An investigation of the annotated data sparsity problem in the medical domain

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Electronics & Communication Engineering by Research

by

Pujitha Appan Kandala
201231211
pujitha.ak@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
Nov 2018

International Institute of Information Technology
Hyderabad, India

CERTIFICATE OF AUTHORSHIP

I, Pujitha Appan Kandala, declare that the thesis, titled 'An investigation of the annotated data sparsity problem in the medical domain', and the work presented herein are my own. I confirm that this work was done wholly or mainly while in candidature for a research degree at IIIT-Hyderabad.

_____ _____

Date Signature of the candidate

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled 'An investigation of the annotated data sparsity problem in the medical domain' by Pujitha Appan Kandala, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. Jayanthi Sivaswamy

Dedicated to my beloved family for their love and support

Acknowledgments

I would like to take this opportunity to express my gratitude to everyone who helped me make my thesis work a success. First, I would like to thank my advisor Prof. Jayanthi Sivaswamy for her constant guidance. She taught me never to give up and always had my back and pushed me through difficult times. Her suggestions and ongoing assessments at weekly meetings have made me go a long way in terms of studies and self-confidence. I was fortunate to have her as my advisor and, to benefit from her vast knowledge and experience. I would also like to thank Prof. Priyanka Srivatsava for her valuable insights on the perception study.

I am grateful to Ujjwal, Tabish, Samruddhi, Karthik, Raghav, Jahnavi and Arunava for the productive discussions and help in my research. I would also like to thank the entire MIP group for providing a comfortable and positive environment that helped me focus on my research.

Without a doubt, the decision to join IIIT was the best decision I ever made. University life in the past five years has been amazing, leaving many memories. This beautiful journey wouldn't have been possible without my friends Anirudh, Sneha, Gayatri, Sharvani, Swetha and Srikar. I especially thank Anirudh and Sneha who helped me maintain calm and rationality in harsh conditions. I am thankful to all my batch-mates for the memorable trips, night walks and long conversations we have had.

I have a deep gratitude to my parents for their continued support and motivation to complete my thesis. Without their assistance, I would not have the ability to finish much of what I have done and become what I am now.Special thanks to my sister who is a doctor and thus helped me a lot to understand medical topics in depth. I am also thankful for her love and support. Last but not the least, I am thankful to God for giving me the ability, knowledge, and strength to undertake this research study and complete it satisfactorily.

Abstract

Diabetic retinopathy (DR) is the most common eye disease in people with diabetes. It affects them for significant number of years and can also lead to permanent blindness if left untreated. Early detection and treatment of DR is of utmost importance for the prevention of blindness. Hence, automatic disease detection and classification have been attracting much interest. High performance is critical in adoption of such systems, which generally rely on training with a wide variety of annotated data. Availability of such varied annotated data in medical imaging is very scarce. The main focus of this thesis is to deal with the sparsity of annotated data and develop computer-aided diagnostic CAD systems which take less annotated data and yet give high accuracies. We propose three different solutions to address this problem.

First, we propose a semi-supervised framework which paves way for including unlabeled data in training. A co-training framework is used in which features are extracted from a limited training set and independent models are learnt on each of the features, later the models are used to predict labels for new data. The highly confident labelled images from unlabelled set are added back to the training set and the process is continued, thus expanding the number of known labels. This framework is showcased on retinal neovascularization (NV) which is a critical stage of proliferative DR. The analysis of the results for detection of NV showed that an AUC of 0.985 with sensitivity of 96.2% at specificity of 92.6% which were superior to the existing models.

Secondly, we propose crowdsourcing as a solution where we obtain annotations from a crowd and use them for training after refining. We employ a strategy to refine/overcome the noisy nature of crowdsourced annotations by i) assigning a reliability factor for each subject of the crowd based on their performance (at global and local levels) and experience and ii) requiring region of interest (ROI) markings rather than pixel-level markings from the crowd. We also show that these annotations are reliable by training a deep neural net (DNN) for detection of hard exudates which occur in mild non-proliferative DR. Experimental results obtained for hard exudate detection showed that training with refined crowdsourced data is effective as detection performance improves by 25% over training with just expert-markings.

Lastly, we explore synthetic data generation as a solution to address this problem. We propose a novel method, based on generative adversarial networks (GAN), to generate images with lesions such that the overall severity level can be controlled. We showcase this approach for hard exudate and haemorrhage detection in retinal images with 4 levels of severity. These vary from mild to severe non-proliferative

DR. The synthetic data were also shown to be reliable for developing a CAD system for DR detection. Hard exudate/ haemorrhage detection was found to improve with inclusion of synthetic data in the training set with improvement in sensitivity of about 25% over training with just expert marked data.

Contents

List of Figures

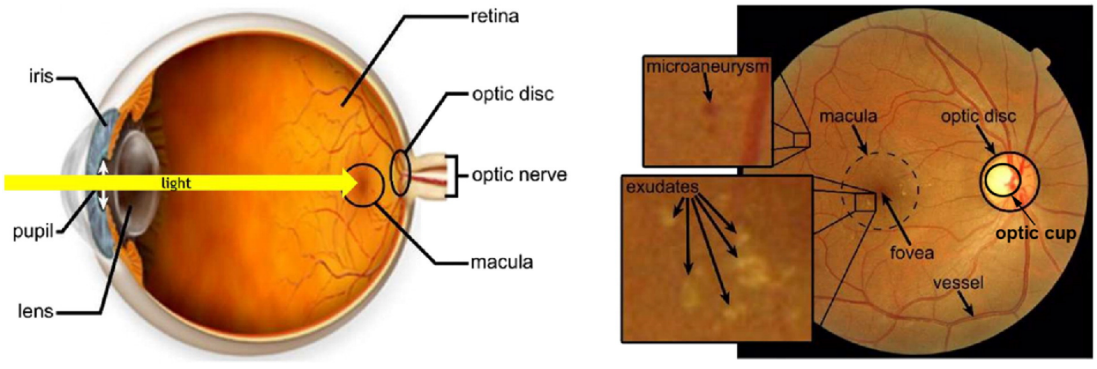
List of Tables

Chapter 1

Introduction

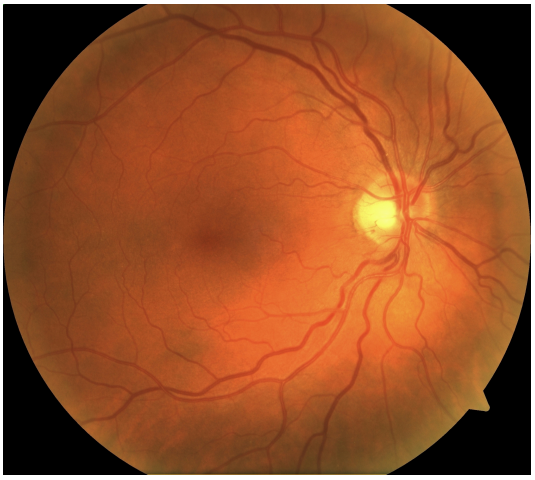
1.1 Retinal Imaging

Medical imaging captures visual portrayals of the interior parts of our body that are not visible to the naked eye. Medical imaging is an important basis to diagnose diseases and encompasses many imaging modalities such as X-RAY, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, retinal imaging, optical coherence tomography (OCT) etc. These techniques produce images of the human body which can be used for diagnosis and treatment. Here, we are particularly interested in retinal imaging. A widely used part of retinal imaging is the fundus image which captures the photograph of the back of the eye. Specialized cameras with flash attached are used, and the rear of the eye is captured through the pupil. The image captures important structures like optic disc, macula and the blood vessels. The retina and its rear projection on to an image can be seen in the Fig. 1.1

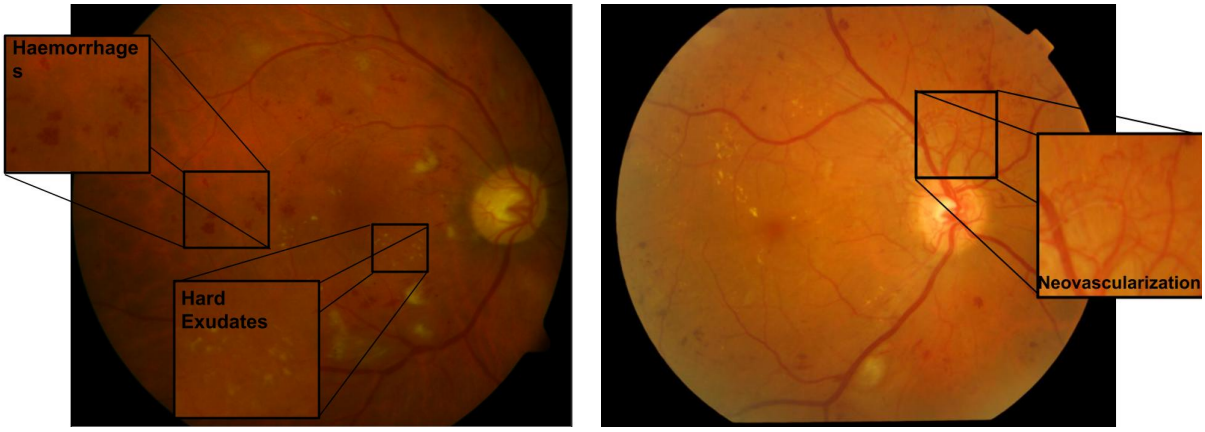


(a) (b)

Figure 1.1: (a) structure of human eye and (b) projected fundus image showing the main structures inside human eye



(a)



(b) (c)

Figure 1.2: (a) Normal retinal image (b) Retinal image containing haemorrhages and hard exudates (c) Retinal image containing neovascularization

The fundus image can reveal presence of many diseases related to the eye such as age-related macular degeneration (AMD), glaucoma, diabetic retinopathy, diabetic macular edema (DME), retinal detachment etc. In this thesis we mainly focus on diabetic retinopathy (DR).

