannan, *Member IEEE*, Caroline Cobb, Stephen Burgess, and Emanuele Trucco, *Member IEEE*

***Abstract*—We propose a novel multiple instance learning method to assess the visibility (visible/not visible) of the retinal nerve fiber layer (RNFL) in fundus camera images. Using only image-level labels, our approach learns to classify the images as well as to localize the RNFL visible regions. We transform the original feature space to a discriminative subspace, and learn a region-level classifier in that subspace. We propose a margin-based loss function to jointly learn this subspace and the region-level classifier. Experiments with a RNFL dataset containing 884 images annotated by two ophthalmologists give a system-annotator agreement (kappa values) of 0.73 and 0.72 respectively, with an inter-annotator agreement of 0.73. Our system agrees better with the more experienced annotator. Comparative tests with three public datasets (MESSIDOR and DR for diabetic retinopathy, UCSB for breast cancer) show that our novel MIL approach improves performance over the state-of-the-art. Our Matlab code is publicly available at https://github.com/ManiShiyam/Sub-category-classifiers-for-Multiple-Instance-Learning/wiki.**

***Index Terms*—multiple instance learning, image classification, retinal nerve fiber layer, retinal image processing, retinal biomarkers for dementia.**

I. Introduction

This paper introduces an automatic system assessing the visibility and location of the retinal nerve fiber layer (RNFL) in fundus camera (FC) images from image-level labels. The optic nerve transmits visual information from the retina to the brain. The expansion of the neural fibers in the optic nerve enters the retina at the optic disc. Its form the RNFL, the innermost retinal layer (Figure 1). The RNFL has been implicated in prediagnostic stages of glaucoma [1] and recently considered as a potential biomarker for dementia [2], by assessing its thickness in optical confocal tomography (OCT) images. However, screening of high numbers of patients would be enabled if the RNFL could be assessed with FC, still much more common than OCT for retinal inspection, already included in large, cross-linked data sets, and increasingly part of routine optometry checks.

Some RNFL-related studies with FC images have been reported, mostly for estimating glaucoma risk [3], but there is very little work on studying associations with dementia with FC images [4]. This is contrast with RNFL analysis via OCT, supported by a rich literature [2], [5]. The RNFL is not always visible in FC images, and its visibility itself has been posited as a biomarker for neurodegenerative conditions. This motivates our work, part of a larger project on multi-modal retina-brain biomarkers for dementia [6].

We report an automatic system to identify FC images with visible RNFL regions and simultaneously localize visible regions. A crucial challenge is obtaining ground truth annotations of visible RNFL regions from clinicians. Region tracing is notoriously a difficult and time-consuming process. We take therefore a Multiple Instance Learning (MIL) approach, requiring only image-level labels (RNFL visible/invisible), which can be generated much more efficiently. In MIL, images are regarded as *bags*, and image regions as *instances*. Each bag has an associated label, and the labels of its instances are unknown.

Visible RNFL regions have significant intra-class variations, and can be difficult to distinguish from other regions. To address this, we embed the instances in a discriminative subspace defined by the outputs of a set of subcategory classifiers. An instance-level (IL) classifier is then learned in that subspace by maximizing the margin between positive and negative bags. A margin-based loss is used to learn the IL and the subcategory classifiers jointly.

This paper brings two main contributions.

1) To our best knowledge, we address a new problem with significant impact potential for biomarker discovery, i.e. classifying FC images as RNFL-visible/invisible, including region localization.
2) As shown in experiments with a local (RNFL) and 3 public data sets, we improve experimental performance compared to state-of-the-art MIL systems by proposing a novel MIL approach with a novel margin-based loss (instead of the cross-entropy loss commonly used in comparable MIL systems).

We evaluated our approach on a local dataset ("RNFL") of 884 FC images, and with three public datasets (MESSIDOR [7] and DR [8] for diabetic retinopathy, UCSB [9] for breast cancer). Table V summarizes the datasets and the experimental

This work was supported by the EPSRC grant EP/M005976/1.

Siyamalan Manivannan is with Department of Computer Engineering, Faculty of Engineering, University of Jaffna, Sri Lanka. (email: siyam@eng.jfn.ac.lk)

Caroline Cobb and Stephen Burgess are with Department of Ophthalmology, NHS Ninewells, Dundee, UK.

Emanuele Trucco is with CVIP, School of Science and Engineering (Computing), University of Dundee, UK.

Manuscript received August 27, 2016; revised December 31, 2016; accepted January 9, 2017

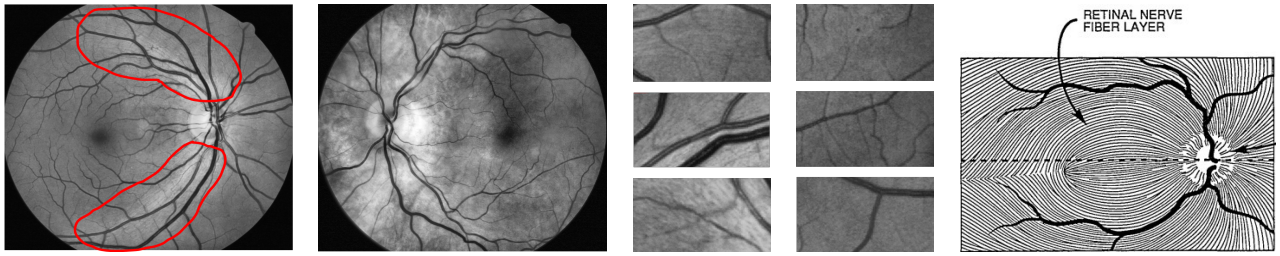


Fig. 1: RNFL visibility in the red-free image: (a) an image with visible RNFL (marked region indicates its visibility), (b) an image with invisible RNFL, (c) examples of RNFL-visible regions, (d) examples of RNFL-invisible regions, (e) a synthetic image showing RNFL and blood vessels [1].

settings used. We collected image-level annotations independently from two practising ophthalmologists (A1 and A2, A1 the more experienced). Overall, they agreed on RNFL visible/invisible image labels $\simeq 91.63\%$ of the time ($\mathcal{P} \simeq 91.63\%$) with a kappa value of $\mathcal{K} \simeq 0.73$. Our experiments suggest that our system agrees very well with both annotators, and better with A1 than A2 (system agreement with A1, $\mathcal{P} \simeq 91.6\%$ with $\mathcal{K} \simeq 0.73$ and A2, $\mathcal{P} \simeq 91.8\%$ with $\mathcal{K} \simeq 0.72$). Our approach also improves state-of-the-art results on the public datasets used (see Table VI).

This paper extends our earlier work [10]. It sets the proposed method in the context of related literature, describes it in more detail, presents more extensive experiments on a larger RNFL dataset with 884 FC images, where image-level annotations were obtained by analysing both the green and the blue channels (previously only 576 images, annotated based on green channel only) to investigate the effect of various components on performance, and summarizes performance in experimental comparisons with other methods.

This paper is organized as follows. The differences between our and recent, comparable work are captured in Section II after a concise discussion of related work. Our proposed approach is explained in detail in Section III followed by experimental validation in Section IV. We conclude the paper and describe future directions in Section V.

II. Related Work

A. RNFL related studies

Some work has been reported on RNFL-related studies with FC images [11–15], mostly as RNFL layer defects are precursors of glaucoma [3]. These approaches can be divided broadly into two categories, (1) image transformation approaches, and (2) patch classification approaches. *Image transformation approaches* [11–13] identify dark stripy patterns in the log-transformed FC images. *Patch classification approaches* [14], [15] are supervised approaches, where annotated image patches from FC images and their associated labels are used to learn a supervised classifier, which in turn is used to predict the label of the given patch (RNFL defect or not).

Our work differs from these approaches mainly in two aspects. First, we focus on identifying RNFL visibility (image-level and region-level) as RNFL is not always visible in the FC images, and its visibility itself could be a candidate bio-marker for neurodegenerative conditions. Second, we propose a novel MIL approach which requires only image-level labels to train the system, hence enables the collection of much larger sets of annotations in the same time span.

B. Multiple instance learning

Various approaches have been proposed since the introduction of MIL by Dietterich et al. [16] in the context of drug activity prediction. Due to the success MIL has been recently explored for several medical imaging problems, for instance cancer detection in digital pathology images [9], [17–19], automated retinopathy screening [7], [8], [19], and lesion detection in lung images [19]. Here we review concisely the most relevant papers for our work (for a general review of MIL, see [20]). MIL approaches can be divided in two broad classes, (1) *instance-level* (IL) and (2) *bag-level* (BL). In both cases a classifier is trained to separate positive bags from negative bags using a loss function defined at the bag level.

IL approaches: the classifier is trained to classify *instances*, obtaining IL predictions. BL predictions are usually obtained by aggregating IL decisions, e.g. DD [16], EM-DD [21], MI-SVM [22], BP-MIL [23], MIL-Boost [24], MCIL-Boost [18]. The *max-rule* is often considered for aggregation, i.e. the prediction of a bag is determined by the top positive instance present in that bag. Under this setting, a bag is considered as positive if it contains at least one positive instance, and all the instances in the negative bags are considered negative. This rule has been widely used, e.g. in DD [16], EM-DD [21], MI-SVM [22] among other methods. This setting, however, discards the information from all other instances except the top positive one. Also, noisy positive instance-level predictions can affect the label of a bag. In some datasets the bag's label is determined by a group of instances instead of one (this is the case in our RNFL dataset, see Figure 8). To overcome these limitations, recent studies adopted relaxed versions of this assumption, in which some or all the instances in a bag contribute to the prediction of that bag [25], [26].

