Classification of SD-OCT Volumes using Local Binary Patterns: Experimental Validation for DME Detection

Guillaume Lemaître[1,a,*], Mojdeh Rastgoo[1,a,*], Joan Massich[1,*], Carol Y. Cheung[c], Tien Y. Wong[c], Ecosse Lamoureux[c], Dan Milea[c], Fabrice Mériaudeau[1], Désiré Sidibé[1]

[a] *ViCOROB, Universitat de Girona, Campus Montilivi, Edifici P4, 17071 Girona, Spain*
[b] *LE2I UMR6306, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté, 12 rue de la Fonderie, 71200 Le Creusot, France*
[c] *Singaore Eye Research Institute, Singapore National Eye Center, Singapore*

Abstract

Keywords: Diabetic Macular Edema, Optical Coherence Tomography, DME, OCT, Local Binary Patterns (LBP)

1. Introduction

Eye diseases such as Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME) are the most common causes of irreversible vision loss in individuals with diabetes. Just in United States alone, health care and associated costs related to eye diseases are estimated at almost $500 M [2]. Moreover, the prevalent cases of DR are expected to grow exponentially affecting over 300 M people worldwide by 2025 [3]. Early detection and treatment of DR and DME play a major role to prevent adverse effects such as blindness. Indeed, the detection and diagnosis of retinal diseases are based on the detection of vascular abnormalities or lesions in the retina.

In past decades, Computer Aided Diagnosis systems devoted to ophthalmology, have been developed focusing on the automatic analysis of fundus im-

[☆] Document source available in GitHub [1]

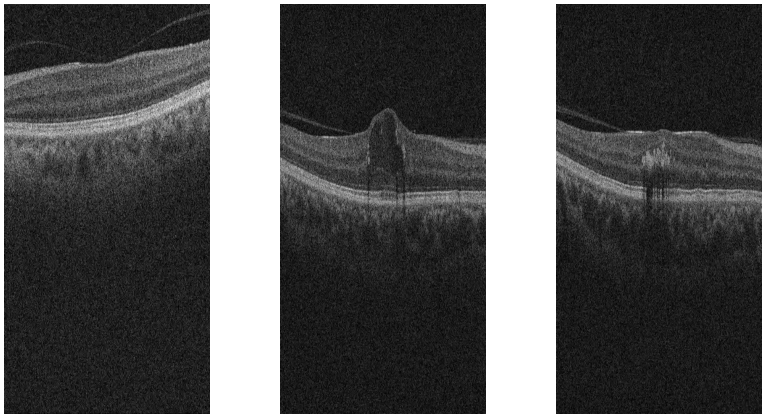
[*] Corresponding author
Email addresses: g.lemaintre58@gmail.com (Guillaume Lemaître), mojdeh.rastgoo@gmail.com (Mojdeh Rastgoo), joan.massich@u-bourgogne.fr (Joan Massich)

ages [4, 5]. However, the use of fundus photography is limited to the detection of signs which are correlated with retinal thickening such as hard and soft exudates, hemorrhages or micro-aneurysms. Moreover, DME is characterized as an increase in retinal thickness within 1 disk diameter of the fovea center with or without hard exudates and sometimes associated with cysts [6]. Therefore, fundus photography cannot always identify the clinical signs of DME; for example cysts, which are not visible in the retinal surface. In addition, it does not provide any quantitative measurements of retina thickness or information about cross-sectional retinal morphology.

Recently, Optical Coherence Tomography (OCT) has been widely used as a valuable diagnosis tool for DME detection. OCT is based on optical reflectivity and produces cross-sectional and three-dimensional images of the central retina, thus allowing quantitative retinal thickness and structure measurements. The new generation of OCT imaging, namely Spectral Domain OCT (SD-OCT) offers higher resolution and faster image acquisition over conventional time domain OCT. SD-OCT can produce $27,000$ to $40,000$ A-scans/seconds with an axial resolution ranging from $3.5\,\mu m$ to $6\,\mu m$ [7]. Figure. 1 shows one normal B-scan and two abnormal B-scans.

Many of the previous works on OCT image analysis have focused on the problem of retinal layers segmentation, which is a necessary step for retinal thickness measurements [8, 9]. However, few have addressed the specific problem of DME and its associated features detection from OCT images.

In this research we focus on the latter problem and propose an automatic framework for identification of DME patients versus normal subjects using OCT volumes. The proposed method, which is an extension of our previous work [10], is based on Local Binary Patterns (LBP) features to describe the texture of OCT images and dictionary learning using the Bag-of-Words (BoW) models [11]. We propose to extract 2D and 3D LBP features from OCT images and volumes, respectively. The LBP descriptors are further extracted from the entire sample or local patches within individual samples. In this research beside the comparison of 2D and 3D features, we also compare the effects of common



(a) Normal (b) DME-cyst (c) DME-exudate

Figure 1: Example of SD-OCT images for normal (a) and DME patients (b)-(c) with cyst and exudate, respectively.

pre-processing steps for OCT data, study the optimal configuration regarding the BoW approach in conjunction with different base classifiers.

This paper is organized as follows, Section 2 presents a summary of the related studies. The proposed framework is explained in Sect. 3, while the experiments and results are discussed in Sect. 4. Finally, the conclusion and avenue for future directions are drawn in Sect. 5.

2. Related Work

This section reviews the works straightly addressing the problem of classifying OCT volumes as normal or abnormal. A summary can be found in Table 1.

Srinivasan *et al.* [12] proposed a classification method to distinguish DME, Age-related Macular Degeneration (AMD) and normal SD-OCT volumes. The OCT images are pre-processed by reducing the speckle noise by enhancing the sparsity in a transform-domain and flattening the retinal curvature to reduce the inter-patient variations. Then, Histogram of Oriented Gradients (HOG) are extracted for each slice of a volume and a linear Support Vector Machines (SVM) is used for classification. On a dataset of 45 patients equally subdivided into the

three aforementioned classes, this method leads to a correct classification rate of 100%, 100% and 86.67% for normal, DME and AMD patients, respectively.