DR causes damage to the eye and is the major cause of blindness. It is generally caused in people with diabetics. More than 400 million people in the world have diabetes, among these more than half of the cases are noticed in older people and about one third of them report having DR. The symptoms of DR are blurring of vision, difficulty in perceiving colors, eye pain, double vision, etc. But these symptoms occur at advance stages. Recently, diabetics affected patients have been given laser treatment to reduce the occurrence of blindness. The laser treatment also helps only when done at the appropriate stage. Hence, it is of utmost importance to identify the changes in retina and treat it immediately.

Retinal imaging offers a safe and non-destructive way to observe any changes in the retina. The retinal images captures the characteristics of these diseases which can be viewed by the doctor. The

diseased eye can have hard exudates which are visible as yellowish blobs caused due to lipid leakage, haemorrhages which are reddish blobs formed due to leakage of blood from rupture of a blood vessel, neovascularization which is due to the formation of new thin blood vessels. Sample retinal images which contain these diseases are shown in Fig. 1.2.

Automatic diagnosis and identification of these diseases from the medical images is possible with a system known as computer aided diagnosis (CAD). Nowadays these CAD systems are built on a machine learning framework due to its success on many computer vision tasks. This has motivated exploration of ML in wide ranging of medical applications from disease detection [19] to segmentation [9]. The ML framework's success is contingent on abundance of training data *with* expert annotations. Acquisition of expert annotations has always been difficult in the medical domain given the tedium of the task and the priority patient care takes over the annotation task.

1.2 Data sparsity

The annotations can be obtained at three different levels: image-level, region/local-level and pixel-level annotations. The image-level annotations give information on whether the image is associated with a particular abnormality, local-level annotations indicate the location of the abnormality, pixel level annotations classify each pixel as belonging to the abnormality or the background. The different types of annotations are shown in Fig. 1.4. From this we can infer that the increasing difficulty level of generating these annotations is as follows: image-level (least difficult), local-level (medium difficulty) and pixel-level (most difficult). The difficulty level of annotations also depends on different factors such as the number, size of lesions in the image and the conspicuity of the lesions. As the number of lesions increase, the time to annotate increases. It is also hard to view lesions which are very small or are close to the boundaries of the image, making it more difficult to annotate. Due to the difficulty of local/pixel-level annotation, majority of the public datasets available are annotated only at image-level. This leads to the sparsity of annotations at local-level and pixel-level. Different datasets and the annotations available are shown in Table. 1.1.

1.2.1 Staging

In general retinal images are graded based on the abnormality present and a stage is assigned. Staging is very important in detection of DR as it helps in determining the severity of the disease. DR is mainly of two types, proliferative and non-proliferative. The presence of neovascularization, growth of abnormal blood vessels indicates proliferative DR (PDR). Early disease detection without the growth of new blood vessels is known as non-proliferative DR (NPDR). NPDR is a precursor to PDR stage with increased severity of the disease. There are different stages in NPDR depending on the number of haemorrhages, hard exudates and microaneurysms present in the retinal image. The early treatment diabetic retinopathy study (ETDRS) grading for DR is shown in Fig. 1.3.

Table 1.1: Popular public datasets

Datasets	No. of images	Image-level (Staging)	Local-level	Pixel-level
DIARETDB-0	130	Yes (No)	-	-
DIARETDB-1	89	Yes (Yes)	Yes	-
MESSIDOR	1200	Yes (Yes)	-	-
Kaggle	$\sim$ 30000	Yes (Yes)	-	-
DRiDB	31	Yes (Yes)	Yes	-

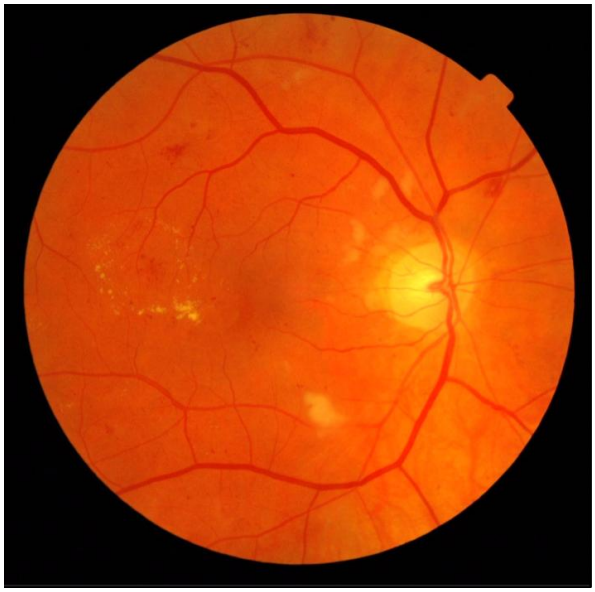
Measure	Score	Observable Findings
ICDR severity level		
No apparent retinopathy	0	No abnormalities (Level 10 ETDRS)
Mild non-proliferative diabetic retinopathy	1	Microaneurysm(s) only (Level 20 ETDRS)
Moderate non-proliferative diabetic retinopathy	2	More than just microaneurysm(s) but less than severe non-proliferative diabetic retinopathy (Level 35, 43, 47 ETDRS)
Severe non-proliferative diabetic retinopathy	3	Any of the following: > 20 intra-retinal haemorrhages in each of 4 quadrants, definite venous beading in $\geq$2 quadrants, prominent intra-retinal microvascular abnormalities in $\geq$1 quadrant, or no signs of proliferative retinopathy. (Level 53 ETDRS: 4-2-1 rule)
Proliferative diabetic retinopathy	4	One or more of the following: neovascularization and/or vitreous or preretinal haemorrhages. (Levels 61, 65, 71, 75, 81, 85 ETDRS)
Macular oedema severity level		
No macular oedema	0	No exudates and no apparent thickening within 1 disc diameter from fovea
Macular oedema	1	Exudates or apparent thickening within 1 disc diameter from fovea

Abbreviations: ETDRS, Early Treatment Diabetic Retinopathy study; ICDR, International Clinical Diabetic Retinopathy

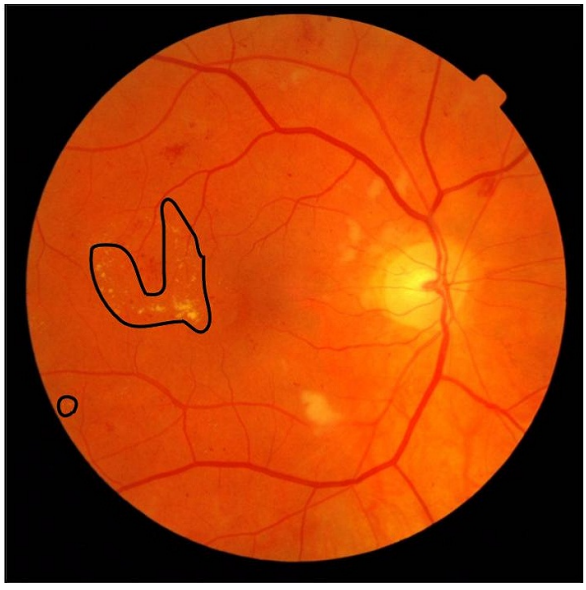
doi:10.1371/journal.pone.0139148.t001

Figure 1.3: ETDRS grading protocol

For staging we need to know the location of the lesions according to the ETDRS grading. Hence, we concentrate on generating the local-level annotations in the entire thesis. As we have seen in Table. 1.1, only a few public datasets have local annotation leading to data sparsity. A popular solution to address this data sparsity issue is data augmentation (via geometric transformations) which is adopted by the computer vision community. However, this has limited success in the medical domain as it does not introduce any real variability that is essential for robust learning of abnormalities, normal anatomy etc.



(a)



(b)

(c)

Figure 1.4: (a) Image-level annotation - abnormal (contains hard exudate) (b) local-level annotation - location of hard exudate (c) pixel-level annotation - pixels belonging to the hard exudate.

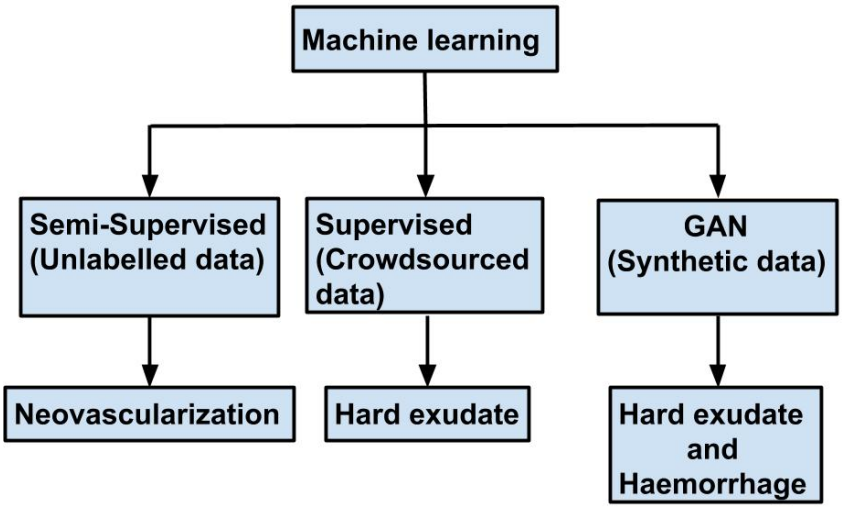


Figure 1.5: Methods followed to address the data sparsity issue

We address this problem of data sparsity using different types of machine learning methods. There are three main approaches in machine learning to build a CAD system: (i) semi-supervised training, (ii) supervised training and (iii) unsupervised learning. Here we focus on semi-supervised, supervised and a system which combines both supervised and unsupervised approaches of machine learning. A semi-supervised approach is one where we use both unlabelled and labelled data for training. This type of approach was used in the first part of the thesis for detection of neovascularization. A supervised approach uses labelled data for training a CAD system. We have conducted a crowdsourcing experiment and considered a heterogeneous mixture of annotations (crowd and expert) to train a supervised model in the second part of the thesis for detection of hard exudates. Finally we have utilized a generative adversarial network (GAN) architecture which combines a supervised and unsupervised approach for detection of haemorrhages and hard exudates in the last part of the thesis.