BL approaches: a classifier is trained to classify *bags*. BL approaches can be further categorized into two categories. In the first, a bag-level feature representation is computed from its instance representations, and a supervised classifier is trained

[1] source: http://doctorbond.in/assessment-retinal-nerve-fiber-layer/

on this representation, e.g. MILES [27], JC^{2}MIL [28], RMC-MIL [29]. In the second category, bag-to-bag similarities are computed based on the instance-level feature representations, and a supervised classifier is trained using this similarity matrix, e.g. mi-graph [30]. Since BL approaches are trained to predict bags, instance-level predictions cannot be obtained directly.

The original feature space may not be sufficiently discriminative for the problem at hand. Hence *embedding-based* (EB) approaches have been proposed to embed the instances in a discriminative space, and subsequently a BL (MILES [27], JC^{2}MIL [28], RMC-MIL [29]) or an IL (BRT [17]) classifier is trained in this space.

MIL approaches have also been explored within the recent, successful Convolutional Neural Networks (CNN) paradigm for visual recognition [26], [31]. Here, a MIL pooling layer is introduced at the end of the deep network architecture to aggregate (pool) IL predictions and compute the BL ones.

Our approach is an EB approach; it learns an IL classifier instead of the BL one, and it can therefore provide both IL and BL predictions. CNN+MIL [26], [31] as well as the EB approaches [17], [28], [29] minimize cross-entropy loss. However, recent studies suggest that margin-based loss may be a better choice than the cross-entropy loss for classification problems [32], [33] as it directly minimizes classification errors. Some authors sought to improve the cross-entropy loss by boosting the importance of wrongly classified data points, e.g. [34]. Considering this, we propose a margin-based loss where the bags which violate the margin are penalized, and show improved performance over the cross-entropy loss.

III. Method

A. Motivation and the overview of the method

Most MIL approaches do not make explicit assumptions about the inter or intra-class variations of the positive and negative bags (e.g. [22], [30]). However, with high intra-class variation and low inter-class distinction these approaches may not perform well. This is the case for our RNFL dataset: the visible RNFL regions have a high intra-class variations, and they are often difficult to distinguish from RNFL-invisible regions (Figure 1). To overcome this, we assume there exists a set of discriminative sub-categories, and learn a set of classifiers for them. These sub-categories, for instance, may capture different variations (or visual appearance) of the RNFL regions and background. Each classifier in this pool is learned specifically to separate a particular sub-category from others. Each instance is thus transformed from its original feature space to a discriminative subspace defined by the output of these classifiers. An IL classifier is then learned in this space based on a margin-based loss which penalizes the bags violating the margin constraints. For each bag, the BL prediction is obtained by aggregating (pooling) the decisions of its instances. An overview of the proposed approach is illustrated in Figure 2. In this section linear classifiers were used for the sub-category and for the IL classifier(s) due to their advantages (simplicity, easy to learn, prone to overfitting, etc.) over the non-linear ones.

Symbols	Definition
B_i	a bag (image)
$y_i \in \{-1, 1\}$	label of B_i
$\mathbf{x}_{ij} \in \mathbb{R}^d$	feature representation of an instance (image patch)
$\mathcal{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K] \in \mathbb{R}^{d \times K}$	a set of sub-category classifiers
$\boldsymbol{\mu}_k \in \mathbb{R}^d$	k^{th} sub-category classifier
q_{ijk}	the probability of $\mathbf{x}_{ij}$ belonging to the k^{th} sub-category vs rest
$\mathbf{z}_{ij} = [q_{ij1}, \ldots, q_{ijk}] \in \mathbb{R}^K$	instance representation in the discriminative space
$\mathbf{w} \in \mathbb{R}^K$	IL classifier
p_{ij}	IL probability
P_i	BL probability
r	pooling parameter
γ	margin parameter
$N_+ (N_-)$	the number of positive (negative) bags in the training set

TABLE I: Main symbols introduced and their definitions

B. Sub-category classifiers for MIL

Let the training dataset contain $\{(B_i, y_i)\}_{i=1}^N$, where B_i is the i^{th} bag (image), $y_i \in \{-1, 1\}$ is its label, and N is the number of bags. Each bag B_i consists of N_i instances (image patches), so that $B_i = \{\mathbf{x}_{ij}\}_{j=1}^{N_i}$, where $\mathbf{x}_{ij} \in \mathbb{R}^d$ is the feature representation of the j^{th} instance of the i^{th} bag.

Let $\mathcal{M} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K] \in \mathbb{R}^{d \times K}$ be a set of sub-category classifiers, where each classifier is learned to separate a particular sub-category from others. The probability of an instance $\mathbf{x}_{ij}$ belonging to the k^{th} sub-category vs rest can be given as

$$q_{ijk} = \frac{1}{1 + \exp(-\boldsymbol{\mu}_k^T \mathbf{x}_{ij})}. \tag{1}$$

The new instance-representation $\mathbf{z}_{ij}$ in the discriminative sub-space is defined by the outputs of these sub-category classifiers, i.e.

$$\mathbf{z}_{ij} = [q_{ij1}, \ldots, q_{ijK}, 1]. \tag{2}$$

Let $\mathbf{w} \in \mathbb{R}^{K+1}$ define the instance-level classifier which is learned in this discriminative subspace, and the probability of the instance $\mathbf{x}_{ij}$ belonging to the positive class, p_{ij}, can be given as

$$p_{ij} = \sigma(\mathbf{w}^T \mathbf{z}_{ij}). \tag{3}$$

where $\sigma(x) = 1/(1 + \exp(-x))$.

The BL probability, P_i, of a bag B_i can be obtained by aggregating (pooling) the probabilities of the instances inside the bag. In this work, we use the *generalized-mean* operator for aggregation, although other pooling operators can be used ([26]).

$$P_i = \left(\frac{1}{N_i} \sum_{j=1}^{N_i} p_{ij}^r \right)^{1/r}, \tag{4}$$

where r is a pooling parameter. When $r = 1$, Eq. (4) becomes *average-pooling*, and large r values (e.g. $r = 100$) approximate *max-pooling*.

The set of the sub-category classifiers ($\mathcal{M}$), the pooling parameter (r), and the IL classifier ($\mathbf{w}$) can be learned using a loss function defined at the BL. In this work we propose a margin-based loss function for this purpose and compare it with the widely-used cross-entropy loss.

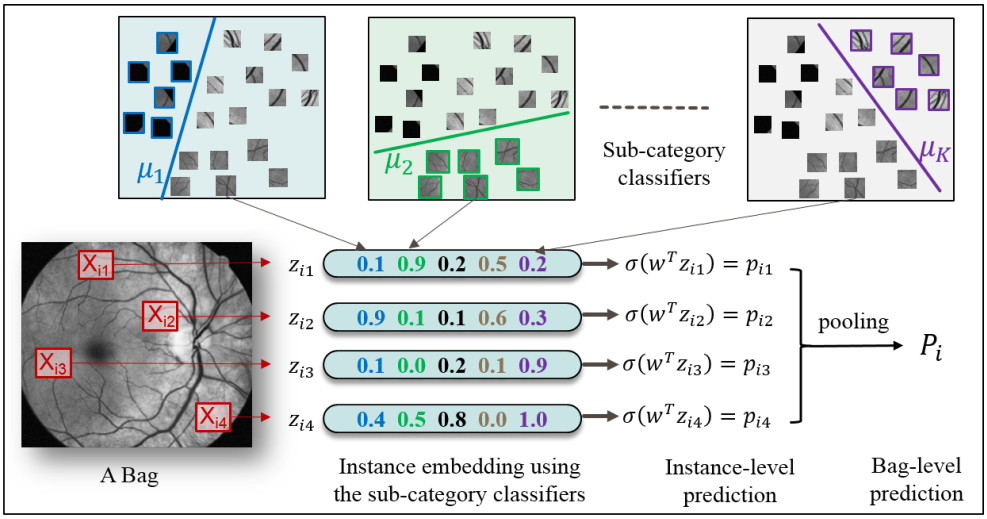


Fig. 2: Overview of the proposed approach, where the set of sub-category classifiers ($[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K]$), the instance-level classifier ($\mathbf{w}$) and the pooling parameter (r) can be learned from weakly-labelled training data.

1) Cross-entropy loss: The cross-entropy loss function can be defined as

$$\mathcal{L}_c(r, \mathcal{M}, \mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 - \frac{\lambda}{N_+}\sum_{i:y_i=1}\log(P_i) - \frac{\lambda}{N_-}\sum_{i:y_i=-1}\log(1-P_i) \tag{5}$$

where $P_i = P_i(y_i = 1|B_i, r, \mathcal{M}, \mathbf{w})$, λ controls the trade-off between regularization (first term) and cross-entropy terms (last two terms), and N_+, N_- are the total number of positive and negative bags in the training set respectively. Note that this loss is widely used by the existing MIL approaches, e.g. [17], [26], [28], [29], [31].

