Classifying DME vs Normal SD-OCT volumes: A review

Joan Massich*, Mojdeh Rastgoo*, Guillaume Lemaître*, Carol Y. Cheung†,
Tien Y. Wong†, Désiré Sidibé*, Fabrice Mériaudeau*‡
*LE2I UMR6306, CNRS, Arts et Métiers, Univ. Bourgogne Franche-Comté,
12 rue de la Fonderie, 71200 Le Creusot, France
†Singapore Eye Research Institute, Singapore National Eye Center, Singapore
‡Centre for Intelligent Signal and Imaging Research (CISIR), Electrical & Electronic Engineering Department,
Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia
¶Corresponding author: joan.massich@u-bourgogne.fr

***Abstract*—This article reviews the current state of automatic classification methodologies to identify Diabetic Macular Edema (DME) versus normal subjects based on Spectral Domain OCT (SD-OCT) data. Addressing this classification problem has valuable interest since early detection and treatment of DME play a major role to prevent eye adverse effects such as blindness.**

The main contribution of this article is to cover the lack of a public dataset and benchmark suited for classifying DME and normal SD-OCT volumes, providing our own implementation of the most relevant methodologies in the literature. Subsequently, 6 different methods were implemented and evaluated using this common benchmark and dataset to produce reliable comparison.

***Index Terms*—Diabetic Macular Edema (DME), Spectral Domain OCT (SD-OCT), Machine Learning (ML), benchmark,**

I. Introduction

Diabetic Retinopathy (DR), and more particularly Diabetic Macular Edema (DME), are leading causes of irreversible vision loss and the most common eye diseases in individuals with diabetes. Taking into account that the number of individuals affected by diabetes dieases are expected to grow exponentially in the next decade [1], developing methodologies for early detection and treatment of DR and DME has become a priority to prevent adverse effects.

The main focus of this work is to describe the actual state of DME detection in Optical Coherence Tomography (OCT) images. DME presents an increase in retinal thickness within 1 disk diameter of the fovea center with or without hard exudates and sometimes associated with cysts [2]. Spectral Domain OCT (SD-OCT) is an emerging eye imaging modality providing cross-sectional retinal morphology information [3], which cannot be estimated from more established eye imaging modalities such as fundus imaging.

The initial efforts of the ophtalmic community in developing technologies for SD-OCT have been placed in segmenting the retinal layers, which is a necessary step for retinal thickness measurements [4, 5]. However, latter efforts address the specific problem of DME automatic detection in OCT volumes. These efforts reveal the needs to address: (i) enhancing the quality of OCT volumes, (ii) finding pathology signs, and (iii) appropriate classification strategies.

Advances in any of those regards is of great interest since (i) manual evaluation of SD-OCT volumetric scans is expensive and time consuming [**?**]; (ii) SD-OCT acquisition has some shortcomings due to eye movements during the scanning [**?**], reflectivity nature of the retina [6], high level of noise and inconsistent quality of the images; (iii) due to the coexistence of multiple pathologies [**?**] as well as large intra-pathology variability, consistently identifying pathology-specific biomarkers remains challenging [**?**].

The rest of this article is structured as follows: Section II offers a general idea of the literature state-of-the-art in SD-OCT volume classification. Section III reviews some publicly available datasets and states the need for another one that suits the classification task here described. Section IV proposes an experimental benchmark to compare different methodologies presented in Sect. II. Section V reports and discusses the obtained results, while Sect. VI wraps up our thoughts regarding this work and its possible direction.

II. Background

This section reviews works straightly addressing the problem of classifying OCT volumes as normal or abnormal, regardless of the targeted pathology. The methods are categorized in terms of their learning strategy, namely supervised or semi-supervised learning.