1.3 Thesis Focus

The main focus of the thesis is to address the issue of data sparsity in medical domain in the context of developing CAD solutions. The solutions for addressing the data sparsity issue are mainly showcased on retinal images. In the first part of the thesis, we develop a solution using semi-supervised approach, showcasing on neovascularization (Fig. 1.2(c)), second part focuses on relying on crowdsourced annotations showcasing on hard exudates and final part address this problem by synthetically generating retinal images showcasing on hard exudates and haemorrhages (Fig. 1.2(b)).

1.3.1 Contributions

The contributions of the thesis are as follows. We have looked at different ways in which the sparsity of data annotations can be resolved. We have used supervised and semi-supervised methods for developing CAD systems.

1. A co-training based semi-supervised approach for neovascularisation detection paves way for including unlabeled data in training.

2. A CAD system which uses refined crowdsourced annotations to detect the presence of DR lesions. A strategy is also proposed to overcome the noisy nature of crowdsourced annotations.

3. A generative adversarial network (GAN), to generate images with lesions such that the severity level of the disease can be controlled. Further the generated synthetic images have proved to be reliable by using then in training a computer aided diagnosis (CAD) system for lesion detection in retinal images

1.4 Organization of the thesis

The thesis is organized as follows: the solutions to the data sparsity problem addressed in chapter-2 using semi-supervised approach which utilizes unlabelled data, in chapter-3 using crowdsourced annotations and in chapter-4 using generative adversarial networks for synthetic image generation. The conclusion and future work is given in chapter-5.

Chapter 2

Semi-supervised learning using unlabelled data for CAD development

2.1 Introduction

Advanced stages of Diabetic Retinopathy (DR) is marked by the formation of new, weak and thin microvascular networks, a phenomenon known as Neovascularization (NV). NV increases the risk of fibrosis, bleeding and ultimately loss of vision and its detection is therefore of interest. However, existing literature is predominantly devoted to detection of lesions that arise in the earlier stage of DR.

Existing approaches for NV detection typically follow a classification route, where the features are either extracted from the segmented vessel map [5] [4] [16] or directly from the raw image [2] [35] [3] [20] [47]. A wide variety of features and their combinations have been used in both approaches. These are extracted in all but one method [2] at a patch level. Features considered in the first approach include a combination of shape, intensity and gradient features [5]; multiscale AM-FM [4]; shape, position, orientation, intensity and line density [16]. In the second approach, features that have been explored include morphological [2], morphological and GLCM [35], multiscale AM-FM [3]; vesselness, power spectrum distribution [20]; LBP and multi-scale Counterlet transform [47]. The final classification is done using LDA [5], SVM [4] [16] [2] , hierarchical clustering [35], random forest classifier [3] [20] and Artificial Neural Network [47].

Accurate vessel segmentation, specifically of thin vessels, is critical to the first type of approach while adequate training data and hand crafting of features is required for both approaches. Large public access datasets for NV detection provide image-level annotation whereas the existing detection methods require patch level labels whose generation is laborious and hence not scalable. Consequently, most methods depend on locally sourced annotations and report only on a selected set of images from public datasets. We propose a patch-based NV detection method which i) uses generic features thereby eliminating the need for hand crafting and ii) leverages the availability of large amount of unlabeled data by employing a co-training based semi-supervised framework. We show that co-training with generic features such as vesselness and oriented local energy leads to consistently good performance on nearly 3000 images from 4 public datasets.

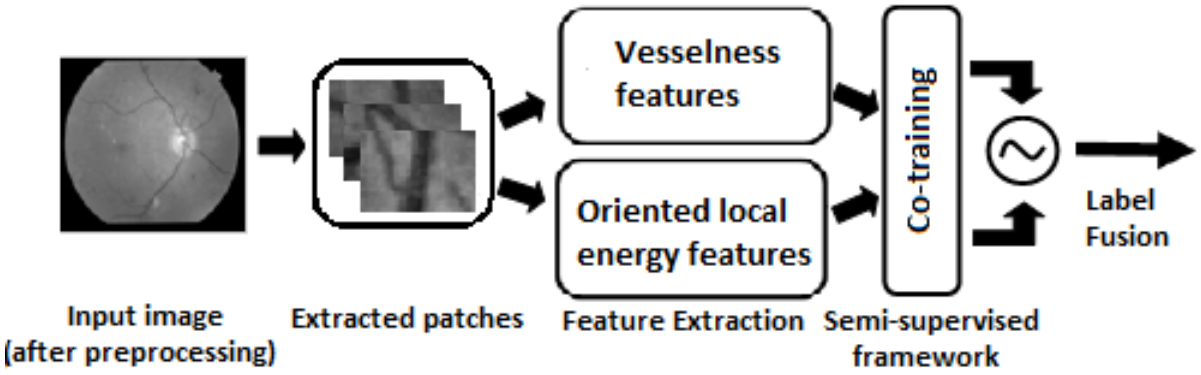


Figure 2.1: Proposed system for NV detection

2.2 Method

The proposed method consists of four stages: pre-processing, feature extraction, co-training and label fusion as shown in Fig. 2.1. All processing was restricted to the green plane of the given image. The stages are described in detail next. As retinal images suffer from non-uniform illumination, this is corrected using luminosity and contrast normalization [24] as a pre-processing step.

2.2.1 Feature extraction

Co-training requires two independent feature spaces to represent a patch. These features also have to be discriminative for NV and Non-NV patches. Two types of features are chosen to satisfy these requirements: (i) a Hessian based vesselness feature which is popular for vessels and (ii) a more generic one for oriented structures to capture coarse texture.

2.2.1.1 Vesselness based features

The vesselness is computed on the basis of eigenvalues of the Hessian as described in [13]. The probability of an image region to contain vessels or other ridges is found based on the eigenvalues and the vesselness at every point X: (x,y) in a patch at a scale s is computed as:

$$v(X,s) = \begin{Bmatrix} 0 & if \quad \lambda_2 > 0, \\ exp(-\frac{\lambda_1^2}{2\lambda_2^2\alpha^2})(1-exp(-\frac{\lambda_1^2+\lambda_2^2}{2c^2})) & otherwise \end{Bmatrix} \quad (2.1)$$

Here, α and c are thresholds; $\lambda_i; i \in \{1,2\}$, are eigenvalues of the Hessian matrix computed at X.

The final vesselness map is obtained by taking the maximum response across all scales. A sample NV patch and the derived vessel map are shown in Fig. 2.4(a,b). This map is row vectorized to form the feature vector g_{vm}.

2.2.1.2 Oriented local energy features (OLE)

NV is characterized by texture at multiple scales and orientations. This texture can be represented by local energy which is defined as a sum of squared responses of a pair of conjugate symmetric filters. A Gabor filter bank is a standard way to compute local energy as it aids determining responses at different orientation and scales. We choose the log Gabor kernel in the frequency (f) domain:

$$K(f,\theta) = exp(-\frac{(log(\frac{f}{f_0}))^2}{2(log(\frac{\sigma_f}{f_0}))^2})exp(-\frac{(\theta-\theta_0)^2}{2\sigma_\theta^2}) \tag{2.2}$$

Here f_0 is the central radial frequency, θ_0 is the orientation of the filter, σ_θ and σ_f are the angular and radial bandwidths.

The OLE at every point X in the patch is computed as:

$$E_{\theta_0}^{f_0}(x,y) = \sqrt{(R_{\theta_0}^{f_0,even}(x,y))^2 + (R_{\theta_0}^{f_0,odd}(x,y))^2} \tag{2.3}$$

Here $R_{\theta_0}^{f_0,even}$ and $R_{\theta_0}^{f_0,odd}$ are the responses of even and odd symmetric log Gabor filters. The total energy of entire patch of size m xn, at a specific θ_0 is found as:

$$\hat{E}(\theta_0, f_0) = \sum_{x=1}^{m}\sum_{y=1}^{n} E_{\theta_0}^{f_0}(x,y) \tag{2.4}$$

This can be considered as the histogram function of energy map which expresses the oriented energy at different scales. The histogram is normalized by dividing it by maximum energy over all orientations at particular scale. The final feature vector g_{ole} is derived by concatenating the energy histogram at different orientations and scales. Fig. 2.4 shows sample Non-NV (d) and NV patches (e) and their energy histograms (c). The two types of patches are clearly distinguishable with the energy plots for NV patches (in red) having higher energy on average.