Venhuizen *et al.* proposed a method for OCT images classification using the BoW models [13]. The method starts with the detection and selection of keypoints in each individual B-scan, by keeping the most salient points corresponding to the top 3% of the vertical gradient values. Then, a texton of size 9×9 pixels is extracted around each keypoint, and Principal Component Analysis (PCA) is applied to reduce the dimension of every texton to get a feature vector of size 9. All extracted feature vectors are used to create a codebook using k-means clustering. Then, each OCT volume is represented in terms of this codebook and is characterized as a histogram that captures the codebook occurrences. These histograms are used as feature vector to train a Random Forest (RF) with a maximum of 100 trees. The method was used to classify OCT volumes between AMD and normal cases and achieved an Area Under the Curve (AUC) of 0.984 with a dataset of 384 OCT volumes.

Liu *et al.* proposed a methodology for detecting macular pathology in OCT images using LBP and gradient information as attributes [14]. The method starts by aligning and flattening the images and creating a 3-level multi-scale spatial pyramid. The edge and LBP histograms are then extracted from each block of every level of the pyramid. All the obtained histograms are concatenated into a global descriptor whose dimensions are reduced using PCA. Finally a SVM with an Radial Basis Function (RBF) kernel is used as classifier. The method achieved good results in detection OCT scan containing different pathology such as DME or AMD, with an AUC of 0.93 using a dataset of 326 OCT scans.

As stated in the previous section, our current research is an extension of our previous work [1] with further contributions and evaluations at every stages of our classification framework.

Table 1: Summary of the state-of-the-art methods.

Ref	Diseases			Data size	Pre-processing				Features	Representation	Classifier	Evaluation		
	AMD	DME	Normal		De-noise	Flatten	Aligning	Cropping				Sensitivity (SE)	Specificity (SP)	AUC
[12]	✓	✓	✓	45	✓	✓		✓	HOG		linear-SVM	86.7%,100%,100%		
[13]	✓		✓	384					Texton	BoW, PCA	RF			0.984
[14]	✓	✓	✓	326		✓	✓		Edge, LBP	PCA	SVM-RBF			0.93
[10]		✓	✓	62	✓				LBP-LBP-TOP	PCA, BoW, histogram	RF	87.5%	75%	

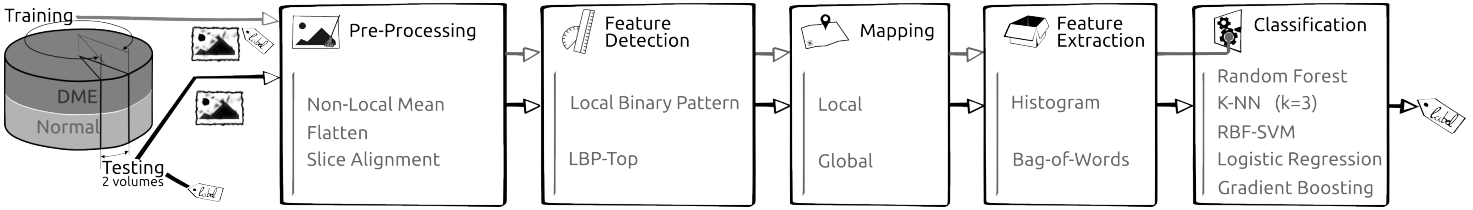
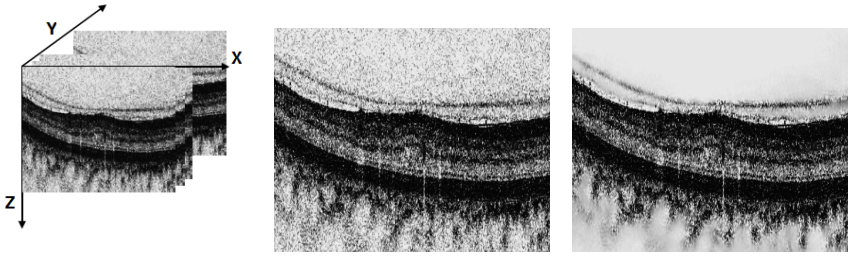


Figure 2: Our proposed classification pipeline.



(a) (b) (c)

Figure 3: OCT: (a) Organization of the OCT data - (b) Original image - (c) NL-means filtering. Note that the images have been negated for visualization purposes.

3. Materials and Methods

The proposed method, as well as, its experimental set-up for OCT volume classification are outlined in Fig. 2. The methodology is formulated as a standard classification procedure which consists of five steps. First, the OCT volumes are pre-processed as presented in details in Sect. 3.1. Then, LBP and LBP-TOP features are detected, mapped and extracted as discussed in depth in Sect. 3.2, Sect. 3.3, and Sect. 3.4, respectively. Finally, the classification step is presented in Sect. 3.5.

3.1. Image pre-processing

This section describes the set of pre-processing techniques which aim at enhancing the OCT volume. The influence of these pre-processing methods and their possible combinations are extensively studied in Sect. 4.4-4.6.

3.1.1. Non-Local Means (NL-means)

OCT images suffer from speckle noise, like other image modalities such as Ultra-Sound (US) [15]. The OCT volumes are enhanced by denoising each B-scan (i.e. each $x - z$ slice) using the NL-means [16], as shown in Fig. 3. NL-means has been successfully applied to US images to reduce speckle noise and outperforms other common denoising methods [17]. NL-means filtering preserves fine structures as well as flat zones, by using all the possible self-predictions that the image can provide rather than local or frequency filters such as Gaussian, anisotropic, or Wiener filters [16].

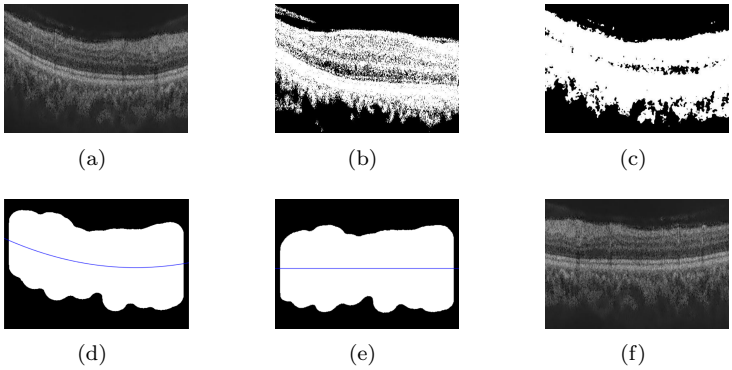


Figure 4: Flattening procedure: (a) original image, (b) thresholding, (c) median filter, (d) curve fitting, (e) warping, (f) flatten image.