2) Margin-based loss: Margin-based loss has some advantages over the cross-entropy loss for classification problems [33]. First, it improves the accuracy of the training data by focussing on the wrongly classified images, instead of making the correct predictions more accurate (as in cross-entropy loss). Second, since margin-based loss maximize the margin between two classes, overfitting can be avoided, leading to a better generalization. Third, it improves training speed, as model updates are only based on the images classified wrongly; the ones classified correctly will not contribute to the model updates, and can be avoided altogether in derivative calculations.

Therefore, we propose the following margin-based loss function, which penalizes the bags violating the margin defined by the parameter γ:

$$\begin{aligned} &\arg\min_{r,\mathcal{M},\mathbf{w}} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda\sum_{i=1}^{N}\xi_i \\ &\text{s.t. } y_i(P_i - 0.5) \geq \gamma - \xi_i, \quad \xi_i \geq 0, \ \forall i \end{aligned} \tag{6}$$

where $\gamma \in (0, 0.5]$ is a tunable margin parameter, ξ_i are the slack variables associated with the misclassified bags, and λ determines the relative importance of the regularization and the misclassification errors.

The functional in Eq. (6) can be transformed to a single objective function without constraints as below,

$$\begin{aligned} \mathcal{L}_m(r, \mathcal{M}, \mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{\lambda}{N_+}\sum_{i:y_i=1}\mathcal{L}_i(y_i, B_i, r, \mathcal{M}, \mathbf{w}) \\ + \frac{\lambda}{N_-}\sum_{i:y_i=-1}\mathcal{L}_i(y_i, B_i, r, \mathcal{M}, \mathbf{w}) \end{aligned}$$
$$\text{where,} \quad \mathcal{L}_i(y_i, B_i, r, \mathcal{M}, \mathbf{w}) = \max\left[0, \ \gamma + y_i(0.5 - P_i)\right]^2 \tag{7}$$

In Eq. (7), each bag B_i falls into one of the two categories. It lies on the margin or beyond the margin if $y_i(P_i - 0.5) \geq \gamma$, in which case B_i will be correctly classified and it will not contribute to the loss defined in Eq.(7). On the other hand, if B_i lies within the margin ($0.5 - \gamma \leq P_i \leq 0.5 + \gamma$) it will be classified wrongly. Since this cost function maximizes the margin between positive and negative bags (in the probability space) and the model updates are only based on the misclassified bags, we expect a good generalization and reduced time for optimization (compared to cross-entropy loss).

Relations to cross-entropy: When $\gamma = 0.5$ the cost function defined by Eq. (7) can be rewritten as

$$\mathcal{L}_m(r, \mathcal{M}, \mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{\lambda}{N_+}\sum_{i:y_i=1}(1-P_i)^2 + \frac{\lambda}{N_-}\sum_{i:y_i=-1}P_i^2 \tag{8}$$

This function makes the probabilities of the positive bags closer to 1 and the negative bags closer to 0. In other words, it maximizes the probabilities of the positive bags while minimizing the probabilities of the negative ones. Hence its objective is very similar to the objective of the cross-entropy loss (Eq. (5)). Setting γ to a larger value (e.g. $\gamma = 0.5$, Eq. (8)) leads to best probability estimates, however, it may give lower classification accuracy as the cost function focuses on improving the probabilities. Smaller γ values (e.g. $\gamma = 0.3$, Eq.(7)) lead to better classification accuracy, but they may not result in best probability estimates as the cost function concentrates on improving the classification accuracies.

Algorithm 1: SCC-MIL: Learn parameters

Input: training data $\{I_i, y_i\}_{i=1}^N$, no. of sub-categories K
Output: $\mathcal{M}, r, \mathbf{w}$
Initialize $\mathcal{M}, \mathbf{w}$ and r, costPre = ∞
while *max no of iterations not reached* **do**
  $\mathbf{w} \leftarrow$update $\mathbf{w}$ while fixing $\mathcal{M}$ and r constant.
  $\mathcal{M} \leftarrow$update $\mathcal{M}$ while fixing $\mathbf{w}$ and r constant.
  $r \leftarrow$update r while fixing $\mathbf{w}$ and $\mathcal{M}$ constant.
  cost = calculate cost using Eq. (7)
  if $|\text{cost} - \text{costPre}| < \epsilon$ **then** return
  else costPre = cost
end

C. Initialization and optimization

Since there are three variables ($\mathcal{M}$, $\mathbf{w}$ and r) to be learned in the cost functions defined in Eq. (5) and (7), we use a coordinate descent method, where we learn one variable at a time while keeping others constant. We use the L-BFGS algorithm [35] for the alternate minimization as it reported to be faster than stochastic gradient descent [36]. Algorithm 1 describes the alternating optimization, where the maximum number of iterations was fixed to 25 and ϵ was set to $\epsilon = 10^{-5}$.

The derivatives of P_i with respect to the variables $\{\boldsymbol{\mu}_k\}, r$ and $\mathbf{w}_k$ can be given as:

$$\frac{\partial P_i}{\partial \mu_k} = \frac{P_i}{\sum_{j=1}^{N_i} p_{ij}^r} \sum_{j=1}^{N_i} p_{ij}^r (1 - p_{ij}) w_k q_{ijk} (1 - q_{ijk}) \mathbf{x}_{ij} \quad (9)$$

$$\frac{\partial P_i}{\partial r} = \frac{P_i}{r} \left[\frac{1}{\sum_{j=1}^{N_i} p_{ij}^r} \sum_{j=1}^{N_i} p_{ij}^r \log(p_{ij}) - \log P_i \right] \quad (10)$$

$$\frac{\partial P_i}{\partial \mathbf{w}} = \frac{P_i}{\sum_{j=1}^{N_i} p_{ij}^r} \sum_{j=1}^{N_i} p_{ij}^r (1 - p_{ij}) \mathbf{z}_{ij} \quad (11)$$

The derivative of the cost functions $\mathcal{L}_c$ (Eq. (5)) and $\mathcal{L}_m$(Eq. (7)) with respect to $\mathbf{w}_k$ can be given as:

$$\begin{aligned} \frac{\partial \mathcal{L}_c}{\partial \mathbf{w}} &= \lambda \mathbf{w} - \frac{1}{N_+} \sum_{i:y_i=1} \frac{1}{P_i} \frac{\partial \mathcal{P}_i}{\partial \mathbf{w}} + \frac{1}{N_-} \sum_{i:y_i=-1} \frac{1}{1 - P_i} \frac{\partial \mathcal{P}_i}{\partial \mathbf{w}} \\ \frac{\partial \mathcal{L}_m}{\partial \mathbf{w}} &= \lambda \mathbf{w} + \frac{1}{N_+} \sum_{i:y_i=1} \frac{\partial \mathcal{L}_i}{\partial \mathbf{w}} + \frac{1}{N_-} \sum_{i:y_i=-1} \frac{\partial \mathcal{L}_i}{\partial \mathbf{w}} \\ \text{where, } \frac{\partial \mathcal{L}_i}{\partial \mathbf{w}} &= -2 y_i \max\left[0, \gamma + y_i (0.5 - P_i)\right] \frac{\partial P_i}{\partial \mathbf{w}} \end{aligned} \quad (12)$$

The derivative of the cost functions with respect to other variables can be computed in a similar manner.

Since the original cost function is non-convex the solution may depend on the initialization. Therefore we propose the following method to initialize $\mathcal{M}$. First the instances from the training set are clustered using k-means with dictionary size K, and a set of one-vs-rest linear SVM classifiers $\mathbf{u}_k$ $(k = 1, \ldots, K)$ are learned to separate each cluster c_k from the rest. To convert each of the binary SVM classifier into a probabilistic model we fit a sigmoid function (Eq. (13)) as explained by Platt et al. [37].

$$\begin{aligned} P(y_{ij} = 1 | \mathbf{u}_k, \mathbf{x}_{ij}) &= \frac{1}{1 + \exp^{(a_k \mathbf{u}_k^T \mathbf{x}_{ij} + b_k)}} \\ \text{where, } y_{ij} &= \begin{cases} 1, & \text{if } \mathbf{x}_{ij} \in c_k \\ -1, & \text{otherwise} \end{cases} \end{aligned} \quad (13)$$

The parameters a_k and b_k are learned using publicly available code [2]. The sub-category classifiers are then initialized as $\boldsymbol{\mu}_k = [a_k \mathbf{u}_{k1}, \ldots, a_k \mathbf{u}_{kd}, b_k]$. $\mathbf{w}$ was initialized to zero. Refer to Section IV-A6 for experiments with different initializations.

	A1	A2
no of images in which RNFL is visible (+)	696	728
no of images in which RNFL is invisible (-)	188	156

TABLE II: RNFL dataset.