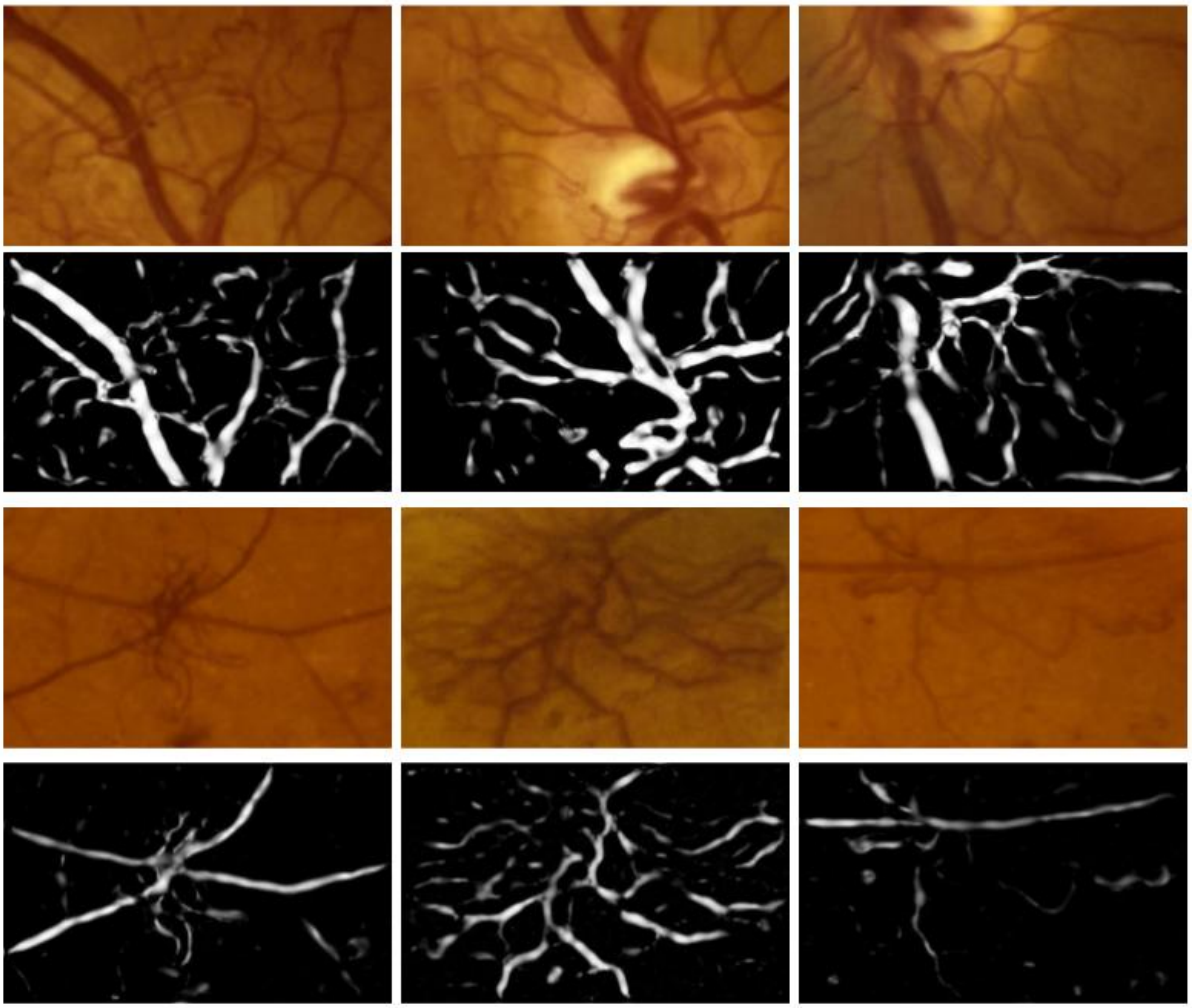


Figure 2.2: Sample patches containing neovascularization and their corresponding vesselness feature

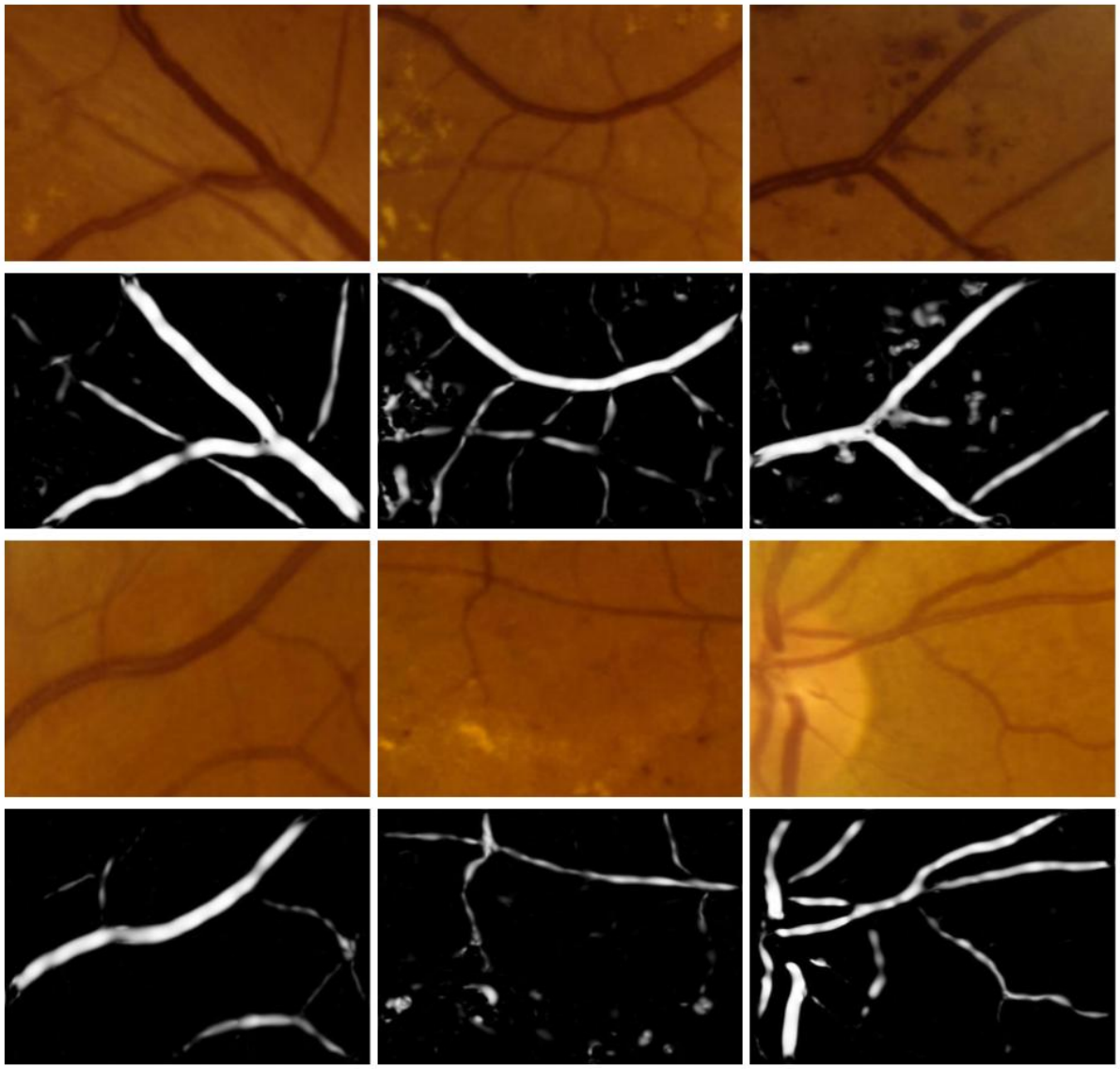


Figure 2.3: Sample normal patches which do not contain neovascularization and their corresponding vesselness feature

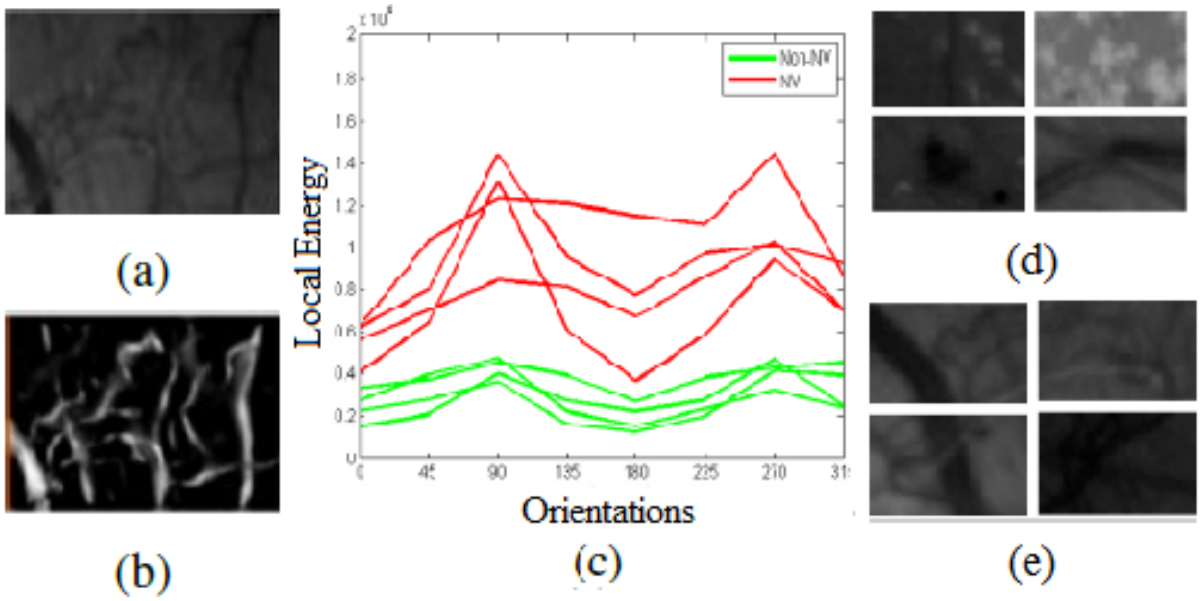


Figure 2.4: Feature representation for patches. (a) NV patch and its (b) vesselness map; (c) Oriented local energy histogram for (d) Non-NV and (e) NV patches.