3.1.2. Flattening

Textural descriptors characterize spatial arrangement of intensities. However, the OCT scans suffer from large type of variations: inclination angles, positioning, and natural curvature of the retina [14]. Therefore, these variations have to be taken into account to ensure a consistent characterization of the tissue disposition, regardless of the location in the retina. This invariance can be achieved from different manners: (i) using a rotation invariant descriptor (cf. Sect. 3.2), or (ii) by unfolding the curvature of the retina. This latter correction is known as image flattening which theoretically consists of two distinct steps: (i) estimate and fit the curvature of the Retinal Pigment Epithelium (RPE) and (ii) warp the OCT volume such that the RPE becomes flat.

Our correction is similar to the one of Liu *et al.* [14]: each B-scan is thresholded using Otsu's method followed by a median filtering to detect the different retina layers (see Fig 4(c) and Fig 4(b)). Then, a morphological closing and opening is applied to fill the holes and the resulting area is fitted using a second-order polynomial (see Fig. 4(d)). Finally, the scan is warped such that the curve becomes a line as presented in Fig. 4(e) and Fig. 4(f).

Table 2: Number of patterns ($LBP_{\#pat}$) for different sampling points and radius ($\{P, R\}$) of the LBP descriptor.

Sampling point for a radius ($\{P, R\}$)			
	$\{8, 1\}$	$\{16, 2\}$	$\{24, 3\}$
$LBP_{\#pat}$	10	18	26

3.1.3. Slice alignment

The flattening correction does not enforce an alignment through the OCT volume. Thus, in addition to the flattening correction, the warped curve of each B-scan are positioned at the same altitude in the z axis.

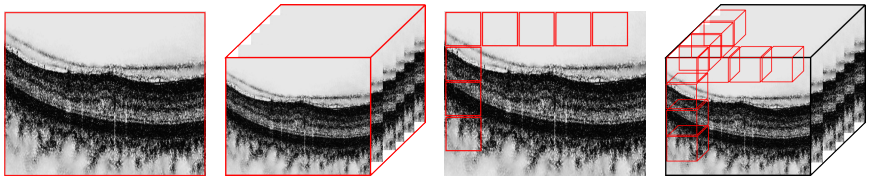
3.2. Feature detection

In this research, we choose to detect simple and efficient LBP texture features with regards to each OCT slice and volumes. LBP is a texture descriptor based on the signs of the differences of a central pixel with respect to its neighboring pixels [18]. These differences are encoded in terms of binary patterns as in Eq. (1):

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \ , \qquad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} , \tag{1}$$

where g_c, g_p are the intensities of the central pixel and a given neighbor pixel, respectively. P is the number of sampling points in the circle of radius R. Ojala *et al.* further extend the original LBP formulation to achieve rotation invariance at the expense of limiting the texture description to the notion of circular "uniformity" [18]. Volume encoding is later proposed by Zhao *et al.* by computing LBP descriptors in three orthogonal planes, so called LBP-TOP [19].

In this research we consider rotation invariant and uniform LBP and LBP-TOP features with various sampling points (i.e., $\{8, 16, 24\}$) with respect to different radius, (i.e., $\{1, 2, 3\}$). The number of patterns ($LBP_{\#pat}$) in regards with each configuration is reported in Table 2.



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

Figure 5: *Global* (a)-(b) and *local* (c)-(d) mapping for LBP and LBP-TOP features (2D B-scan and 3D volume, respectively).

3.3. Mapping

The mapping stage is used to partition the previously computed feature images to later extract the final descriptor as presented in the next section. For this work, two mapping strategies are defined: (i) *global* and (ii) *local* mapping.

Global mapping considers to extract the final descriptors from the 2D feature image for LBP and 3D volume for LBP-TOP. Therefore, for a volume with d slices, the *global*-LBP mapping will lead to the extraction of d elements. While the *global*-LBP-TOP represents the whole volume as a single element. The *global* mapping for 2D images and 3D volume is shown in Fig. 5(a) and 5(b).

Local mapping considers to extract the final descriptors from a set of $(m \times m)$ 2D patches for LBP and a set of $(m \times m \times m)$ sub-volumes for LBP-TOP. Given N and N' the total number of 2D patches and 3D sub-volumes respectively, the *local*-LBP approach provides $N \times d$ elements, while *local*-LBP-TOP provides N' elements. This mapping is illustrated in Fig. 5(c) and 5(d).

3.4. Feature extraction

Two strategies are used to describe each OCT volume texture.

Low-level representation The texture descriptor of an OCT volume is defined as the concatenation of the LBP histograms with the *global*-mapping.

The LBP histograms are extracted from the previously detected LBP images (see Sect. 3.2). Therefore, the LBP-TOP final descriptor is computed through the concatenation of the LBP histograms of the three orthogonal planes with the final size of $3 \times LBP_{\#pat}$. Similarly, the LBP descriptor is defined through concatenation of the LBP histograms per each slice with the final size of $d \times LBP_{\#pat}$.

High-level representation The concatenation of histograms employed in the low-level representation in conjunction with either *global*- or *local*-mapping can lead lead to a high dimensional feature space. For instance, *local*-mapping results to a size of $N \times d \times LBP_{\#path}$ for the final LBP descriptor and $N' \times LBP_{\#path}$ for the final LBP-TOP descriptor. High-level representation simplifies this high dimensional feature space into a more discriminant lower space. BoW approach is used for this purpose [11]. This model represents the features by creating a codebook or visual dictionary, from the set of low-level features. The set of low-level features are clustered using k-means to create the codebook with k clusters or visual words. After creating the codebook, each of the training example is represented as a histogram of size k. The histogram is obtained by calculating the frequency of occurrences of each of the k words in the extracted features from the training example.

3.5. Classification

Classification corresponds to the mapping of a set of inputs $\mathbf{x}$ into a set of categorical outputs $\mathbf{y}$ using a linear or non-linear function $f(\cdot)$. In supervised learning methods, this function is defined by providing a training set of N samples $\mathbf{x}_{tr}$ with their associated labels $\mathbf{y}_{tr}$. In the remainder of this section, we briefly summarize the supervised classification methods used in the experiments. Details regarding the parameters used in our experiments are provided in Sect. 4.

k-Nearest Neighbor (NN) is a non-parametric classification method in which an unlabeled feature vector x is assigned to the majority class of its k

nearest-neighbors from the training set. To avoid a tie case, the parameter k is set to an odd number.