A. Supervised methods

Supervised learning is based on a fully annotated and labeled training set. In this approach, the labeled training data are used to train the classifier function later used for prediction. Figure 1 illustrates a prevalent framework for supervised learning. Each SD-OCT volume undergoes: (i) *preprocessing* to reduce noise and other acquisition deficiencies which alter the images; (ii) *feature detection* to quantify visual cues like appearance, texture, shape, etc.; (iii) *mapping* in which a sample is either considered as whole (i.e., global) or partitioned into a set of sub-elements (i.e., local dense/sparse patches, pyramid, etc.); (iv) *feature representation* to associate a descriptor (e.g., concatenation, statistics, histogram, Principal Component Analysis (PCA), Bag-of-Words (BoW),

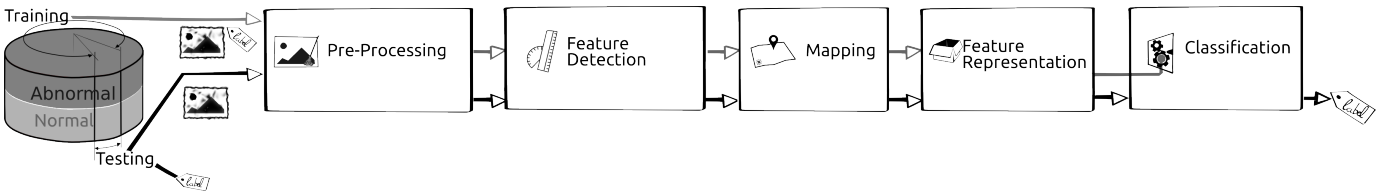


Fig. 1. Common framework

etc.) for each element from the *mapping-stage*. This descriptor packages the visual cues related to the sample; (v) *classification* to determine the associated class of each sample.

Venhuizen *et al.* propose a classification method to distinguish between Age-related Macular Degeneration (AMD) and normal SD-OCT volumes using BoW models [**?**]. A set of keypoints are detected and selected at each individual B-scan, by keeping the salient points included in the top 3% of the vertical gradient values. Around each of these keypoints, a $9\,\text{px} \times 9\,\text{px}$ texton is extracted, generating a feature vector of 81 dimensions, later reduce to 9 using PCA. All extracted feature vectors are used to create a codebook using *k*-means clustering. Then, each OCT volume descriptor is represented as a histogram that captures the codebook occurrences and are classified by a Random Forest (RF) composed of 100 trees. The method is tested using a publicly available dataset of 384 OCT volumes [7], achieving an Area Under the Curve (AUC) of 0.984.

Srinivasan *et al.* propose a classification method to distinguish DME, AMD, and normal SD-OCT volumes [**?**]. Each OCT slice is pre-processed using Block Matching 3D filtering (BM3D) to reduce the speckle noise and is flattened to reduce the inter-patient retinal curvature variations. A multi-resolution pyramid is generated for each pre-processed slice and a Histogram of Oriented Gradients (HOG) feature is computed for each layer. These features are classified using a linear Support Vector Machines (SVM). Note that each individual B-scan is classified into one of the three categories, namely DME, AMD, and normal, and a volume is label to a given class by taking the majority vote of all B-scans. This method is also tested using a publicly available dataset, composed of 45 patients equally subdivided into the three targeted classes. Correct classification rates of 100%, 100% and 86.67% are obtained for normal, DME, and AMD patients, respectively.

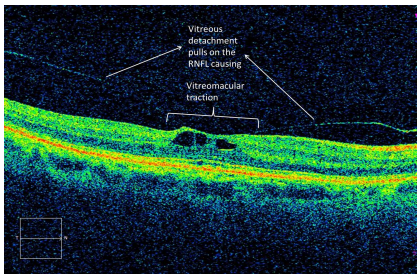
Extending the previous work, Alsaih *et al.* aggregate Local Binary Patterns (LBP) to HOG in order to add texture information and reduce the number of dimension using PCA [**?**].

Lemaître *et al.* propose a method based on LBP features to describe the texture of OCT images and dictionary learning using the BoW models [**?**]. In this method, the OCT images are first pre-processed using Non-Local Means (NLM) filtering, to reduce the speckle noise. Then, the volumes are mapped into a discrete set of structures: (i) local corresponding to patches, or (ii) global corresponding to volume slices or the whole volume. According to the chosen mapping, LBP or LBP from Three Orthogonal Planes (LBP-TOP) texture features are extracted and represent each volume through histogram, PCA, or BoW representation. The final feature descriptors are classified using RF classifier. This methodology is tested against Venhuizen *et al.* [**?**] using public and non-public datasets showing an improvement within the results by achieving a Sensitivity (SE) of 87.5% and a Specificity (SP) of 75%.