2.2.2 Co-training

Availability of annotations is limited in the medical domain as it is tedious to generate them and training with this limited data has the potential to over-fit and lacks robustness. Hence, the proposed method uses co-training [8] to utilise unlabeled images which are more widely available. Co-training works best with independent features. Table.2.1 shows the normalized values obtained after applying PCA on the features extracted from sample patches. Fig. 2.5 shows sample patches consisting of all possible combinations of following structures relevant to DR: vessels, NV, hemorrhages, hard exudates and background.The last three are irrelevant to NV detection and can be seen to be marked by low values of g_{vm}. Our feature choice satisfies the requirement for co-training as, the pure vessel (denoted as V) and NV (denoted as NV+V) patches are adequately separated from each other, as well as from other patches.

Table 2.1: Feature Space Representation Values

	NV+V	**V**	**HEM+V**	**HE+V**	**HEM**	**HE**	**BG**
g_{vm}	1	0.8	0.65	0.25	0.02	0.01	0.01
g_{ole}	0.9	0.02	0.7	0.17	0.37	0.7	0.02

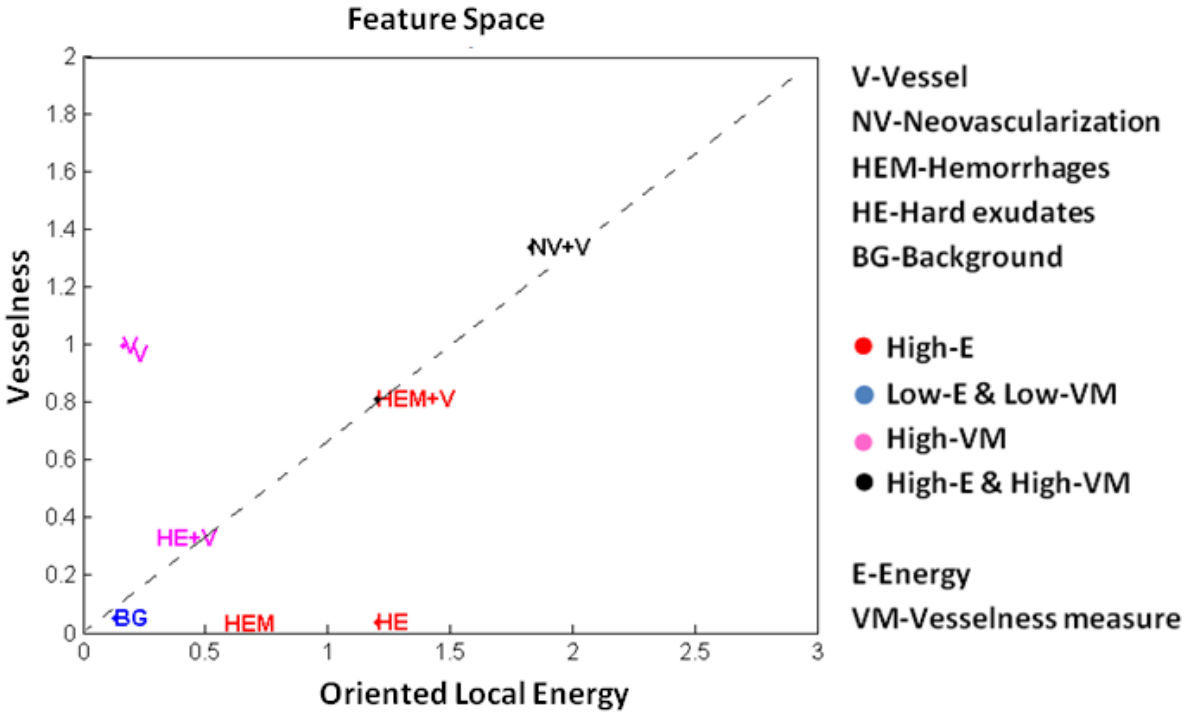


Figure 2.5: Feature Space Representation

Hence, OLE and vesselness based features can be used to independently predict the confidence score for a patch with NV. Two classifiers C_1 and C_2 are trained separately on these features using gradient boosting [21] which is a way to design a strong classifier by fusing weak classifiers. The resultant strong classifier is used to predict the class probabilities.

In the co-training framework, let L be a set of labeled data (-1 for Non-NV and +1 for NV patches) and U be the set of unlabeled data. A classifier h_1 is trained using g_{vm} and classifier h_2 is trained using g_{ole}. The trained classifier h_1 is used to label the unlabeled data. The most confidently predicted patches (p positive and n negative) are fed back to the training set to update h_2. Similarly, the most confidently predicted patches by h_2 are fed back to update h_1. This process is repeated until all the unlabeled patches are labeled.

2.2.3 Label fusion

The labels obtained from co-training are fused using nearest neighbour information in feature space . A rule-based voting system is considered for fusion as described. In a test patch, for each feature, the K nearest features are extracted from the updated training set. Each retrieved similar feature, along with its labels and weights, contributes a vote. The weights are computed as :

$$\widehat{w}_{g_i} = exp(\frac{-d(x_{g_i}, \widehat{y}_{Kg_i})}{\sigma^2}) \tag{2.5}$$

Here $g_i \in \{g_{vm}, g_{ole}\}$, $d(x_{g_i}, \widehat{y}_{Kg_i})$ is the Euclidean distance between the test feature x_{g_i} and K nearest training features $\widehat{y}_{Kg_i}$. The aggregation of votes results in a probabilistic decision as given:

$$P(x, y_K) = \frac{\widehat{w}_{Kg_{vm}} \widehat{l}_{Kg_{vm}} + \widehat{w}_{Kg_{ole}} \widehat{l}_{Kg_{ole}}}{\sum_{k=1}^{K} (\widehat{w}_{kg_{vm}} \widehat{l}_{kg_{vm}} + \widehat{w}_{kg_{ole}} \widehat{l}_{kg_{ole}})} \tag{2.6}$$

Here $\widehat{l}_{Kg_i}$ denotes K training labels of feature g_i. The probabilities obtained are thresholded to get the final class labels.

2.3 Datasets

Four datasets (1 private and 3 public) are considered for the evaluation of the proposed method. KPHDR [47] is a private dataset with local annotations (only for NV). The patch level ground truth needed for evaluation was obtained such that atleast 30% of the patch was annotated as having NV. All 3 public datasets for DR , namely, MESSIDOR [12], Kaggle [1] (a challenge set) and DIARET-DB0 [27] provide only image-level annotations. MESSIDOR and Kaggle have 4-level annotations regarding DR severity and albeit different, which are roughly: No DR, mild, moderate, severe and PDR. DIARET-DB0 provides annotations in terms of various abnormalities present in an image. Images from all datasets were re-sized to $\sim$0.8 times the given image size (maintaining the aspect ratio) prior to patch extraction for computational efficiency. A window of 100x150 dimensions was used to divide the image into patches with a stride of 75 pixels.

2.3.1 Training and Testing datasets:

KPHDR has a total of 2575 abnormal patches. Training/testing sets were constructed for this dataset with 3000/55765 patches with 1:1 ratio of NV:Non-NV for training to address class imbalance problem. The data samples chosen fortraining, testing and unlabelled data are disjoint sets and the details are reported in Table. 2.2.

2.4 Experiments and Results

2.4.1 Experiments

Log-Gabor filters were considered at 12 scales and 24 orientations as given in equation 2.2. Vesselness maps were computed at 5 scales with sigma varying between 1 to 10 (step size 2). Given an image patch, oriented local energy feature (of size 288) and vesselness based feature (of size 15000) are extracted to derive patch level predictions.

Training was done on features extracted from a set of 3000 patches (L) selected randomly from KPHDR dataset, 1:1 ratio of Non-NV and NV patches. The unlabeled (U) dataset of 4000 patches is

taken from the remaining 3 datasets (1:1:1 ratio from each dataset). Gradient boosting for training was done with exponential loss function and decision trees were used as base-learners. The shrinkage factor which reduces the impact of potentially unstable regression coefficients was varied for different values and finally set to 0.1. For each iteration, the unlabeled dataset was tested using the obtained models, from which samples with top 10 confidence score of both models were used for updating the training set.

Table 2.2: Total number of data used for evaluation

Dataset	**Training**	**Unlabelled** $N_I : N_P$	**Testing** $N_I : N_P$
KPHDR	3000*	-	55765*
MESSIDOR	-	12 : 1296	469 : 46431
DIARET-DB0	-	7 : 1309	34 : 6358
Kaggle	-	29 : 1305	2081 : 93645
Total	3000*	48 : 3910	2584 : 202199

*number of patches; N_I, N_P denotes the number of images and the number of patches extracted from those images respectively.

Image level decision of NV/Non-NV was done by considering patch level predictions on the entire image and thresholding the number of NV patches detected out of total patches in the image. Experimentally this threshold was determined to be 2%-4% of total number of patches.

2.4.2 Results :

The performance was assessed using the following evaluation metrics: Sensitivity (SN), Specificity (SP), Receiver Operating characteristics (ROC) and Area Under Curve (AUC). To underscore the contribution from co-training, the obtained values are reported without/with co-training (I/II) in Table. 2.3.