Logistic Regression (LR) is a linear classifier which uses the logistic function to estimate the probability of x to belong to a particular class c_i [20]. Thus, the posterior probability is expressed as:

$$p(c_i|x) = \frac{1}{1 + \exp(-w^T x)} \tag{2}$$

where w is a vector of the regression parameters to obtain a linear combination of the input feature vector x. The vector w can be inferred by finding the maximum likelihood estimates via optimization methods such as quasi-Newton method [21]. Once the vector w is found, an unlabeled feature vector is assigned to the class which maximizes the posterior probability.

Random Forest (RF) is an ensemble of decision trees [22] which generalizes the classification process by applying two types of randomization: at the tree level, each tree is fed by a bootstrap made of S' samples which are built from the original data of size S such that $S = S'$, and at the node level, a subset of feature dimensions m is randomly selected from the original dimension M such that $m << M$. The trees in RF are grown to their maximum length without any pruning. In the testing stage, each tree in the ensemble casts a unit vote in the final prediction and the final prediction is based on combination of all the votes.

Gradient Boosting (GB) is a reformulation of AdaBoost [23] in which the problem of finding an ensemble of real-valued weak learners is tackled as a numerical optimization [24]. A strong learner is built by iteratively finding the best pair of real-valued weak learner function and its corresponding weight which minimizes a given differentiable loss function. Common choice for weak learners is decision stumps or regression trees while the loss function is generally an exponential or logarithmic loss [25], minimized via gradient descent or quadratic approximation.

Support Vector Machines (SVM) is a sparse kernel classification method which aims at finding the best linear hyperplane which separates two classes by maximizing the margin between them [26]. SVM becomes a non-linear classifier by using the kernel trick [27] which consists in replacing each inner product by a non-linear kernel function such as RBF or polynomial kernels.

4. Experiments and Validation

An experimental suit is designed to test the influence of the different blocks composing our framework, using different datasets (see Table 10). TThe rest of this section details the common configuration parameters across all the experiments, while the following subsections focus on the specific aim of each experiment.

Unless stated otherwise, all the experiments are run using our own dataset alone, SERI. Only for the sake of comparison, *Experiment #1* is performed on the public Duke dataset. Acquisition details regarding SERI and Duke datasets are reported in Sect. 4.1 and Sect. 4.2, respectively.

For all the experiments, LBP and LBP-TOP features are extracted for different sampling points of 8, 16, and 24 for radius of 1, 2, and 3, respectively. As previously mentioned, the *local*- and *global*-mapping strategies are used. The partitioning for *local*-mapping is set to (7×7) patch for 2D LBP and $(7 \times 7 \times 7)$ sub-volume for LBP-TOP.

All the experiments are evaluated using Leave-One-Patient Out Cross-Validation (LOPO-CV) strategy. At each round, a pair DME-normal volume is selected for testing while the remaining volumes are used for training. The use of this method implies that no variance in terms of Sensitivity (SE) and Specificity (SP) can be reported. Despite this limitation, LOPO-CV has been employed due to the small size of the dataset.

All the experiments are evaluated in terms of SE and SP, which are statistics driven from the confusion matrix (see Fig. 6) as stated in Eq. 3. The SE evaluates

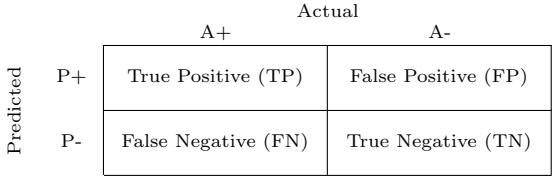


Figure 6: Confusion matrix with true and false positive detected samples (TP, FP) in the first row, from left to right and the false and true negative detected samples (FN, TN) in the second row, from left to right.

the performance of the classifier with respect to the positive class, while the SP evaluate its performance with respect to negative class.

$$SE = \frac{TP}{TP + FN} \qquad SP = \frac{TN}{TN + FP} \tag{3}$$

Among all, one experiment is carried out using Accuracy (ACC) and F1-score (F1) as formulated in Eq. 4. ACC is used to have an overall sense of the classifier performance, and F1 is used to see the trade off between SE and precision.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

The details of the experiments are presented from Sect. 4.3 to Sect. 4.6 and summarized in Table 10. *Experiment #1* (Sect. 4.3) is according to the experiment reported in [10] to evaluate the effects of different feature representations and compares the results to those obtained by Venhuizen *et al.* [13]. *Experiment #2 and #3* (Sect. 4.4 & 4.5) studies the high-level feature representation using BoW. The former experiment studies the effect of the codebook size in order to find the optimal number of words using a linear classifier; while the latter compares different classifiers (see Sect. 3.5). *Experiment #4* (Sect. 4.6) solely focus on the low-level representation.

A summary of the most relevant findings can be found in Sect. 5.

Table 3: The outline and summary of the performed experiments. $\sim$ indicate that common configuration applies.

	Dataset	Pre-processing	Features	Mapping	Representation	Classification	Evaluation
Common:	SERI	NL-means	LBP,LBP-TOP $S = \{8, 16, 24\}$ $R = \{1, 2, 3\}$				LOPO-CV SE, SP
Experiment#1: Goal: Evaluation of features, mapping and representation	+ Duke	$\sim$	$\sim$	*global* *local*	BoW Histogram	RF	+[13]
Experiment#2: Goal: Finding the optimum number of words	$\sim$	+ F + F+A	$\sim$	*global* *local*	BoW $k \in K$	LR	+ACC, F1
Experiment#3: Goal: Evaluation of different pre-processing for high-level features	$\sim$	+F +F+A	$\sim$	*global* *local*	BoW optimal k	3-NN RF SVM GB	$\sim$
Experiment#4: Goal: Evaluation of different pre-processing for low-level features	$\sim$	+F +F+A	$\sim$	*global*	Histogram	3-NN RF SVM GB	$\sim$

4.1. SERI-Dataset

This data was acquired by the Singapore Eye Research Institute (SERI), using CIRRUS TM (Carl Zeiss Meditec, Inc., Dublin, CA) SD-OCT device. The datasets consist of 32 OCT volumes (16 DME and 16 normal cases). Each volume contains 128 B-scan with resolution of 512 $\times$ 1024 pixels. All SD-OCT images are read and assessed by trained graders and identified as normal or DME cases based on evaluation of retinal thickening, hard exudates, intraretinal cystoid space formation and subretinal fluid.