IV. Experiments

A. RNFL visibility classification

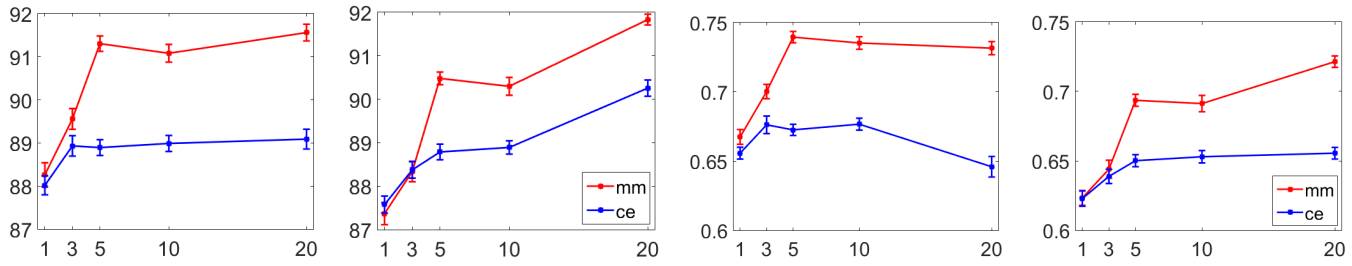
1) Dataset: This dataset contains 884 high-quality color fundus images collected from the Tayside diabetic retinopathy screening programme in Scotland via the GoDARTS bioresource [3]. Images were obtained in accordance to the current regulations for clinical studies (ethics, Caldicott, anonymization). We do not include patient characterization as we do not compute associations with clinical parameters, but only test RNFL visibility. Each image in this dataset was independently annotated by two practising ophthalmologists (A1 and A2, A1 the more experienced), who provided binary (RNFL visible or not) image-level annotations by analysing the red-free (green and blue channels) images. Table II shows the statistics of the annotations.

2) Instance representation: We resized the images preserving their aspect ratio so that their maximum dimension (row or column) was 1000 pixels. As the RNFL is observed clinically in red-free FC images, we considered the green and the blue channel for processing. The contrast of the green channel was enhanced using an adaptive histogram-equalization method [38]. We found that enhancing contrast of the blue channel led to inferior performance, therefore no pre-processing was applied on the blue channel. Instances (square image patches) of size 200×200 pixels with an overlap of 100 pixels were extracted from each color channel independently, leading to $\sim 150\,(75 \times 2)$ instances per image (bag). Inside each instance, SIFT features [4] (patch size of 24×24 pixels, overlap 16 pixels) were computed and encoded using Locality Constrained Linear Coding (LLC) [39], with a dictionary size of 500. Average-pooling was applied to get a feature representation for each instance. The open source library *vlfeat* [40] was used for SIFT feature extraction and dictionary learning. The public code [39] was used for LLC encoding. We used the *L2-and-power* normalisations [41] to normalize the representation of each instance.

[2] http://www.work.caltech.edu/~htlin/program/libsvm/doc/platt.m

[3] http://medicine.dundee.ac.uk/godarts

[4] refer the supplementary material for experiments with different features. Supplementary materials are available in the supplementary files /multimedia tab.



(a) K vs $\mathcal{P}_1$ **(b)** K vs $\mathcal{P}_2$ **(c)** K vs $\mathcal{K}_1$ **(d)** K vs $\mathcal{K}_2$

Fig. 3: Comparison of margin-based and cross-entropy loss functions with different number of subcategories (K in horizontal axis, vertical bars represent the standard errors.)

3) Experimental settings and evaluation measures: We used a 3 fold cross-validation repeated 3 times to evaluate the performance of different approaches, and report the mean and the standard errors of the performance measures obtained over these experimental runs. In each experimental run the training images with consensus labels from the annotators were used for training.

Table III shows an example confusion matrix based on system-annotator agreements. We use the following evaluation measures to compare the system with the annotators.

- Accuracy ($\mathcal{P}_i$): the percentage of the images agreed by an annotator, A_i, and the system. $\mathcal{P}_i = (a+d)/n \times 100\%$
- Kappa values ($\mathcal{K}_i$) [42] for an annotator, A_i, vs system agreements: $\mathcal{K}_i = (p_o - p_e)/(1-p_e)$, where p_o is the proportion of the observed agreements, and p_e is the proportion of the expected agreements, defined as $p_o = (a+d)/n$, and $p_e = (f_1 g_1 + f_2 g_2)/n^2$.

		Annotator		total
		+	-	
System	+	a	b	g_1
	-	c	d	g_2
total		f_1	f_2	n

TABLE III: Example confusion matrix.

4) Margin-based vs cross entropy loss function: This section compares the proposed margin-based loss function (Eq. (6)) with the widely used cross-entropy loss function (Eq. (5)), and shows experimentally that the margin-based loss function gives better system-annotator agreements compared to the cross-entropy one.

Figure 3 reports the system-annotator agreements for different number of sub-categories (K). Overall, increasing K improves the classification accuracy ($\mathcal{P}$) regardless of the loss functions used. However, when $K = 20$ the cross-entropy loss function gives lower system-annotator kappa values compared to the values obtained for smaller K ($K < 20$). This may be due to overfitting. In all cases the proposed margin-based loss performs similarly (for $K \leq 3$) or considerably better (for $K > 3$) to the cross-entropy loss function as it directly maximizes the classification accuracy. When $K = 20$, the proposed margin-based loss gives a system-annotator agreement of $\mathcal{P}_1 = 91.6 \pm 0.19$ and $\mathcal{K}_1 = 0.732 \pm 0.004$ with A1, and $\mathcal{P}_1 = 91.8 \pm 0.12$ and $\mathcal{K}_1 = 0.721 \pm 0.004$ with A2, which is similar to the inter-annotator agreement on the entire dataset ($\mathcal{P} = 91.63$ and $\mathcal{K} = 0.73$). Figure 8 shows some region-level predictions by the proposed approach.

In this experiment the parameter γ (Eq. (6)) was set to $\gamma = 0.3$ (see Section IV-A5 for the effect of the γ values), the pooling parameter r (Eq. (4)) was set to $r = 3$ (see Section IV-A6 for the effect of the r values), and the regularization parameter λ (Eq. (6) and Eq. (5)) was determined based on applying a 3-fold cross-validation on the training set of each experimental run.

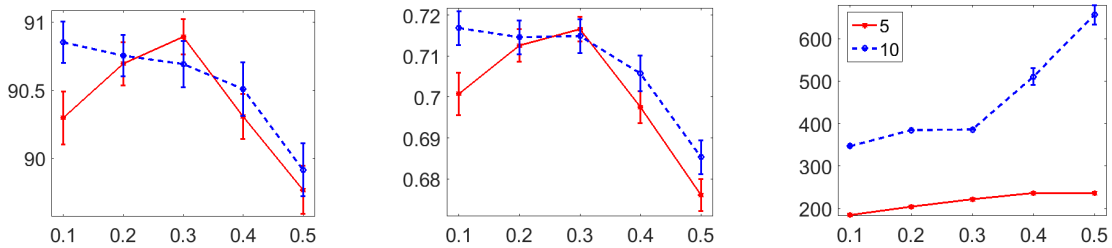
5) Effect of γ in the proposed margin-based loss function: This section compares the system-annotator agreements with different margin parameter (γ in Eq. (6)).

Figure 4 and Figure 5 respectively report the averaged system-vs-annotator agreements (averaged over A1 and A2) and the probability distributions obtained for training and testing sets for different γ values. Note that the probability distributions in Figure 5 was aggregated over all the experimental runs and are based on the images with consensus ground truth. For both training and testing, increasing γ from 0.1 to 0.5 leads to improved probability distributions. The best probability distributions were obtained for $\gamma = 0.5$. However, $\gamma = 0.5$ also gives lower classification performance as it maximizes the probability outputs on the training set, instead of directly minimizing the classification errors. Note that when $\gamma = 0.5$ the margin-based loss function becomes similar to the cross-entropy loss (see Section III-B2 for discussion). Figure 4 also reports the computational time (on a core i7 CPU with 32GB RAM and Matlab 2015a) required to optimize Eq. (6). Small γ values not only give better agreements, but also take less computational time compared to larger values.

As can be seen from Figures 4 and 5, $\gamma = 0.3$ is a better compromise between a good classification accuracy and reasonable probability estimates. Therefore in the following sections we fix γ to $\gamma = 0.3$ with the margin-based loss function.

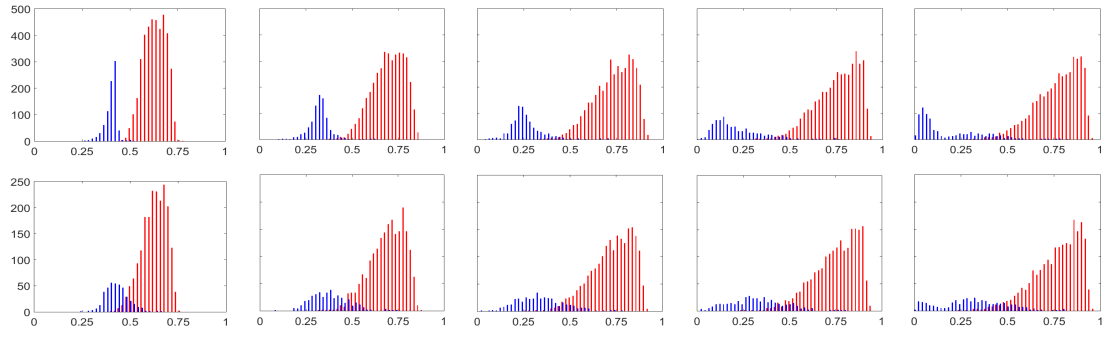
In this experiment, following the previous experiment r was fixed to $r = 3$ and the λ parameter was learned by applying a 3-fold cross validation on the training set of each experimental run.