The tabulated results show that the average SN and SP values *without* co-training are comparable to other methods demonstrating the effectiveness of the selected features which are generic in nature. Overall, we can note that co-training results in an improvement of 5%-11% in NV detection performance.

The performance metrics reported by 4 other methods are listed in Table 2.4. As mentioned in the introduction, the testing results have been reported in literature not on the entire dataset, but on a selected number of patches/images as indicated in the table. Hence, a direct comparison is not possible.

The proposed method has SN/SP of 96.2/92.6% on KPHDR, which appears to be lower than the best results of 99.62/96.61% reported in [47]. However, the test set sizes for these results differ by several orders (200). Likewise, on the MESSIDOR dataset, the proposed method achieves a SN/SP of 97.56/93.4% against 98/97% reported in [5], however there is roughly a 5-fold difference in the test set

Table 2.3: NV detection results for the proposed method, without (I) and with co-training (II) on 4 datasets.

Dataset	SN(%)		SP(%)		AUC	
	(I)	(II)	(I)	(II)	(I)	(II)
KPHDR**	89.8	96.2	92.7	92.6	0.9378	0.9850
MESSIDOR*	97.56	97.56	82.5	93.4	0.9659	0.9877
DIARET-DB0*	100	100	86.38	90.91	0.9605	0.9868
Kaggle*	92.91	92.73	74.8	91.25	0.9274	0.9725
Average*	96.82	96.76	81.22	91.85	0.951	0.982

**at patch level; *at Image level

sizes. Even though Kaggle has high inter-image variations, the proposed method is able to perform well. Thus, we can say that this method is robust as well as superlative in performance.

Table 2.4: NV detection results of existing methods

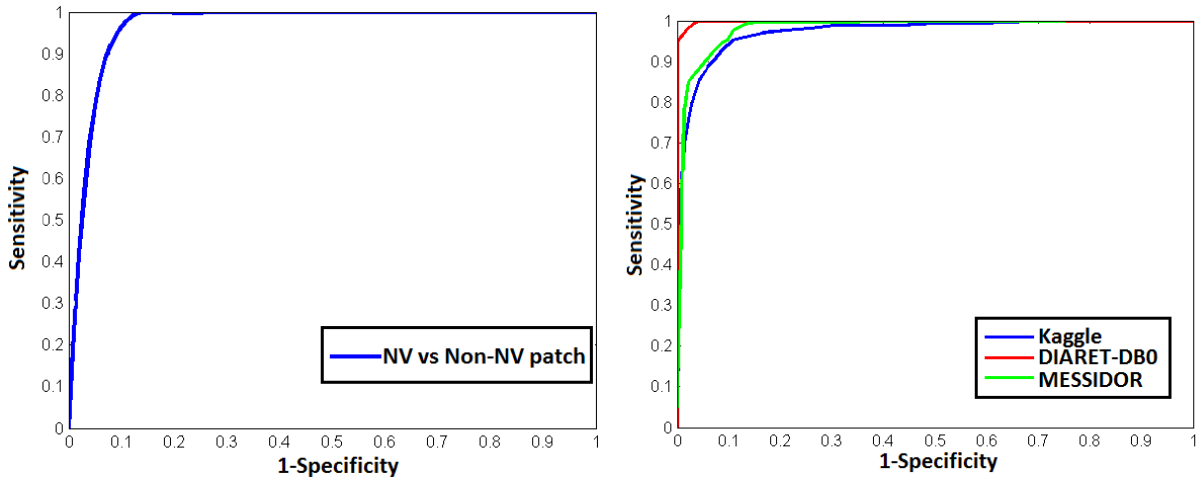
Method	Dataset	Test Samples	SN(%)	SP(%)	AUC
[45]	**KPHDR**	200 patches	94	85	0.92
[47]	**KPHDR**	322 patches	99.62	96.61	-
[20]	**MESSIDOR**	98 images	100	87	0.98
[5]	**MESSIDOR**	130 images	98	97	-

Patch level ROC was computed by varying the threshold on the probabilities obtained after label fusion and is shown in Fig. 2.6(a). Image-level NV detection was also analyzed and the ROC was derived by varying thresholds on the number of abnormal patches detected in a given image (see Fig. 2.6(b). The proposed method achieves an average AUC of 0.982 over 4 datasets which indicates consistency and robustness.

2.5 Concluding Remarks

Automatic detection of NV is a difficult task as it is characterised by complex texture changes and learning is impeded by limited availability of annotated data. These constraints are overcome in our proposed method with the use of vesselness and Gabor features along with co-training. Co-training was seen to improve the AUC value from 0.95 to 0.98 which is significant. Further improvements are possible with appropriate selection of unlabeled data during co-training. Consistent good performance of the method across datasets (over nearly a quarter million patches), at both patch and image levels,

depicts its robustness to changes in resolution, illumination, tissue type (due to population difference) and noisy conditions. These results demonstrate that the proposed system can be used in automated detection and grading of DR.



(a) (b)

Figure 2.6: ROC curve for predictions at the (a) patch level (KPHDR) and (b) image level (other 3 datsets).

Chapter 3

Crowdsourced annotations as an additional form of data augmentation for CAD development

3.1 Introduction

Crowdsourcing has been considered as a solution to address the issue of sparsity of annotated data. It has been shown to be reliable [14, 30, 34] and useful to train classifiers [29]. In [14, 30, 34], annotations were crowd-sourced from fundus images, endoscopy and MRI of brain, while in [29], crowd-sourced data was explored to train a random forest to segment surgical instrument from Laparoscopic images. Recent work has examined the utilization of such crowdsourced data for machine learning further [31] [6]. Active learning is the mode of choice of these approaches. Accordingly, only low confident samples predicted by a model are given to the crowd and their annotations are fedback to update the model. Atlas forests are used in [31] and based on crowd refined annotations (on instrument boundary), a new atlas is generated and added to the forest. Similarly, a convolutional neural network (CNN) is trained in [6] and the crowdsourced mitosis candidates (in a patch of size 33×33) are merged with an aggregation layer for updating the model. The issue of merging crowd annotations for an image to derive a single ground truth (GT) for training a model is an important challenge to overcome the inherently noisy nature of the crowdsourced annotations. Methods for merging ranges from simple Majority Voting (MV) [29] to a stochastic modeling of the crowdsourced information using Expectation Maximization [30] and introducing an aggregation layer in a CNN [6].

Involving the crowd in an active learning mode requires some synchronization between the crowd and model training, which is not always possible in a real-world scenario. Further, the types of annotations to be collected have implications. Pixel level markings are tedious while patch level labeling requires patch selection by a model/human. A high initial annotation load is very much possible even with a model-based selection if the initial training set is sparse. A judicious choice of the patch size (which is problem-dependent) is also required to minimize the load on the crowd.

We propose a novel, crowdsourcing based solution to address the need for large amount of data for DL-based computer aided detection (CAD) systems. We consider crowdsourcing as an independent (of

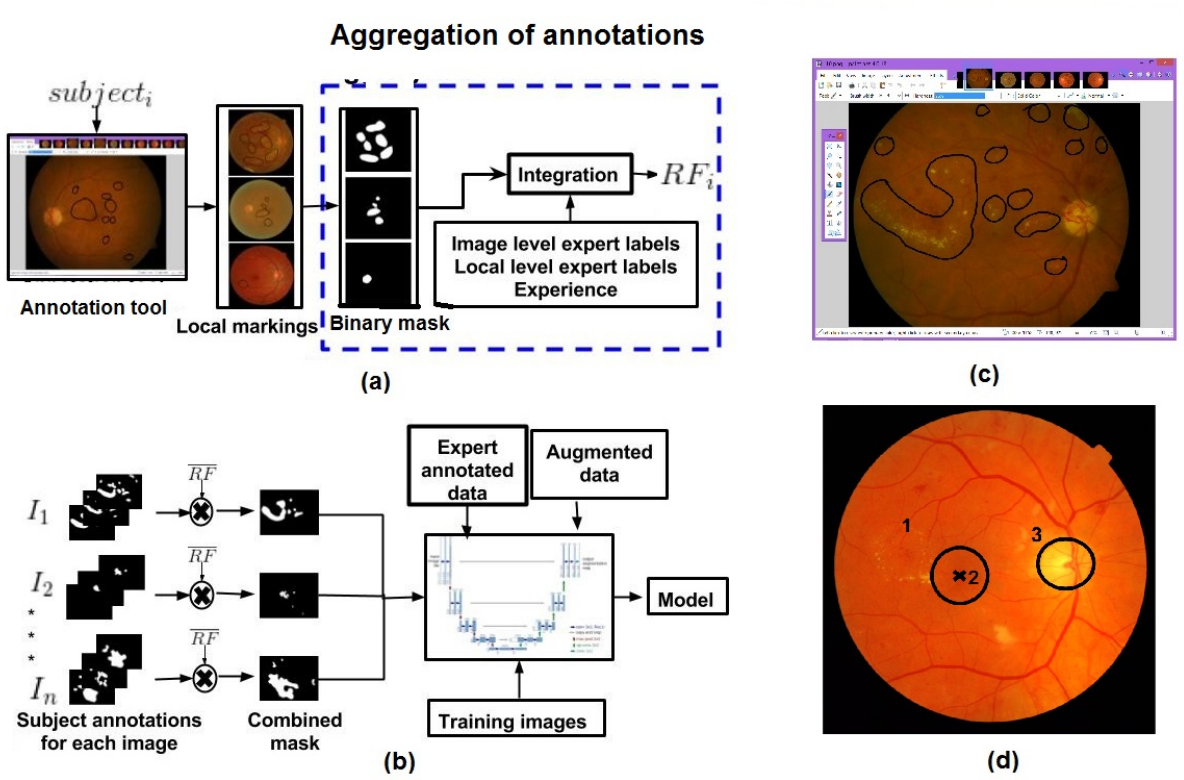


Figure 3.1: Scheme for (a) Reliability Factor (RF) computation for each subject (b) Aggregation of annotations using RF and training U-Net with heterogeneous mixture of annotations; (c) Screenshot of annotation tool. Lesions area marked with black boundary by a subject (d) Fundus image with labeled regions: 1 and 2 are zones of interest centered on macula and 3 is the optic disc.