4.2. Duke-Dataset

This dataset, published by Srinivasan *et al.* [12], was acquired in Institutional Review Board-approved protocols using Spectralis SD-OCT (Heidelberg Engineering Inc., Heidelberg, Germany) imaging at Duke University, Harvard University and the University of Michigan. This datasets consist of 45 OCT volumes (15 AMD, 15 DME and 15 normal). In this study we only consider a subset of the original data containing the 15 DME and the 15 normal OCT volumes.

4.3. Experiment #1

This experiment replicates some of the experiments reported in [10], using the SERI and Duke datasets. The volumes are pre-processed using NL-means. LBP and LBP-TOP descriptors are detected using the default configuration in conjunction with local and global mapping. Volumes are described using both low-level and high-level feature representation. In accordance with [10], BoW is used with a codebook of size 32 words, and the volumes are classified using RF classifier with 100 un-pruned trees.

Results are listed in Table 4. The two configurations achieving the best results in Table 4 are compared to Venhuizen *et al.* [13] in Table 5. Overall, the obtained results indicate that features driven from LBP descriptors are highly discriminative. Nevertheless, Table 5 indicates a substantial performance difference between SERI and Duke dataset. This is attributed to the fact that the volumes in Duke dataset are provided with embedded pre-processing steps.

Table 4: Experiment #1 - Obtained results of classification using SERI and Duke datasets.

Features	SERI dataste						Duke dataset					
	$\{8,1\}$		$\{16,2\}$		$\{24,3\}$		$\{8,1\}$		$\{16,2\}$		$\{24,3\}$	
	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP	SE	SP
global-LBP-TOP	56.2	62.5	**87.5**	**75.0**	68.7	68.7	80.0	93.3	73.3	86.6	73.3	86.6
local-LBP	**75.0**	**87.5**	81.2	75.0	68.7	62.5	**80.0**	**86.6**	**86.7**	**100**	93.3	86.6
local-LBP-TOP	62.5	68.7	56.2	37.5	37.5	43.7	80.0	86.6	86.6	86.6	60.0	80.0

Table 5: Experiment #1 - Comparing the proposed method by [13] on SERI and Duke datasets.

Data sets	SERI		Duke	
	SE	SP	SE	SP
Venhuizen *et al.* [13]	61.5	58.8	71.4	68.7
{*local*-LBP},{8, 1}	**75.0**	**87.5**	**86.6**	**100.0**
{*global*-LBP-TOP},{16, 2}	**75.0**	**87.5**	80.0	86.6

4.4. Experiment #2

In order to determine the optimal size of the codebook when using BoW, this experiment evaluates several codebook sizes on SERI dataset.

Several pre-processing strategies are evaluated: (i) NL-means, (ii) a combination of NL-means and flattening, and (iii) a combination of NL-means, flattening and aligning. LBP and LBP-TOP descriptors are detected using the default configuration. Volumes are represented using BoW, where the codebook size ranging for $k \in \{10, 20, 30, \cdots, 100, 200, \cdots, 500, 1000\}$. Finally, the volumes are classified using LR. The choice of this linear classifier avoids that the results get boosted by the classifier. In this manner any improvement would be linked to the pre-processing and the size of the codebook.

The usual build of the codebook consists of clustering the samples in the feature space using k-means (see Sect. 3.4). However, this operation is rather computationally expensive and convergence of the k-means algorithm for all codebook sizes is not granted. Nonetheless, Nowak *et al.* [28] pointed out that randomly generated codebooks can be used at the expenses of accuracy. Thus, the codebook are randomly generated since the final aim is to asses the influence of codebook size and not the performance of the framework. For this experiment, the codebook building is carried out using random initialization k-means++ algorithm [29], which is usually used as a k-means initialization algorithm.

Table 6: Experiment #2 - Optimum number of words for each configuration as a result of LR Classification, for high-level feature extraction of *global* and *local*-LBP, and *local*-LBP-TOP features with different pre-processing. The pre-processing includes: NF, F, and F+A.

Features	Pre-processing	$\{8,1\}$			$\{16,2\}$			$\{24,3\}$		
		ACC%	F1%	W#	ACC%	F1%	W#	ACC%	F1%	W#
global-LBP										
	NF	81.2	78.5	500	62.5	58.06	80	62.5	62.5	80
	F	71.9	71	400	68.7	66.7	300	68.7	66.7	300
	F+A	71.9	71	500	71.9	71	200	75	68.7	500
local-LBP										
	NF	75	75	70	65.6	64.5	90	62.5	60	30
	F	75	73.3	30	71.8	61	70	62.5	62.5	100
	F+A	75	69	40	71.9	71	200	68.7	66.7	10
local-LBP-TOP										
	NF	68.7	68.7	400	75	75	500	71.9	71	60
	F	68.7	68.7	300	68.7	66.7	50	75	76.5	80
	F+A	75	73.3	100	75	73.3	90	75	69	70

Table 7: Experiment #2 - The obtained results, using the optimal number of words in terms of SE and SP.

Features	Pre-processing	$\{8,1\}$			$\{16,2\}$			$\{24,3\}$		
		SE%	SP%	W#	SE%	SP%	W#	SE%	SP%	W#
global-LBP										
	NF	68.7	93.7	500	56.2	62.5	80	62.5	62.5	80
	F	68.7	75.0	400	62.5	75.0	300	62.5	75.0	300
	F+A	68.7	75.0	500	68.7	75.0	200	68.7	68.7	500
local-LBP										
	NF	75.0	75.0	70	62.5	68.7	90	56.2	68.7	30
	F	68.7	81.2	30	68.7	75.0	70	62.5	62.5	100
	F+A	62.5	81.2	40	68.7	75.0	200	68.7	62.5	10
local-LBP-TOP										
	NF	68.7	68.7	400	75.0	75.0	500	68.7	75.0	60
	F	68.7	68.7	300	62.5	75.0	50	81.2	68.7	80
	F+A	68.7	81.2	100	68.7	81.2	90	62.5	81.2	70

Figure 7 shows the ACC and F1 score graphs obtained for a single case [1] in [1], while the optimal number of words for all the configuration are reported in a compact manner in Table 6. Table 7 reports the performance of the optimal codebook size in terms of SE and SP.