6) Effect of initializations: In the proposed system two variables, the pooling parameter r and the sub-category classifiers $\mathcal{M}$, need to be initialized. This section investigates how different initializations affect the system-annotator agreements.



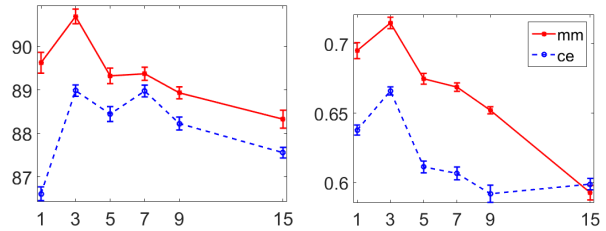
(a) γ vs $\frac{\mathcal{P}_1+\mathcal{P}_2}{2}$ **(b)** γ vs $\frac{\mathcal{K}_1+\mathcal{K}_2}{2}$ **(c)** Computational time in sec.

Fig. 4: Effect of γ (horizontal axis) in the proposed margin-based loss function (Eq. (6)) for different number of sub-categories ($K = 5$ and $K = 10$). Vertical bars represent the standard errors.



$\gamma = 0.1$ $\gamma = 0.2$ $\gamma = 0.3$ $\gamma = 0.4$ $\gamma = 0.5$

Fig. 5: Probability distributions for the training (top row) and testing (bottom row) sets (visible-red, invisible-blue) for different γ values with $K = 10$. $\gamma = 0.5$ leads to better probability estimates, and lower classification performance compared to other γs.

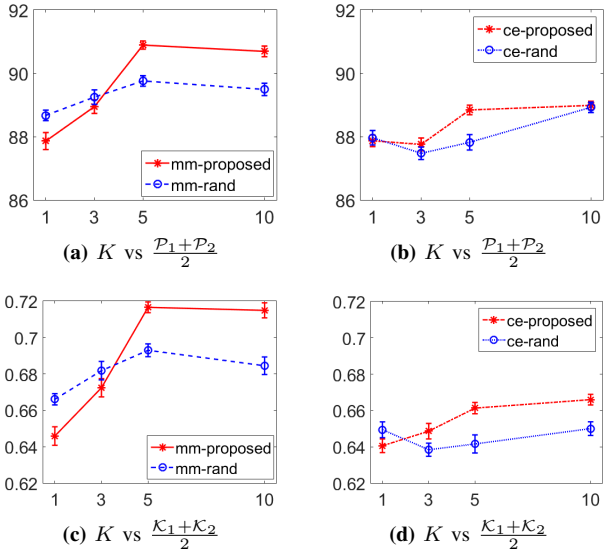


(a) r vs $\frac{\mathcal{P}_1+\mathcal{P}_2}{2}$ **(b)** r vs $\frac{\mathcal{K}_1+\mathcal{K}_2}{2}$

Fig. 6: Effect of initialization for r values (horizontal axis) for the margin-based and the cross-entropy loss functions ($K = 10$).

Figure 6 reports the system-annotator agreements for different initializations of r. Regardless of the loss function used, $r = 3$ gives the best agreements. When r takes a high value (e.g. $r = 15$) image-level predictions will be approximated by one or few top region-level (instance-level) predictions, as larger r values approximate max-pooling. This may lead to noisy image-level level labels, as some noisy regions can easily affect the image-level probability. When $r = 1$ all the regions including the background will contribute to the image-level predictions, as when $r = 1$ the pooling function will be equal to sum-pooling. Therefore $r = 3$ is a good compromise between max-pooling or sum-pooling to determine the image-level predictions.

From the experiments we noticed that the final value of r does not change significantly from its initialization. However, learning r gives improved performance than fixing it. For example, in one of this experimental run r converges to $r = 3.58$ when it is initialized to $r = 3$.



(a) K vs $\frac{\mathcal{P}_1+\mathcal{P}_2}{2}$ **(b)** K vs $\frac{\mathcal{P}_1+\mathcal{P}_2}{2}$

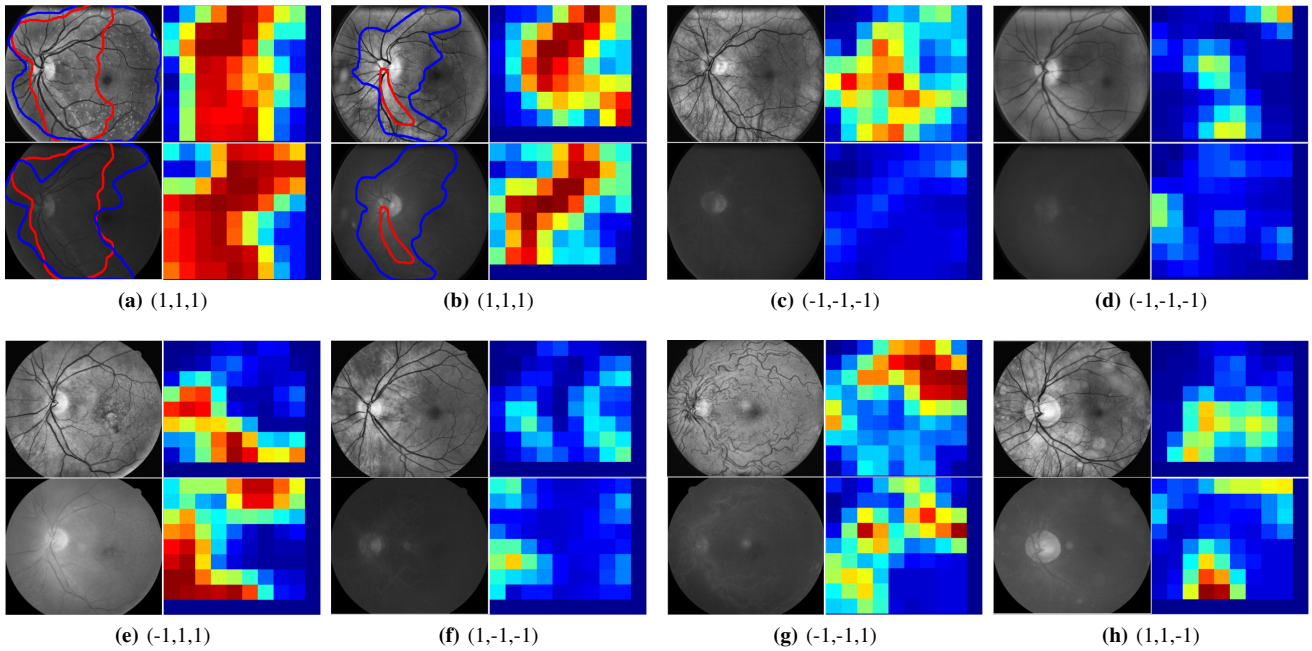
(c) K vs $\frac{\mathcal{K}_1+\mathcal{K}_2}{2}$ **(d)** K vs $\frac{\mathcal{K}_1+\mathcal{K}_2}{2}$

Fig. 7: Effect of initialization for the sub-category classifiers: Proposed vs random initialization for the sub-category classifiers with the margin-based (first column) and the cross-entropy (last column) loss functions (number of sub-categories vs system-annotator agreements).

Figure 7 reports the system-annotator agreements for the proposed and the random initializations for the sub-category classifiers. For both loss functions the proposed initialization gives better system-annotator agreements compared to random

Method	System vs A1		System vs A2		Average training time (in sec)
	$\mathcal{A}$	$\mathcal{K}$	$\mathcal{A}$	$\mathcal{K}$	
mi-SVM	86.71 ± 0.46	0.6283 ± 0.0137	86.05 ± 0.53	0.5900 ± 0.0141	157 ± 2
MI-SVM	88.52 ± 0.62	0.6545 ± 0.0152	89.22 ± 0.46	0.6531 ± 0.0119	43 ± 16
L-MIL-SVM	89.83 ± 0.65	0.6773 ± 0.0210	91.05 ± 0.30	0.6929 ± 0.0107	7 ± 1
MIL-BOOST	89.55 ± 0.77	0.7030 ± 0.0215	88.82 ± 0.63	0.6637 ± 0.0191	111 ± 1
MCIL-Boost	88.62 ± 0.93	0.6850 ± 0.0233	87.47 ± 0.79	0.6350 ± 0.0200	1735 ± 14
SCC-MIL (proposed, K=5)	91.30 ± 0.17	0.7395 ± 0.0123	90.47 ± 0.14	0.6936 ± 0.0129	221 ± 1
SCC-MIL (proposed, K=20)	91.61 ± 0.19	0.7321 ± 0.0041	91.82 ± 0.12	0.7212 ± 0.0040	1200 ± 350

TABLE IV: Different MIL approaches and their agreements ($\mathcal{P}$ and $\mathcal{K} \pm$ standard error) with different annotators (A1 and A2) for RNFL visibility (image-level) classification. Note that the inter-annotator agreement on the entire dataset is $\mathcal{P} = 91.63\%$ and $\mathcal{K} = 0.73$.