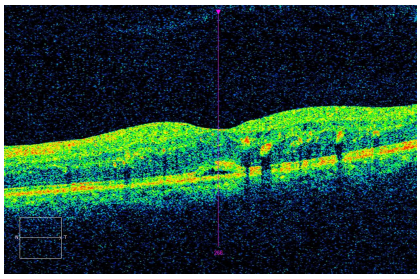
Liu *et al.* propose a methodology aiming at classifying B-scan rather than volume. The classification goal is to distinguish between macular pathology and normal OCT B-scan images using LBP and gradient information as attributes [**?**]. Each OCT slice is flattened before to create a 3-level mutlti-scale spatial pyramid. From each layer of this pyramid, edges are extracted and LBP descriptors are computed for the flattened slice and the edge map. All the obtained histograms are concatenated into a global descriptor whose dimensions are reduced using PCA. Finally, a SVM with a Radial Basis Function (RBF) kernel is used as classifier. A detection rate with an AUC of 0.93 is achieved, using a dataset of 326 OCT scans with various pathologies.

Albarrak *et al.* propose another classification framework to differentiate AMD and normal volumes [**?**]. Each OCT slice undergoes two pre-processing routines: (i) a joint denoising and cropping step using the split Bregman isotropic total variation algorithm and (ii) a flattening step by fitting a second-order polynomial using a least-square approach. Then, LBP-TOP and HOG combined with LBP-TOP features are extracted from individual sub-volumes from each original cropped volume. These features are concatenated into a single feature vector per OCT volume and its dimension is reduced using PCA. Finally, a Bayesian network classifier is used to classify the volumes. The classification performance of the framework in terms of SE and SP achieves 92.4% and 90.5%, respectively, outperforming the method of Liu *et al.* [**?**], using a dataset composed of 140 OCT volumes.

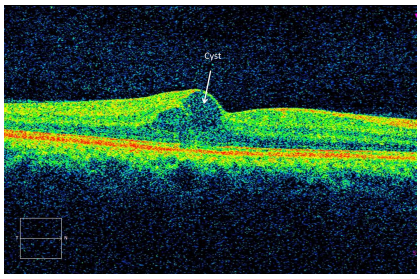
Anantrasirichai *et al.* propose to detect glaucoma in OCT images based on a variety of texture descriptor [**?**]. The texture information is described through LBP, Gray-level co-occurrence matrix (GLCM), wavelet, granulometry, run length measures, and intensity level distributions in combination with retinal layer thickness estimation, without any pre-processing. Each feature vector is projected using PCA before to be classified using an SVM with both linear and RBF kernel. Testing with rather a small dataset of 24 OCT volumes, their proposed method achieves an Accuracy (ACC) of 85 % while using layer thickness and textural informations.



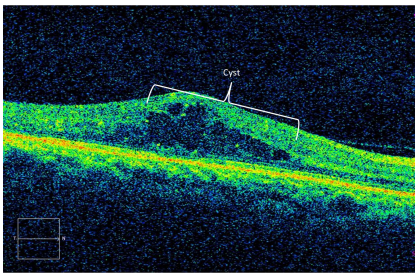
(a) Vitreomacular traction.



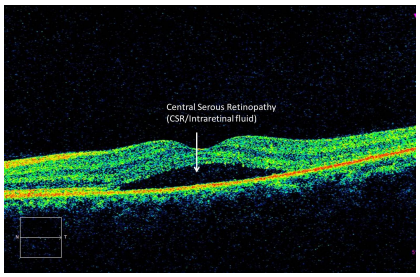
(b) Rethinal thickening.



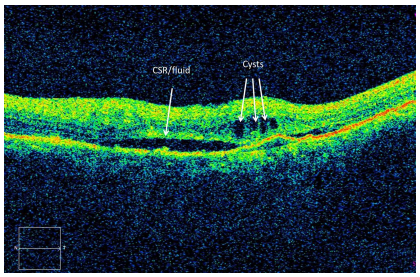
(c) Cyst spaces, causing central and parafoveal retina thickening.