The obtained results show that commonly less number of words is required when higher number of sampling points and radius ($\{P, R\} = \{24, 3\}$) are used. The required number of words decreases for *local*-LBP in comparison with *global*-LBP. Although it was expected that the use of different pre-processing steps affect the optimal number of words, this influence is not substantial nor consistent over all the obtained results.

[1] Full set of scores can be found at the github repository

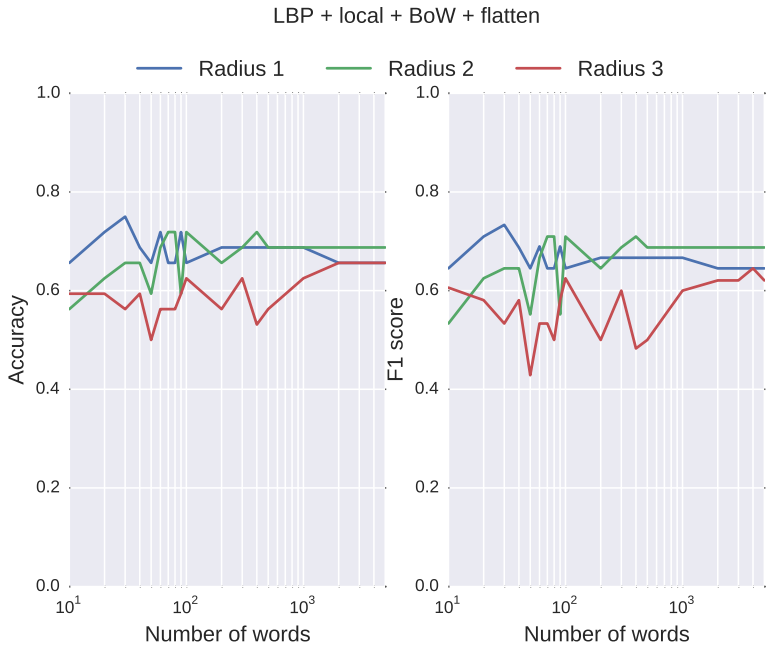


Figure 7: The performance of LR with NL-means+F pre-processing for different P and R.

4.5. Experiment #3

After studying the impact of the codebook size in Sect. 4.4, this experiment explores the improvement associated with: (i) different pre-processing, and (ii) using larger range of classifiers (i.e. linear and non-linear).

For this experiment, several pre-processing strategies are evaluated: (i) NL-means, (ii) a combination of NL-means and flattening, and (iii) a combination of NL-means, flattening and aligning. LBP and LBP-TOP features are detected using the default configuration and volumes are represented using BoW. The codebooks are computed using regular k-means algorithm which is initialized using k-means++, where k is chosen according to the findings of *Experiment #2*. Finally, the volumes are classified using k-NN, RF, GB, and SVM.

Regarding the classification strategies, k-NN classifier is trained by considering the 3 nearest neighbor. The RF and GB classifier are trained using 100 un-pruned trees, while SVM classifier is trained with RBF kernel and its pa-

rameters C, and γ are optimized through grid-search.

Table 8 shows the obtained results from this experiment, where the most relevant configurations are shaded and the highest results are highlighted in **bold**. Regarding the effects of pre-processing, the performance of the most configurations decreases by aligning or flattening the B-scan (i.e. light shaded configurations in Table 8). However, the two best configurations (i.e. dark shaded in Table 8), achieve better results when adding flattening or flattening and alignment as pre-processing. A small radius and small number of samples in feature detection tends to increase the classification performance. Regarding the mapping strategy, local mapping tends to produce better results than global mapping. In terms of choosing a classifier, SVM provides the best results, followed by RF.

The best results (81.2% SE and 93.7% SP) are achieved using NL-means, flattening, LBP detection using $\{P, R\} = \{8, 1\}$, local mapping, high-level representation, using a codebook with $k = 70$, and SVM classifier. This result can be compared with other relevant results in Table 10.

4.6. Experiment #4

This experiment replicates the *Experiment #3* for the case of low-level representation features from the volumes.

For this experiment, several pre-processing strategies are evaluated: (i) NL-means, (ii) a combination of NL-means and flattening, and (iii) a combination of NL-means, flattening and aligning. LBP and LBP-TOP descriptors are detected using the default configuration. Volumes are represented using low-level feature representation of the *global* mapping. Finally, the volumes are classified using k-NN, RF, GB, and SVM, similarly to *Experiment #3*.

The obtained results from this experiment is listed in Table. 9. The most relevant configurations are shaded and the highest results are highlighted in **bold**. Similarly to the results reported in Sect. 4.5, the effect of flattening the B-scan boosts the results for the best performing configuration, but this effect is not consistent across all the configurations. For this experiment, LBP-TOP

outperforms LBP and larger P and R values for feature detection tends to obtain better results. In terms of classifier, RF have better performance than the others but the highest SP is achieved using SVM.

The best results (81.2% SE and 81.2% SP) are achieved when using NL-means, flattening, LBP-TOP detection using $\{P, R\} = \{24, 3\}$, global mapping, low-level representation, and RF classifier. This result can be compared with other relevant results in Table 10

Table 8: Experiment #3 - k-NN and SVM classification with BoW for the *global* and *local* LBP and *local* LBP-TOP features with different pre-processing. The optimum number of words were selected based on the previous experiment.