(a) (1,1,1) **(b)** (1,1,1) **(c)** (-1,-1,-1) **(d)** (-1,-1,-1)

(e) (-1,1,1) **(f)** (1,-1,-1) **(g)** (-1,-1,1) **(h)** (1,1,-1)

Fig. 8: Region-level predictions for example test images by the SCC-MIL: Both the annotators and the system annotate images as RNFL visible (a-b), and invisible (c-f), one annotator annotates as RNFL visible and the other one as RNFL invisible (e-f), images where system disagrees with the annotators (g-h). Under each sub-figure annotations by A1, annotations by A2 and the system's image-level predictions are respectively given inside the brackets. System predictions for the green (top-left) and the blue (bottom-left) channels of each image is given in the right hand side (top-right and bottom-right) of each sub-figure. In (a-b) the traced regions indicate the expert region-level annotations, red by A1 and blue by A2. In the system predictions, red and blue colors respectively indicate high and low RNFL visibility.

initializations when K is large ($K > 5$). It can be also noted that the margin-based loss performs considerably better than the cross-entropy loss regardless of initialization.

With the random initialization, the sub-category classifiers were initialized by sampling from a Gaussian distribution with zero-mean and a standard deviation of 0.1. When $K = 1$ the proposed initialization initializes the sub-category classifier with the mean of all the instance-level features.

7) Comparison with other MIL approaches: In this section we compare SCC-MIL with other approaches for RNFL visibility classification and show that SCC-MIL performs considerably better than others.

We used the public code from [18] for MIL-Boost and MCIL-Boost. We implemented mi-SVM and MI-SVM following [22]. For all the baselines we take care to select the parameters guaranteeing the fairest possible comparison. We also implemented a latent version of MIL-SVM (Eq. (14), L-MIL-SVM), which is similar to the Latent-SVM proposed in [43]. It can be written as:

$$\mathcal{L}_l(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{\lambda}{N_+}\sum_{i:y_i=1}\max(0,\ 1 - f_w(B_i))^2 + \frac{\lambda}{N_-}\sum_{i:y_i=-1}\max(0,\ 1 + f_w(B_i))^2 \quad (14)$$

$$\text{where,} \quad f_w(B_i) = \max_{\mathbf{x}_i \in B_i}(\mathbf{w}^T\mathbf{x}_i)$$

We initialize $\mathbf{w}$ in Eq. (14) by learning a SVM classifier which separates all the instances in positive bags from all the instances in the negative bags. We use the L-BFGS [35] algorithm to optimize Eq. (14), where at each iteration, for each bag B_i, we calculate $f_w(B_i)$ based on the instance which gives the highest score. λ was learned based on applying a 3-fold cross validation on the train set of each iteration.

For MIL-Boost and MCIL-Boost we select the pooling parameter $r \in \{1, 3, 5, 7, 20, 100\}$ such that it gives the best kappa value on a subset of the entire dataset. We found that $r = 5$ is the best choice overall.

Dataset	no of images			Exp. setup
	positive	negative	total	
Messidor [7]	654	546	1200	10 times 2-FCV
DR [8]	265	160	425	fixed train($\frac{2}{3}$)-test($\frac{1}{3}$) split
UCSB Breast cancer [9]	26	32	58	10 times 4-FCV

TABLE V: Datasets and experimental settings (FCV-fold cross validation).

Method	Acc.
MI-SVM [22]	54.5
EMDD	55.1
SIL-SVM	58.4
GP-MIL	59.2
citation k-NN	62.8
MILBoost	64.1
mi-Graph [30]	72.5
Ours	**72.8**

(a) Messidor dataset [7]

Method	Acc.
DD	61.29
mi-SVM [22]	70.32
MILES	71.00
EMDD	73.50
citation k-NN	78.70
SNPM [8]	81.30
mi-Graph [30]	83.87
Ours	**87.93**

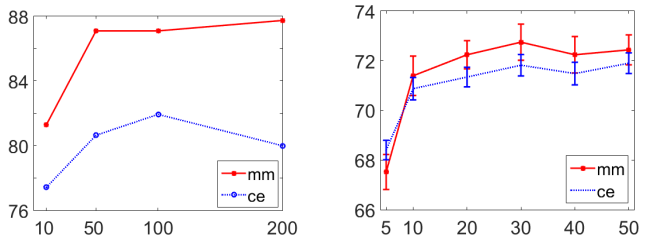
(b) DR dataset [8]

Method	AUC
MILBoost	0.83
GPMIL [9]	0.86
MI-SVM [22]	0.90
RGPMIL [9]	0.90
BRT [17]	0.93
mi-Graph [30]	0.946
JC^{2}MIL [28]	0.95
Ours	**0.967**

(c) UCSB cancer [9]

TABLE VI: Results on the public datasets. All the results except ours and mi-Graph were copied from [7–9]. Some references were omitted due to space. Different evaluation measures were used as they were reported in [7–9].

Table IV reports the results. Our approach gives better agreement with the annotators compared to all other approaches even with small K ($K = 5$). Although MCIL-Boost is also designed to capture sub-category information with a boosting classifier, it is not a EB approach (Section II) and considers the sub-categories present in the positive bags only. Our experiments shows that MCIL-Boost gives lower performance compared to MIL-Boosting. Our approach performs considerably better than all the approaches considered and gives state-of-the-art results on the RNFL dataset.



(a) Messidor dataset **(b)** DR dataset

Fig. 9: Number of sub-categories (horizontal axis) vs accuracy (vertical axis) using SCC-MIL with different loss functions [5].

B. Experiments with public medical image datasets

The following sections explain the three public datasets and the experiments based on them. The experimental settings for these datasets are summarized in Table V.

(1) Messidor [7]: A public diabetic retinopathy screening dataset, contains 1200 eye fundus images, where 654 images are from diseased eyes and 546 images are from healthy eyes. This dataset is well studied in [7] for BL classification. Each image was rescaled to 700×700 pixels and split into 135×135 regions. Each region was represented by a set of features including intensity histograms and texture.

(2) The diabetic retinopathy (DR) screening dataset [8]: 425 FC images where 160 are normal and 265 are from diabetic retinopathy patients. This dataset was constructed from 4 publicly available datasets (DiabRetDB0, DiabRetDB1, STARE and Messidor). Each image is represented by a set of 48 instances.

(3) UCSB breast cancer [9]: 58 TMA H&E stained breast images (26 malignant, 32 benign). Used in [9], [17], [28] to compare different MIL approaches; each image was divided into 49 instances, and each instance is represented by a 708-dimensional feature vector which includes SIFT and local binary patterns.

For fair comparison we use the features made publicly available [6], and follow the same experimental set-up used by the existing approaches. The features were normalized such that each feature dimension will have zero mean and unit variance.

Figure 9 compares margin-based and cross-entropy losses with public datasets. The proposed margin-based loss performs similar or better than the widely used cross-entropy loss, and increasing the number of subcategories improves the classification, although the classification performance saturates for larger K values. Table VI reports the comparative results with other MIL approaches on the public datasets. With Messidor, our approach gives a competitive accuracy of 72.8% (with a standard error of ± 0.4) compared to the accuracy obtained by mi-Graph, which however cannot provide IL predictions as a BL approach. With DR, our approach improves the state-of-the-art accuracy by $\sim 4\%$. With UCSB, our approach achieves an AUC of 0.967 with a standard error of 0.007. Our Equal

[5] refer the supplementary material for the results on UCSB cancer dataset. Supplementary materials are available in the supplementary files /multimedia tab.

[6] Messidor and UCSB cancer: http://www.miproblems.org/datasets/; DR: https://github.com/ragavvenkatesan/np-mil

Error Rate was 0.07 ± 0.01, much smaller than the one reported in [17] (0.16 ± 0.03). Note that SNPM [8] and JC^{2}MIL [28] are two recent approaches, which achieve state-of-the-art results on various non-medical MIL datasets. However, our approach beats these two approaches with a considerable margin.

Since the Messidor dataset has a fixed training set it makes easier to cross-validate to learn the best values for the free parameters r, λ and γ. We applied a 3-fold cross validation on the training set to select the best parameters from the following ranges: $r \in \{1, 2, 5\}$, $\lambda \in \{10^2, 10^3, 10^4\}$ and $\gamma \in \{0.1, 0.3, 0.5\}$. The multiple train-test folds for the DR and UCSB cancer datasets (Table V) make the parameter learning (by applying cross-validation on each of the training set) expensive. Therefore we fixed the parameters ($\lambda = 10^2$, $\gamma = 0.1$, and $r = 1$) and report the classification performance for different number of sub-categories.

V. Conclusions

The RNFL thickness and its visibility have been posited as biomarkers for neurodegenerative conditions. We have proposed a novel MIL method to assess the visibility (visible/not visible) of the RNFL in fundus camera images, which would enable screening of large patient volumes considering that large bioresources exist with FC images but without up-to-date OCT scans, and recalling patients is not always feasible or timely. In addition, our approach locates visible RNFL images from image-level training labels.