model learning) activity and propose a scheme wherein only regions of interest (ROI) are marked by the crowd to reduce the burden. A solution for merging crowd annotations is proposed based on assigning a Reliability factor (RF) for each subject of the crowd. This leverages abundant availability of image-level annotations to assess the subjects. Finally, we show how a heterogeneous mixture of annotations derived from experts and crowd, can be used to train a deep neural network (DNN). The CAD problem taken up to showcase our solution is that of hard exudate (HE) detection and localization from color fundus images. Though this is important in diabetic retinopathy staging, very few images are publicly available with local expert annotations. HE appear as small yellowish blobs in isolation or clusters in images. Our results demonstrate that using crowdsourced data as another form of data augmentation, leads to an improvement in detection performance by 11-25%.

3.2 Methods

We begin with a description of the method adopted for collecting crowd annotations and present a scheme for merging these annotations with increased reliability. We then demonstrate a training regime for a DNN using heterogeneous mixture of annotations as shown in Fig. 3.1 (a,b).

3.2.1 Collection of crowd Annotation

The subjects of the crowd are given a free hand annotation tool (Paint.Net [1]) for the task. Fig. 3.1(c) shows a screenshot of the annotation tool. Every member is asked to first determine whether the given image is normal/abnormal and if abnormal, mark the ROI containing HE. In our work, the crowd had 11 engineering students, 4 of whom were familiar with fundus images (L_k) and others who did not have any knowledge of medical images (L_{nk}). 100 images were given to each subject and for each image: user ID, ROI and time taken to complete the task were recorded. Out of the 100 images taken, 6 images were from DIARETDB1 [25] which provides ROI markings from 4 experts; 94 images were from MESSIDOR [12] which provides annotations at the image-level. Of the 94 images, 70 had HE and 24 were normal. HE and the relevant landmarks are shown on a sample image Fig. 3.1(d). The crowdsourced annotations for a sample image are shown in Fig. 3.2.

3.2.2 Aggregating and Improving quality of Crowd Annotations

The aim is to assign a reliability factor (RF) to every subject i. Ideally, the RF should rely on 3 factors: experience of the subject, their performance at image level and local level. The former can be obtained with explicit queries. The assessment of the latter two has to be done by observation and preferably using experts as benchmark. We propose a strategy which rewards a subject for good performance at both local ROI level (based on performance on the 6 images whose local markings are

[1] http://www.getpaint.net/download.html

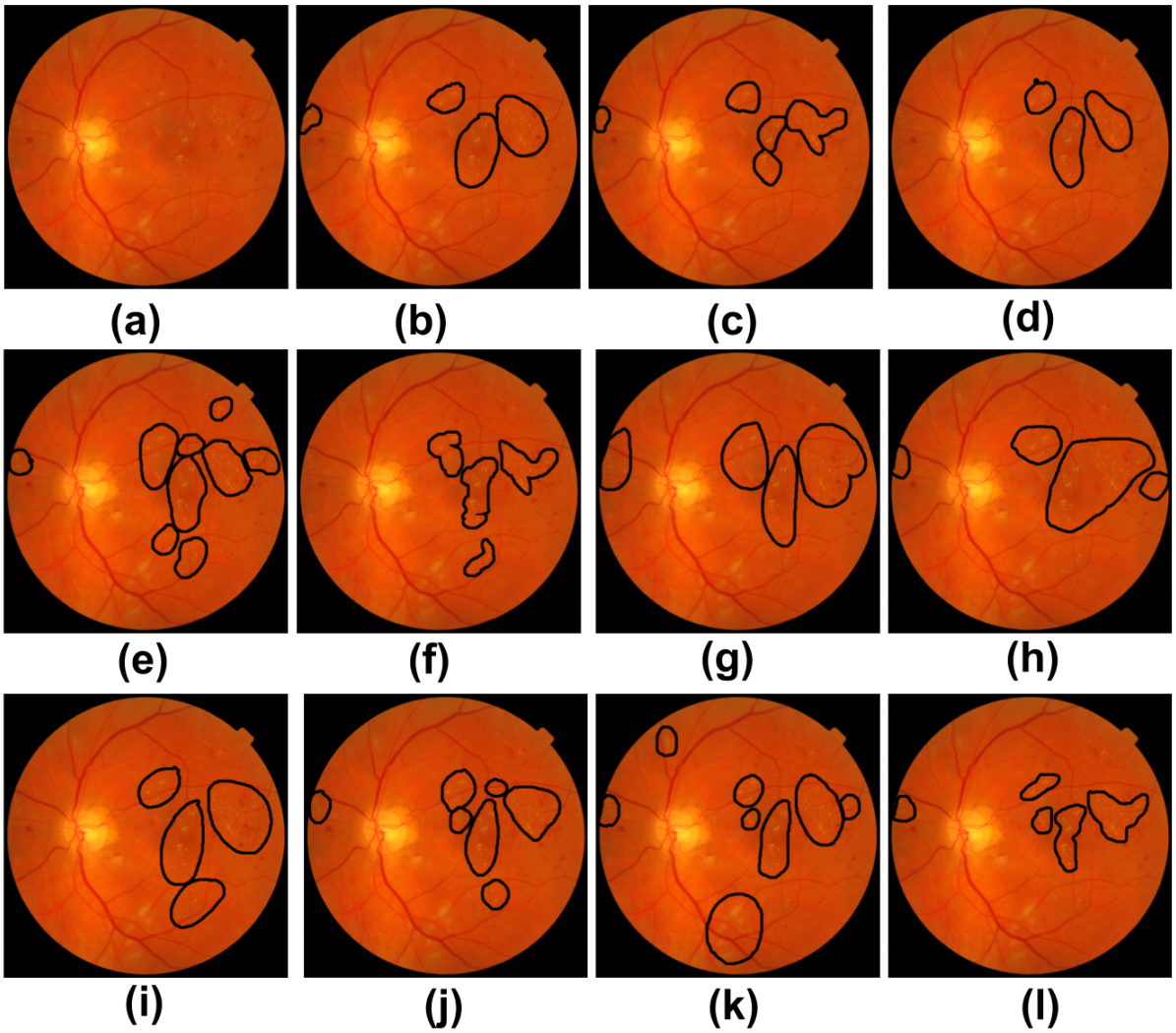


Figure 3.2: Sample image (a) from MESSIDOR dataset followed by region of interest (ROI) markings collected from 11 subjects (b-l) for the same image.

known a priori) and image-level (annotations being available for all the images). A score is given for each of these factors and the final RF is computed as a weighted sum of these scores. The reliability factor RF for the i^{th} subject is defined as :

$$RF(i) = \beta_1 S_1(i) + \beta_2 S_2(i) + \beta_3 S_3(i) \tag{3.1}$$

where $S_j \, \epsilon \, [0, 2]$ are scores determined based on the factors mentioned above and described in detail next; $\beta_i \, \epsilon \, [0, 1]$ are the weights. It is possible to use EM type of techniques to find the optimal weights. In our experiments, weights are explicitly chosen to be 0/1 to evaluate the impact of individual factors on RF.

Performance at image-level: The annotation obtained from the crowd at an image-level is binary. The expert annotation for MESSIDOR is a zone-based label based on the location of HE (standard grading [12]): 0 indicating a normal image, 1 if the lesions are outside a circular region (of diameter equal to optic disc) surrounding the macula and 2 if they are inside this circular region. Hence, we assign a score to a subject not only based on correct labeling of normal images but rewarding them when their ROI is in the correct zone. The score is based on the true poitive rate (TPR) and false poitive rate (FPR) (Eq. 3.7) for each subject which are obtained by comparing the ROI location given by a subject (i) with the zonal labels (j) from MESSIDOR. Specifically, the score for each subject is calculated as follows:

$$S_1(i) = \frac{\sum_{j=0}^{2}(TPR_j(i) - FPR_j(i) + 1)}{3} \tag{3.2}$$

Performance at local level: The local level performance is assessed and a score S_2 is assigned using the 6 images from DIARETDB1. Once again this is based on the TPR/FPR calculated by comparing the ROI marked by a subject with that of (consensus among 3) experts as follows:

$$S_2(i) = TPR(i) - FPR(i) + 1 \tag{3.3}$$

Experience level: This data is gathered with an explicit query on subject's familiarity with medical images in general and fundus image in particular. A score of 2 is assigned to subjects familiar with fundus images and the rest are assigned 1.

Merged output : The merged output annotation of the crowd is a heat map (H) obtained as a weighted (by RF) sum of individual subject annotations for each image j:

$$H_j = \sum_{i=1}^{11} RF(i) I_{ji} \tag{3.4}$$

Here, I_{ji} is the annotated mask for the j^{th} image by the i^{th} subject. On the off chance that none of the data is accessible, regular strategy of majority voting can be used to aggregate the labels, where the heat map is calculated as:

$$H_j = \sum_{i=1}^{11} I_{ji} \tag{3.5}$$

The obtained heat map is finally binarised by thresholding.