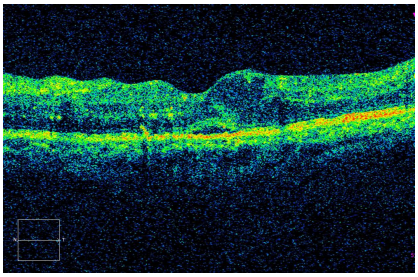
(d) Cyst spaces and hard exudates, causing central retinal thickening.



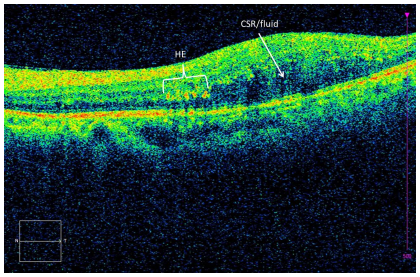
(e) CSR (subretinal fluid), causing central and parafoveal thickening.



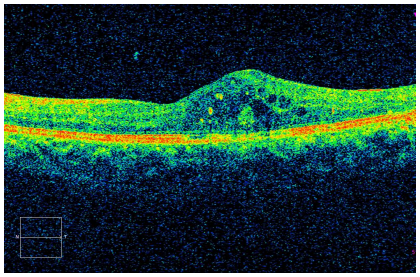
(f) CSR, hard exudates and cyst spaces.



(g) Cyst spaces, causing retinal thickening.



(h) CSR and hard exudates, causing retinal thickening.



(i) Cyst spaces causing parafoveal thickening.

Fig. 2. Examples of DME cases in Singapore Eye Research Institute (SERI) dataset.

B. Semi-supervised methods

Sankar *et al.* propose to use a semi-supervised strategy to classify DME *vs.* normal OCT volumes based on appearance modeling of normal OCT images using Gaussian Mixture Model (GMM) [8]. The main difference between this method and the supervised methodologies lies in the fact that only normal volumes are used to train the system.

For each OCT volume, the B-scans are denoised using NLM filtering, flattened, and resized to ensure homogeneous dimension across all volumes. Each B-scan is vectorized and projected into a lower-dimensional space with p dimensions using PCA. Subsequently, normal B-scans are modelled using a GMM in which the number of mixture components K is determined on a validation set. At the testing stage, a scan is classified as normal or DME depending of its Mahalanobis distance to the learnt model; if the distance is greater than the 97.5% quantile of the Chi-squared distribution with p degree of freedom. Therefore, a volume is classified as abnormal if the number of abnormal slice is greater than a given threshold, previously determined during the validation procedure. A SE and SP of 93.8% and 80.0% are respectively achieved on a cohort of 32 patients.

[t]

III. Data

A common dataset is required to compare different methodologies. Despite the fact that lack of public data is a common claim in the medical image community [9], the ophtalmic community has recently made available public dataset, mainly gathered at Duke University [7, **?**]. Although these datasets have been used by Venhuizen *et al.* [**?**] and Srinivasan *et al.* [**?**], they are not suitable for our problem.

Venhuizen *et al.* have evaluated their framework on a large public dataset of 384 OCT annotated volumes classified either as AMD or normal cases. Our goal, however, remains to focus on the detection of DME rather than AMD, despite the interest of testing the frameworks against a large dataset.

TABLE I
CORRESPONDENCE BETWEEN THE MOST RELEVANT METHODOLOGIES REVIEWED IN SECT. II AND THE PROPOSED EXPERIMENTAL FRAMEWORK.