Experiments suggest that our margin-based loss solution performs better than the cross-entropy loss used by existing EB MIL approaches [17], [28], [29]. Experiments with a local RNFL and 3 public medical image datasets show considerable improvements compared to the state-of-the-art. Future work will address the associations of RNFL visibility with brain features and patient outcome.

References

[1] L. Zangwill and C. Bowd, "Retinal nerve fiber layer analysis in the diagnosis of glaucoma," *Current Opinion in Ophthalmology*, vol. 17, no. 2, pp. 120 – 131, 2006.

[2] K. L. Thomson, J. M. Yeo, B. Waddell, J. R. Cameron, and S. Pal, "A systematic review and meta-analysis of retinal nerve fiber layer change in dementia, using optical coherence tomography," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 1, no. 2, pp. 136 – 143, 2015.

[3] J. B. Jonas and D. Schiro, "Localised wedge shaped defects of the retinal nerve fibre layer in glaucoma," *The British Journal of Ophthalmology*, vol. 78, pp. 285–290, 1994.

[4] H. R. Hedges, R. P. Galves, D. Speigelman, N. R. Barbas, E. Peli, and C. J. Yardley, "Retinal nerve fiber layer abnormalities in Alzheimer's disease," *Acta Ophthalmologica Scandinavica*, vol. 74, no. 3, pp. 271 – 275, 1996.

[5] F. Berisha, G. Feke, C. Trempe, J. McMeel, and C. Schepens, "Retinal abnormalities in early alzheimer's disease," *Invest Ophthalmol Visual Science*, vol. 48, no. 5, pp. 2285–9, 2007.

[6] "Multi-modal retinal biomarkers for vascular dementia: developing enabling image analysis tools," http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/M005976/1, 2015.

[7] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: A benchmarking study," *Computerized Medical Imaging and Graphics*, vol. 42, pp. 44–50, 2015.

[8] R. Venkatesan, P. Chandakkar, and B. Li, "Simpler non-parametric methods provide as good or better results to multiple-instance learning." in *IEEE International Conference on Computer Vision*, 2015.

[9] M. Kandemir, C. Zhang, and F. A. Hamprecht, "Empowering multiple instance histopathology cancer diagnosis by cell graphs," in *Medical Image Computing and Computer-Assisted Intervention*, 2014, pp. 228–235.

[10] S. Manivannan, C. Cobb, S. Burgess, and E. Trucco, "Sub-category classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification," in *Medical Image Computing and Computer-Assisted Intervention*, 2016.

[11] J. Oh, H. Yang, K. Kim, and J. Hwang, "Automatic computer-aided diagnosis of retinal nerve fiber layer defects using fundus photographs in optic neuropathy," *Investigative Ophthalmology and Visual Science*, vol. 56, pp. 2872–9, 2015.

[12] C. Muramatsu, Y. Hayashi, A. Sawada, Y. Hatanaka, T. Hara, T. Yamamoto, and H. Fujita, "Computerized detection of retinal nerve fiber layer defects in retinal fundus images by modified polar transformation and gabor filtering," in *World Congress on Medical Physics and Biomedical Engineering*, vol. 25, 2009, pp. 124–126.

[13] Y. Hayashi, T. Nakagawa, Y. Hatanaka, A. Aoyama, M. Kakogawa, T. Hara, H. Fujita, and T. Yamamoto, "Detection of retinal nerve fiber layer defects in retinal fundus images using gabor filtering," vol. 6514, 2007, pp. 65 142Z–65 142Z–8.

[14] J. Odstrčilík, R. Kolář, V. Harabiš, J. Gazárek, and J. Jan, "Retinal nerve fiber layer analysis via markov random fields texture modelling," in *European Signal Processing Conference*, 2010, pp. 1650–1654.

[15] J. Odstrčilík, R. Kolář, V. Harabiš, J. Gazárek, and J. Jan, "Retinal nerve fiber layer analysis via Markov Random fields texture modelling," in *18th European Signal Processing Conference*, 2010, pp. 504–507.

[16] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 1998, pp. 570–576.

[17] W. Li, J. Zhang, and S. J. McKenna, "Multiple instance cancer detection by boosting regularised trees," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 645–652.

[18] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *IEEE conference on Computer Vision and Pattern Recognition*, 2012.

[19] V. Cheplygina, L. Sørensen, D. M. J. Tax, M. de Bruijne, and M. Loog, "Label stability in multiple instance learning," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 539–546.

[20] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81 – 105, 2013.

[21] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *In Advances in Neural Information Processing Systems*, 2001, pp. 1073–1080.

[22] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems 15*, 2003, pp. 561–568.

[23] J. Ramon and L. D. Raedt, "Multi instance neural networks," in *International Conference on Machine Learning, Workshop on Attribute-Value and Relational Learning*, 2000, pp. 53–60.

[24] P. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems 18*, 2007, pp. 1417–1426.

[25] X. Xinn, "Statistical learning in multiple instance problems," Master's thesis, The University of Waikato, 2003.

[26] O. Z. Kraus, L. J. Ba, and B. J. Frey, "Classifying and segmenting microscopy images using convolutional multiple instance learning," *CoRR*, vol. abs/1511.05286, 2015.

[27] Y. Chen, J. Bi, and J. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[28] K. Sikka, R. Giri, and M. Bartlett, "Joint clustering and classification for multiple instance learning," in *British Machine Vision Conference*, 2015.

[29] A. Ruiz, J. Van de Weijer, and X. Binefa, "Regularized multi-concept MIL for weakly-supervised facial behavior categorization," in *British Machine Vision Conference*, 2014.

[30] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *International Conference on Machine Learning*, 2009, pp. 1249–1256.

[31] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *IEEE Computer Vision and Pattern Recognition*, 2015.

[32] Y. Tang, "Deep learning using support vector machines," in *International Conference on Machine Learning: Challenges in Representation Learning Workshop*, 2013.

[33] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991–2000, 2014.
[34] Z. Huang, J. Li, C. Weng, and C.-H. Lee, "Beyond cross-entropy: Towards better frame-level objective functions for deep neural network training in automatic speech recognition," in *Interspeech*, 2014.
[35] S. M, "minFunc: unconstrained differentiable multivariate optimization in Matlab," http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html, 2005.
[36] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *International Conference on Machine Learning*, 2011, pp. 265–272.
[37] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, no. 3, pp. 267–276, 2007.
[38] K. Zuiderveld, "Graphics gems iv," P. S. Heckbert, Ed., 1994, ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485.
[39] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Computer Vision and Pattern Recognition*, 2010.
[40] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.
[41] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010.
[42] A. Viera and J. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, 2005.
[43] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2017.2653623, IEEE Transactions on Medical Imaging

Sub-Category Classifiers for Multiple-Instance Learning and its Application to Retinal Nerve Fiber Layer Visibility Classification - Supplementary Material

Siyamalan Manivannan, *Member IEEE*, Caroline Cobb, Stephen Burgess, and Emanuele Trucco, *Member IEEE*

A. Effect of λ

This section investigates the system-annotator agreements with different values of the λ parameter (Eq. (5) and (7)).

Figure 1 reports the system-annotator agreements and the required training time for different λ values. For both loss functions, larger λ values lead to increased training time, as larger values focus more on reducing the wrongly classified data points (in the margin-based loss function) or improving the probability estimates of the training data (in the cross-entropy loss). When λ is fixed, the margin-based loss function gives similar (when $\lambda = 10^2$) or better (when $\lambda = 10^3$ or 10^4) agreements with the annotator compared to the cross-entropy one. In all cases, the margin-based loss function takes similar (when $\lambda = 10^2$) or much lower (when $\lambda = 10^3$ or 10^4) computational time for training compared to the cross-entropy one as it only concentrate on correcting misclassifications (see Section III in the paper for discussion).

B. Comparison with bag-level classification approaches

This section compares different bag-level classification methods such as LLC+SVM and CNN with SCC-MIL, and shows that the proposed SCC-MIL performs considerably better than these bag-level approaches for image-level classification. Notice that region-level predictions cannot be obtained using these approaches as these are BL approaches.

LLC+SVM is a supervised linear SVM classifier (*libLinear* [1]) trained on the image-level feature representations obtained by average-pooling the dictionary-encoded (size 500) hand-crafted local features. Different hand-crafted local features such as SIFT, multi-resolution local patterns (mLP [2]), Random Projection (RP [3]), raw patch (RAW-PATCH) were used with LLC+SVM. We also tried different local filter banks such as Schmid Filters (S-FILTER [4]) and Leung-Malik Filters [5] and report the one which gave the best results. We used the public code[1] for these filter banks. To guarantee a fair comparison all the local features were extracted from patches of size 24×24 pixels with an overlap of 16 pixels. Each of the vectorized local patch of dimension 24×24 is projected to a compressed space of dimension 200 using a random projection matrix [3] to get a RP feature.

In addition to LLC+SVM, we also trained a CNN to evaluate its performance on our RNFL dataset, as recently CNN has been widely used for medical image classification [6]. We used the ImageNet (1.2 Million images) trained model "AlexNet" [7] with the Caffe library [8] for this purpose[2]. Data augmentation (horizontally mirrored images, and randomly cropped image regions of size 450×450 from images of size 500×500) was used to fine-tune the CNN using the RNFL dataset. The initial learning rate and the maximum number of iterations were set at 10^{-4} and $10,000$ respectively for fine tuning.