Features	Pre-processing	k-NN						SVM					
		$\{8,1\}$		$\{16,2\}$		$\{24,3\}$		$\{8,1\}$		$\{16,2\}$		$\{24,3\}$	
		SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%
global-LBP													
	NF	43.7	93.7	43.7	87.5	43.7	62.5	68.7	87.5	62.5	62.5	50.0	56.2
	F	43.7	56.2	50.0	75.0	62.5	56.2	56.2	56.2	56.2	75.0	56.2	68.7
	FA	56.2	62.5	43.7	81.2	68.7	56.2	56.2	68.7	68.7	68.7	56.2	75.0
local-LBP													
	NF	**75.0**	**87.5**	50.0	68.7	43.7	43.7	**75.0**	**93.7**	50.0	75.0	56.2	56.2
	F	56.2	56.2	50.0	50.0	50.0	43.7	**81.2**	**93.7**	68.7	68.7	68.7	75.0
	FA	56.2	43.7	50.0	75.0	50.0	62.5	**75.0**	**93.7**	75.0	68.7	68.7	68.7
local-LBP-TOP													
	NF	56.2	75.0	56.2	75.0	62.5	56.2	**81.2**	**87.5**	**75.0**	**100**	56.2	75.0
	F	62.5	43.7	37.5	68.7	43.7	62.5	**81.2**	**81.2**	75.0	68.7	81.2	68.7
	F+A	56.2	56.2	68.7	50.0	43.7	62.5	62.5	75.0	68.7	75.0	62.5	81.2

Features	Pre-processing	RF						GB					
		8^{riu2}		16^{riu2}		24^{riu2}		8^{riu2}		16^{riu2}		24^{riu2}	
		SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%
global-LBP													
	NF	**68.7**	**93.7**	43.7	62.5	50.0	68.7	56.2	50.0	37.5	31.2	50.0	43.7
	F	56.2	50.0	56.2	75.0	50.0	75.0	50.0	56.2	56.2	75.0	43.7	62.5
	FA	68.7	50.0	56.2	62.5	62.5	56.2	56.2	50.0	68.7	50.0	43.7	75.0
local-LBP													
	NF	**81.2**	**81.2**	62.5	56.2	56.2	56.2	75.0	62.5	68.7	87.5	50.0	75.0
	F	56.2	81.2	62.5	68.7	68.7	62.5	68.7	75.0	50.0	75.0	50.0	62.5
	FA	68.7	62.5	62.6	68.7	43.7	43.7	56.2	50.0	68.7	56.2	50.0	50.0
local-LBP-TOP													
	NF	68.7	62.5	**68.7**	**81.2**	68.7	68.7	37.5	68.7	62.5	81.2	62.5	50.0
	F	50.0	62.5	62.5	62.5	43.7	75.0	50.0	56.2	43.7	62.5	50.0	62.5
	F+A	50.0	62.5	**81.2**	**87.5**	50.0	68.7	56.2	62.5	81.2	68.7	75.0	68.7

Table 9: Experiment #4 - Classification results obtained from low-level representation of global LBP and LBP-TOP features with different pre-processing. Pre-processing steps include: NF, F, F+A. Different classifiers such as RF, GB, SVM, and k-NN are used.

Features	Pre-processing	k-NN						k-SVM					
		$\{8,1\}$		$\{16,2\}$		$\{24,3\}$		$\{8,1\}$		$\{16,2\}$		$\{24,3\}$	
		SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%
global-LBP													
	NF	37.5	50.0	25.0	50.0	37.5	68.7	56.2	62.5	56.2	43.7	56.2	68.7
	F	62.5	50.0	56.2	75.0	62.5	68.7	75.0	68.7	62.5	62.5	62.5	68.7
	FA	56.2	50.0	56.2	75.0	62.5	68.7	75.0	68.7	62.5	62.5	62.5	68.7
global-LBP-TOP													
	NF	31.2	93.7	37.5	100.0	37.5	81.2	62.5	75.0	**62.5**	**93.7**	56.2	87.5
	F	50.0	56.2	56.2	75.0	56.2	62.5	68.7	75.0	43.7	68.7	68.7	56.2
	F+A	75.0	43.7	56.2	43.7	68.7	50.0	68.7	62.5	62.5	56.2	56.2	68.7
Features	Pre-processing	RF						GB					
		8^{riu2}		16^{riu2}		24^{riu2}		8^{riu2}		16^{riu2}		24^{riu2}	
		SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%	SE%	SP%
global-LBP													
	NF	43.7	62.5	43.7	62.5	56.2	75	43.7	43.7	43.7	37.5	37.5	31.25
	F	56.2	56.2	68.7	62.5	62.5	68.7	25	56.2	50.0	43.7	25.0	43.7
	F+A	65.2	56.2	50.0	50.0	56.2	68.7	43.75	62.5	62.5	50.0	31.2	31.2
global-LBP-TOP													
	NF	56.2	68.7	**68.7**	**87.5**	**68.7**	**81.2**	68.7	68.7	75.0	50.0	56.2	43.7
	F	56.2	62.5	81.2	68.7	**81.2**	**81.2**	56.2	62.5	62.5	68.7	68.7	81.2
	F+A	68.7	62.5	75.0	68.7	**75.0**	**81.2**	56.2	43.7	62.5	62.5	**75.0**	**75.0**

5. Conclusions

The work presented here addresses automatic classification of SD-OCT volumes as normal or DME. In this regard, an extensive study is carried out covering (i) the effects of different pre-processing steps, (ii) the influence of different mapping and feature extraction strategies, (iii) the impact of the codebook in BoW, (iv) the comparison of different classification strategies.

Table 10 summarizes the most relevant aspects of all the experimentation here reported, showing a clear improvement with respect to previous studies [10, 13]. Based on the reported results, the 3D features and high level 2D features using patches achieve the most desirable results.

Table 10: Summary of all the results

Experiment 3	81.2	93.7	NLM+F	LBP	✓			local	High	SVM	30
	75.0	93.7	NLM+F+A	LBP	✓			local	High	SVM	40
	75.0	93.7	NLM	LBP	✓			local	High	SVM	70
	75.0	100	NLM	LBP-TOP		✓		local	High	SVM	500
	81.2	87.5	NLM	LBP-TOP	✓			local	High	SVM	400
	81.2	87.5	NLM+F+A	LBP-TOP		✓		local	High	RF	90
	81.2	81.2	NLM	LBP	✓			local	High	RF	70
Experiment 4	81.2	81.2	NLM	LBP-TOP			✓	global	Low	RF	
	81.2	81.2	NLM+F	LBP-TOP	✓			local	High	SVM	300
	81.2	81.2	NLM+F+A	LBP-TOP			✓	global	Low	GB	
	81.2	81.2	NLM+F	LBP-TOP			✓	global	Low	RF	
	75.0	87.5	NLM	LBP	✓			local	High	k-NN	70
Experiment 1	75.0	87.5	NLM	LBP	✓			local	High	RF	32
Experiment 1	75.0	87.5	NLM	LBP-TOP		✓		global	High	RF	32
	68.7	93.7	NLM	LBP	✓			global	High	RF	500
	75	81.2	NLM+F+A	LBP-TOP			✓	global	Low	RF	
	68.7	81.2	NLM	LBP-TOP		✓		local	High	RF	500
	62.5	93.7	NLM	LBP-TOP		✓		global	Low	SVM	
	68.7	87.5	NLM	LBP-TOP		✓		global	Low	RF	
	68.7	81.2	NLM	LBP-TOP				global	Low	RF	
	75.0	75.0	NLM	LBP-TOP				global	Low	RF	
	68.7	75.0	NLM+F	LBP-TOP	✓			global	Low	SVM	
	56.2	75.0	NLM	LBP			✓	global	Low	RF	
	56.2	75.0	NLM+F	LBP		✓		global	Low	k-NN	
	56.2	75.0	NLM+F+A	LBP		✓		global	Low	k-NN	
Venhuizen *et al.* [13]	61.5	58.8									