Ref	Pre-processing	Features	Mapping	Representation	Classification
Venhuizen *et al.* [**?**, **?**]		Texton	Local	PCA BoW	RF
Srinivasan *et al.* [**?**, **?**]	Denoising (BM3D) Flattening Cropping	HOG	Global		Linear-SVM
Lemaître *et al.* [**?**, **?**]	Denoising (NLM)	LBP LBP-TOP	Local Global	PCA BoW Histogram	RF
Alsaih *et al.* [**?**]	Denoising (BM3D) Flattening Cropping	LBP HOG	Local	PCA Histogram	Linear-SVM
Liu *et al.* [**?**, **?**]	Flatten Aligned	Edge LBP	Local	PCA BoW	RBF-SVM
Sankar *et al.* [8, **?**]	Denoising (NLM) Flattening Cropping	Pixel intensities	Global	PCA	Mahalanobis-distance to GMM

In the contrary, Srinivasan *et al.* have tested their framework using a public dataset containing AMD, DME, and normal volumes [**?**]. The data, however, have been denoised, aligned, and cropped, without access to the original set of images.

Therefore, we use the SERI dataset to conduct this study [**?**]. This dataset has been acquired by the SERI, using CIRRUS TM (Carl Zeiss Meditec, Inc., Dublin, CA) SD-OCT device. The dataset consists of 32 OCT volumes, subdivided into 16 DME and 16 normal cases. Each volume contains 128 B-scans with a resolution of $512\,\text{px} \times 1{,}024\,\text{px}$. All SD-OCT images have been 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 (see Fig. 2).

IV. EXPERIMENTAL SETUP

The experimental set-up is summarized in Table III, where the most relevant works in Sect. II are formulated as the 5-steps standard classification procedure described in Fig. 1.

A. Implementation details

The experiments, described in this work, are publicly available at [**?**] allowing for further comparisons and improvements. All the methods in Table III have been developed using *protoclass* [**?**], a rapid prototyping toolkit to perform image processing and Machine Learning (ML) tasks. Furthermore, each method has been implemented as a plug-in to [**?**], so that all methods can be evaluated in a common framework [1].

Note that Liu *et al.* train the algorithm at the B-scan level, and SERI dataset provides Ground Truth (GT) at volume level only. Thus, two strategies have been explored to solve this issue: (i) similarly to Srinivasan *et al.*, at training stage, all B-scans are considered as abnormal for a DME volume and at testing stage, a majority vote rule is applied to whether label a volume as abnormal or not; (ii) similarly to Venhuizen *et al.*, an approach using BoW is used. From the methods reviewed in Sect. II, we decline to implement Albarrak *et al.* and Anantrasirichai *et al.*. The former do not provide sufficient implementation details to replicate their results [**?**]; while, the latter use a descriptor based on the layer thickness which require a layer segmentation stage using a generic segmentation algorithm and further user validation [**?**].

[1] See table III for standalone repositories of each method. All repositories provide tests to ensure that our implementation comply with the original work.

B. Evaluation

All the experiments are evaluated in terms of SE and SP (see Fig. 3) using the Leave-Two-Patient Out Cross-Validation (LTPO-CV) strategy, in line with [**?**]. Therefore, at each cross-validation iteration, a DME and normal volumes are kept for testing, while the remaining volumes are used as training. The SE evaluates the performance of the classifier with respect to the positive class, while the SP evaluates its performance with respect to negative class.

Subsequently, no SE or SP variance can be reported.

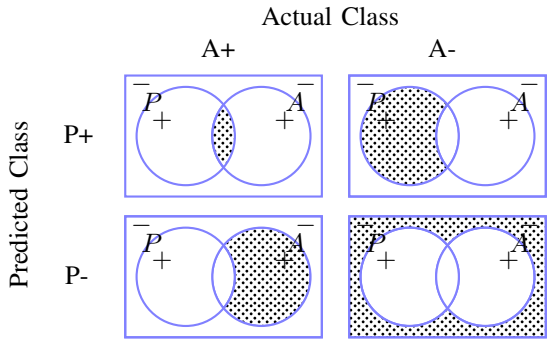
V. RESULTS AND DISCUSSION

The entire set of experiments with their associated results can be found in [**?**], while Table III shows the configuration leading to the best results of each method. The results are reported in terms of SE and SP (see Sect. IV-B).