Table I reports the results. SIFT feature performs considerably better than other hand-crafted features as it capture local texture. CNN performs better than all the hand-crafted features as it learns discriminative features at different scales and an image-level classifier jointly. However, compared to LLC+SVM, CNN is computationally expensive to train even on GPUs, needs very large amounts of training data and needs a good initialization for its parameters.

It is interesting to note that SCC-MIL (Table IV in the main paper) gives considerably better performance compared to all of these bag-level approaches as (1) it learns an instance-level classifier which weights the importance of the image regions when making the image-level prediction, and hence the non-discriminative background features can be eliminated from determining the label of the images, and (2) it learns a discriminative sub-space and transforms the original feature space to this discriminative subspace, making the instances more discriminative. The non-discriminative background features can easily make the image-level feature representation obtained by LLC+SVM less discriminative, hence leads to lower performance than SCC-MIL.

C. Visualization of sub-category classifiers' responses

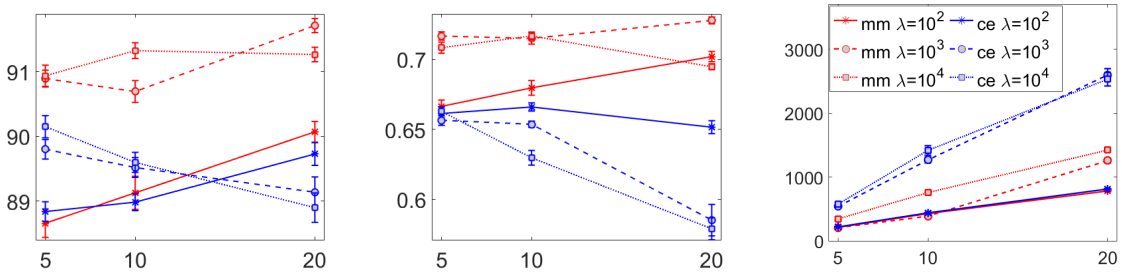
Figure 2 shows the responses of the sub-category classifiers on some example images. The probability maps for RNFL visibility are obtained as a weighted combination of these responses (refer Eq. (3) of the main paper).

D. Experiments with public medical image datasets

Figure 3 compares margin-based and cross-entropy losses with UCSB cancer dataset. Results for other public datasets are given in Figure 9 of the main paper.

[1] http://www.robots.ox.ac.uk/~vgg/research/texclass/filters.html

[2] The Tesla K40 GPU used for this research was donated by the NVIDIA Corporation

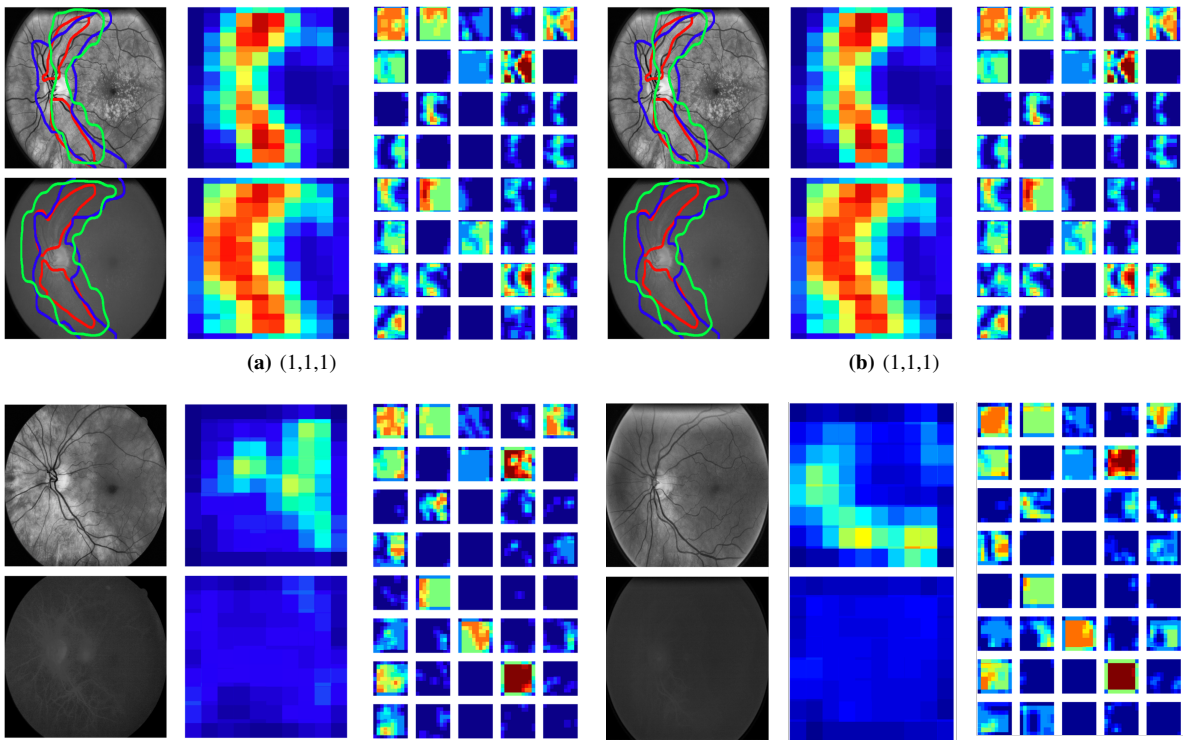


(a) K vs $\frac{\mathcal{P}_1+\mathcal{P}_2}{2}$ **(b)** K vs $\frac{\mathcal{K}_1+\mathcal{K}_2}{2}$ **(c)** K vs time in sec

Fig. 1: Effect of λ: (a,b) K vs system-annotator agreements, and (c) K vs computational time for different λ values and for different loss functions.

Features	A1 vs system		A2 vs system	
	$\mathcal{A}$	$\mathcal{K}$	$\mathcal{A}$	$\mathcal{K}$
CNN	90.69 ± 0.29	0.709 ± 0.008	90.37 ± 0.28	0.677 ± 0.009
LLC+SVM (SIFT)	89.65 ± 0.13	0.706 ± 0.003	87.92 ± 0.13	0.638 ± 0.003
LLC+SVM (mLP [2])	86.10 ± 0.20	0.610 ± 0.005	84.58 ± 0.19	0.541 ± 0.005
LLC+SVM (RP [3])	85.93 ± 0.16	0.611 ± 0.004	84.16 ± 0.16	0.539 ± 0.004
LLC+SVM (RAW-PATCH)	84.77 ± 0.10	0.558 ± 0.003	84.12 ± 0.15	0.512 ± 0.004
LLC+SVM (S-FILTER [4])	86.83 ± 0.10	0.633 ± 0.003	84.64 ± 0.13	0.550 ± 0.004

TABLE I: System vs annotator agreements ($\pm$ standard errors) for different image-level classification approaches.



(a) (1,1,1) **(b)** (1,1,1) **(c)** (-1,-1,-1) **(d)** (-1,-1,-1)

Fig. 2: Four example images (first column of each sub-figure), their probability maps obtained by SCC-MIL (second column of each sub-figure), and the set of sub-category ($K = 20$) classifiers' responses (last column of each sub-figure) are given in this figure. The top-left and the bottom-left corners of each sub-figure show the green and the blue channels of each image respectively. In (a,b) the visible RNFL regions by A1, A2 and the system are traced using red, blue and green colors. In the probability maps red values indicate high RNFL visibility and the blue values indicate low visibility. These probability maps were obtained as a weighted combination of the sub-category classifiers' responses (shown in the last column). Under each sub-figure the image-level annotations by A1, A2, and the system's predictions are respectively given inside brackets.

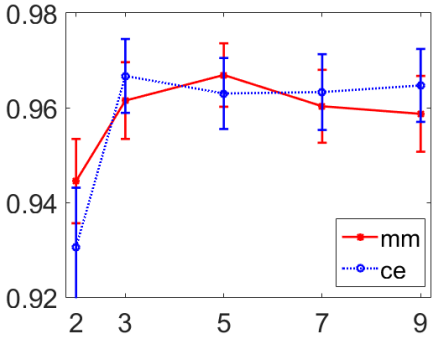


Fig. 3: Number of sub-categories (horizontal axis) vs AUC (vertical axis) for UCSB cancer dataset (results for other public datasets are given in Figure 9 of the main paper).

REFERENCES

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[2] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, and S. J. McKenna, "An automated pattern recognition system for classifying indirect immunofluorescence images of hep-2 cells and specimens," *Pattern Recognition*, vol. 51, pp. 12–26, 2016.

[3] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *ACM Knowledge Discovery and Data Mining*, 2001, pp. 245–250.

[4] C. Schmid, "Constructing models for content-based image retrieval," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, 2001, pp. II–39–II–45 vol.2.

[5] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[6] W. Li, S. Manivannan, J. Zhang, E. Trucco, and S. J. McKenna, "Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks," in *International Symposium on Biomedical Imaging (ISBI)*, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.