References

[1] G. Lemaître, M. Rastgoo, J. Massich, retinopathy: Miccai-omia-2015 (Jul. 2015). doi:10.5281/zenodo.22195.
URL http://dx.doi.org/10.5281/zenodo.22195

[2] S. Sharma, A. Oliver-Hernandez, W. Liu, J. Walt, The impact of diabetic retinopathy on health-related quality of life, Curr. Op. Ophtal. 16 (2005) 155–159.

[3] S. Wild, G. Roglic, A. Green, R. Sicree, H. King, Global prevalence of diabetes estimates for the year 2000 and projections for 2030, Diabetes Care 27 (5) (2004) 1047–1053.

[4] M. D. Abramoff, M. K. Garvin, M. Sonka, Retinal image analysis: a review, IEEE Review Biomed. Eng. 3 (2010) 169–208.

[5] E. Trucco, A. Ruggeri, T. Karnowski, L. Giancardo, E. Chaum, J. Hubschman, B. al Diri, C. Cheung, D. Wong, M. Abramoff, G. Lim, D. Kumar, P. Burlina, N. M. Bressler, H. F. Jelinek, F. Meriaudeau, G. Quellec, T. MacGillivray, B. Dhillon, Validation retinal fundus image analysis algorithms: issues and proposal, Investigative Ophthalmology & Visual Science 54 (5) (2013) 3546–3569.

[6] Early Treatment Diabetic Retinopathy Study Group, Photocoagulation for diabetic macular edema: early treatment diabetic retinopathy study report no 1, Arch. Ophtalmol. 103 (12) (1985) 1796–1806.

[7] T. C. Chen, B. Cense, M. C. Pierce, N. Nassif, B. H. Park, S. H. Yun, B. R. White, B. E. Bouma, G. J. Tearney, J. F. de Boer, Spectral domain optical coherence tomography: ultra-high speed, ultra-high resolution ophtalmic imaging, Arch. Ophtalmol. 123 (12) (2005) 1715–1720.

[8] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, S. Farsiu, Automatic segmentation of seven retinal layers in sd-oct images congruent with expert manual segmentation, Optic Express 18 (18) (2010) 19413–19428.

[9] R. Kafieh, H. Rabbani, M. D. Abramoff, M. Sonka, Intra-retinal layer segmentation of 3d optical coherence tomography using coarse grained diffusion map, Medical Image Analysis 17 (2013) 907–928.

[10] G. Lemaître, M. Rastgoo, J. Massich, S. Sankar, F. Mériaudeau, D. Sidibé, Classification of SD-OCT volumes with LBP: Application to dme detection, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Ophthalmic Medical Image Analysis Workshop (OMIA), 2015.

[11] J. Sivic, A. Zisserman, Video google: a text retrieva approach to object matching in videos, in: IEEE ICCV, 2003, pp. 1470–1477.

[12] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, S. Farsiu, Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images, Biomedical Optical Express 5 (10) (2014) 3568–3577.

[13] F. G. Venhuizen, B. van Ginneken, B. Bloemen, M. J. P. P. van Grisven, R. Philipsen, C. Hoyng, T. Theelen, C. I. Sanchez, Automated age-related macular degeneration classification in OCT using unsupervised feature learning, in: SPIE Medical Imaging, Vol. 9414, 2015, p. 94141l.

[14] Y.-Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, R. J. M., Automated macular pathology diagnosis in retinal oct images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding, Medical Image Analysis 15 (2011) 748–759.

[15] J. M. Schmitt, S. Xiang, K. M. Yung, Speckle in optical coherence tomography, Journal of biomedical optics 4 (1) (1999) 95–105.

[16] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, IEEE, 2005, pp. 60–65.

[17] P. Coupe, P. Hellier, C. Kervrann, C. Barillot, Nonlocal means-based speckle filtering for ultrasound images, IEEE TIP (2009) 2221–2229.

[18] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (7) (2002) 971–987.

[19] G. Zhao, T. Ahonen, J. Matas, M. Pietikäinen, Rotation-invariant image and video description with local binary pattern features, Image Processing, IEEE Transactions on 21 (4) (2012) 1465–1477.

[20] D. R. Cox, The regression analysis of binary sequences, Journal of the Royal Statistical Society. Series B (Methodological) (1958) 215–242.

[21] R. H. Byrd, J. Nocedal, R. B. Schnabel, Representations of quasi-newton matrices and their use in limited memory methods, Mathematical Programming 63 (1-3) (1994) 129–156.

[22] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[23] J. H. Friedman, Stochastic gradient boosting, Computational Statistics & Data Analysis 38 (4) (2002) 367–378.

[24] G. Lemaitre, J. Massich, R. Marti, J. Freixenet, J. C. Vilanova, P. M. Walker, D. Sidibe, F. Meriaudeau, A boosting approach for prostate cancer detection using multi-parametric mri, in: International Conference on Quality Control and Artificial Vision (QCAV2015), SPIE, 2015.

[25] C. Becker, R. Rigamonti, V. Lepetit, P. Fua, Supervised feature learning for curvilinear structure segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer, 2013, pp. 526–533.

[26] V. Vapnik, A. Lerner, Generalized portrait method for pattern recognition, Automation and Remote Control 24 (6) (1963) 774–780.

[27] A. Aizerman, E. M. Braverman, L. I. Rozoner, Theoretical foundations of the potential function method in pattern recognition learning, Automation and Remote Control 25 (1964) 821–837.

[28] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: Computer Vision–ECCV 2006, Springer, 2006, pp. 490–503.

[29] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.