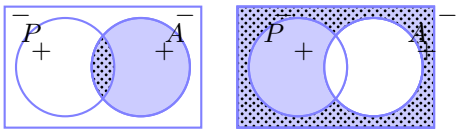
Lemaitre *et al.* achieve the best results when using LBP-TOP features, a global mapping, and histogram representation [**?**]. Alsaih *et al.* perform better when using HOG features with PCA representation [**?**]. Our interpretation of Liu *et al.*,

TABLE II
SUMMARY OF THE CLASSIFICATION PERFORMANCE IN TERMS OF SE AND SP IN (%).

	Lemaitre *et al.* [**?**]	Sankar *et al.* [8]	Alsaih *et al.* [**?**]	Srinivasan *et al.* [**?**]	Liu *et al.* [**?**]	Venhuizen *et al.* [**?**]
SE	87.5	81.3	75.0	68.8	68.8	61.5
SP	75.0	62.5	87.5	93.8	93.8	58.8



(a) Confusion matrix with truly and falsely positive samples detected (TP, FP) in the first row, from left to right and the falsely and truly negative samples detected (FN, TN) in the second row, from left to right.



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

(b) SE and SP evaluation, corresponding to the ratio of the doted area over the blue area.

Fig. 3. Evaluation metrics: (a) confusion matrix, (b) SE - SP

as proposed in Sect. IV, achieves the best results when using majority voting instead of BoW models. Refer to Table III for configuration details of the remaining methods.

Results in [**?**] indicate two main findings with major impact: (i) features describing the entire volume rather than each B-scan are more discriminative; and (ii) a pre-processing stage with denoising is fundamental.

Other observations include the facts that (i) to represent B-scans, local mapping in conjunction with dimension reduction, either using PCA or BoW, improve the results. However, the combination of both decreases the performance in comparison to non reduced histogram representation; (ii) building BoW models from concatenated detected features, might lead to the curse of dimensionality, which would explain why the RBF-SVM overfits in [**?**]; (iii) building BoW models from the concatenation of all features for each B-scan, might lead to the curse of dimensionality since 128 samples per volume is not enough to describe a space with a number of dimensions of the order of thousands; which could explain the over-fitting using RBF-SVM as in [**?**].

VI. CONCLUSION AND FURTHER WORK

The work here presented states the relevance of developing methodologies to automatically differentiate DME *vs.* normal SD-OCT scans. This article offers an overview of the state-of-the-art of DME detection and provides a public benchmarking to facilitate further studies. In this regard, there are two crucial aspects to improve the work here presented: (i) enlarge the dataset. (ii) reach out to other authors in order to enlarge this benchmark with additional methods and improve the existing approaches.

REFERENCES

[1] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes estimates for the year 2000 and projections for 2030," *Diabetes care*, vol. 27, no. 5, pp. 1047–1053, 2004.

[2] Early Treatment Diabetic Retinopathy Study Group, "Photocoagulation for diabetic macular edema: early treatment diabetic retinopathy study report no 1," *JAMA Ophthalmology*, vol. 103, no. 12, pp. 1796–1806, 1985.

[3] Y. T. Wang, M. Tadarati, Y. Wolfson, S. B. Bressler, and N. M. Bressler, "Comparison of Prevalence of Diabetic Macular Edema Based on Monocular Fundus Photography vs Optical Coherence Tomography," *JAMA Ophthalmology*, pp. 1–7, Dec 2015.

[4] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in sd-oct images congruent with expert manual segmentation," *Optic Express*, vol. 18, no. 18, pp. 19 413–19 428, 2010.

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

[6] J. S. Schuman, C. A. Puliafito, J. G. Fujimoto, and J. S. Duker, *Optical coherence tomography of ocular diseases*. SLACK incorporated, 2004.

[7] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, C. A. Toth, A.-R. E. D. S. . A. S. D. O. C. T. S. Group *et al.*, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, 2014.

[8] S. Sankar, D. Sidibé, Y. Cheung, T. Wong, E. Lamoureux, D. Milea, and F. Meriaudeau, "Classification of sd-oct volumes for dme detection: an anomaly detection approach," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2016, pp. 97 852O–97 852O.

[9] M. L. Giger, H.-P. Chan, and J. Boone, "Anniversary paper: History and status of CAD and quantitative image analysis: the role of medical physics and AAPM," *Medical physics*, vol. 35, no. 12, p. 5799, 2008.