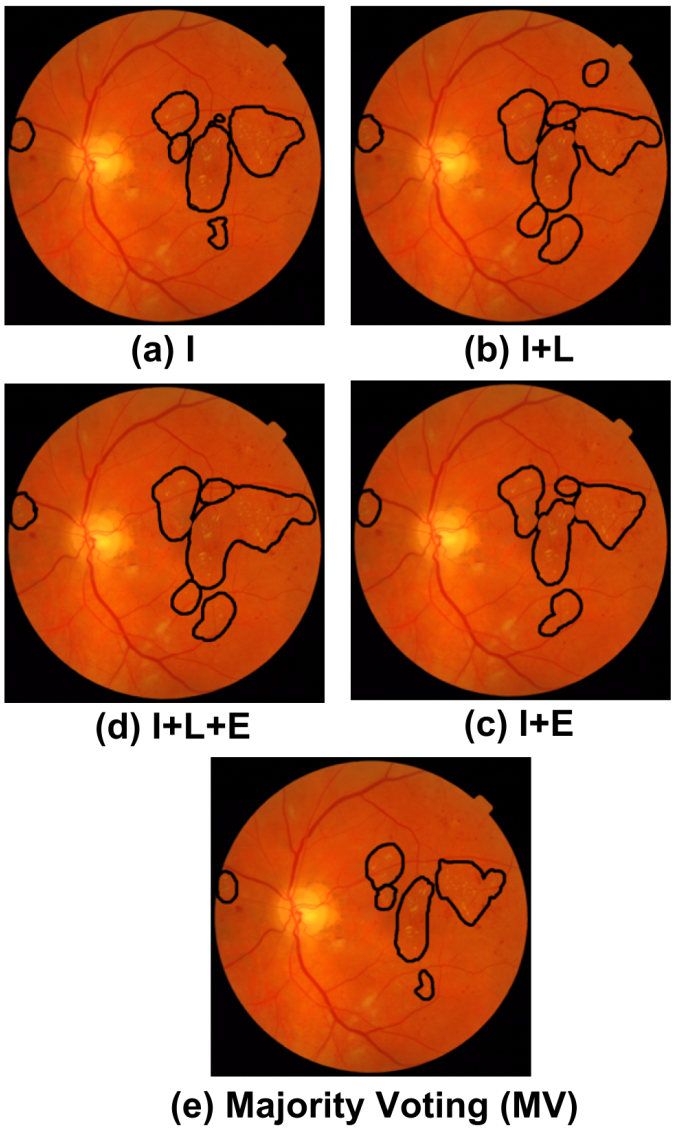


Figure 3.3: The result of aggregation of the subject annotations considering different factors: I - Image level performance, L - Local level performance and E - Experience of the subject. Majority Voting is taken as baseline when none of the above information is available

3.2.3 DNN for aggregation of crowd annotations

We propose an alternate strategy to aggregate crowd annotations using DNN to train different models with different crowd annotations as ground truth. The performance of the subject is assessed based on the model performance on images which have local markings from the expert.

Table 3.1: Assessment of the scheme for Label Aggregation

	TPR_0	FPR_0	TPR_1	FPR_1	TPR_2	FPR_2	**Accuracy**
I ($\beta_2 = 0, \beta_3 = 0$)	100	1.7	87.9	3.3	90.9	6.6	86.2
I + L ($\beta_3 = 0$)	100	15.3	100	16.6	93.9	0	97.8
I + E ($\beta_2 = 0$)	100	7.57	97	0	87.9	0	90
I + L + E	100	6	100	0	87.9	0	91.8
MV ($RF(i) = 1 \forall i$)	89.3	3.5	78.8	5.2	91	13.5	75.7

*I and L denote image and local level performance and E denotes experience of subjects. MV denotes majority voting. All values are in %

In this approach, we chose U-net to train the models. Let C_i be a subject and I_{ij}be the local annotation given by the subject i on image j. Here, $i\epsilon \{1, 2, \cdots, 11\}$ as there are 11 subjects and $j\epsilon \{1, 2, \cdots, 70\}$ as 70 abnormal images from MESSIDOR are considered for training. Each U-net (U_i) is trained to detect hard exudates using the above 70 images for training and the corresponding crowd annotations I_{ij} as ground truth. As there are 11 subjects we obtain a total of 11 U-net models. Now, each of the U-net model U_i is tested on DMED and DRiDB images to obtain pixel wise classification. The SN and PPV values are calculated (Eq. 3.9) for each model by comparing against the local ground truth marked by experts. The RF for each subject is given based on these values as:

$$RF(i) = \frac{SN(i) + PPV(i)}{2} \tag{3.6}$$

3.2.4 DNN for hard exudate detection

We chose the U-Net [42] to demonstrate the proposed solution for crowdsourcing based training. The architecture is modified in terms of the number of filters at each convolutional layer. The number of filters at each stage are reduced to half as there is less variability in lesions to be learnt. Binary cross entropy is used as the loss function. *Preprocessing:* Fundus images suffer from non-uniform illumination due to image acquisitions, camera limitations etc. This is corrected using luminosity and contrast normalization [22]. The optic disc region in every image is masked out and inpainted. Fundus extension is applied to remove the black mask region and all images are normalized to have zero mean and unit variance.

Data Augmentation: Data augmentation is done by applying random transformations to the images. This included random rotation between -25^o to 25^o, random translation in vertical / horizontal directions in the range of 50 pixels, and random horizontal / vertical flips. For fairness, the number of images used for data augmentation are chosen to be the same as that of crowdsourced images.

3.3 Implementation and evaluation details

3.3.1 Datasets

:Four public datasets, (DRIDB [38], DMED [15], MESSIDOR and DIARETDB1) were considered for the evaluation of DNN for HE detection. DMED (1 expert) has pixel level annotations whereas DIARETDB1 (4 experts) and DRiDB (1 expert) have ROI markings. We considered the consensus marking of 3 experts to derive a binary mask in case of DIARETDB1. The obtained binary mask was overlapped on the image and thresholded to get pixel level lesion mask. The MESSIDOR dataset was used for crowdsourcing and has only image-level labels for HE. Images from all the datasets were cropped and re-sized to 256$\times$256 before feeding to the DNN.

3.3.2 DNN for aggregation of crowd annotations

The training of each U-net consisted of 70 abnormal images from MESSIDOR given to the crowd for annotation. After augmentation it accounts to a total of 140 images for training with the corresponding crowd annotations as ground truth. The testing consists of 84 abnormal images, 31 from DRiDB and 53 from DMED.

3.3.3 DNN for HE detection

Since the problem of interest is HE detection, only pathological images with HE (considered abnormal) were included for all training and testing. A total of 154 images were collected for training: DriDB and DMED had a total of 31 and 53 abnormal images with expert annotations; 94 images were randomly chosen from MESSIDOR such that 70 were abnormal with crowd annotations. Including data augmentation, the total number of training images count to 308. DIARETDB1 had 48 abnormal images when consensus of 3 expert marking is taken and out of these 42 were considered for testing since 6 were given to the crowd for local annotation. The testing set size is limited only by the paucity of images with local markings available for public access.

3.3.4 Implementation details

The UNET model was implemented in python using Keras with Theano as backend and trained on a NVIDIA GTX 970 GPU, 4GB RAM. Training was done with random initialized weights for 2000 epochs by minimizing the loss function using Adam optimizer. For model parameters learning rate was initialized to $0.5X10^{-5}$, batch size is 4 and others were left at default values. Class weights were defined as inverse ratio of the number of positive samples to negative samples and modified empirically.

Table 3.2: HE detection performance with different training regimes.

Trained data (total number of images)	SN(%)	PPV(%)	AUC
Expert (84)	90	60.3	0.750
Expert + Augmentation (154)	89.8	61.6	0.765
Expert + Crowd (I+L) (154)	90.1	71.5	0.869
Expert + Crowd (MV) + Augmentation (308)	90	84.6	83.9
Expert + Crowd (I) + Augmentation (308)	90	85	0.879
Expert + Crowd (DNN) + Augmentation (308)	90.7	85.1	0.891
Expert + Crowd (I+L) + Augmentation (308)	90.1	90.4	0.932

3.3.5 Evaluation metrics

Assessment of the crowdsourced annotations was done with TPR, FPR and accuracy as evaluation metrics. As the image-level labels available from the experts is for 3 classes (labeled i: 0, 1 and 2), TPR, FPR and accuracy were calculated as follows:

$$\begin{aligned} TPR_i &= \frac{N_{ii}}{\sum_{j=0}^{2} N_{ij}} \\ FPR_i &= \frac{\sum_{j=0,j\neq i}^{2} Nij}{\sum_{j=0,j\neq i}^{2} Nij + \sum_{k=0,\neq i}^{2}\sum_{j=0,\neq i}^{2} N_{jk}} \end{aligned} \tag{3.7}$$

$$Accuracy = \frac{\sum_{i=0}^{2} N_{ii}}{\sum_{i=0}^{2}\sum_{j=0}^{2} N_{ij}} \tag{3.8}$$

Here N_{mn} denotes the number of images with disagreement, the crowd label is m and the expert label is n.

The HE detection performance was evaluated using Sensitivity (SN), Positive Predictive Value (PPV) which are defined as:

$$SN = \frac{TP}{TP + FN}, PPV = \frac{TP}{TP + FP} \tag{3.9}$$

The pixel wise detection by U-net was converted to region wise by apply connected component analysis to evaluate against the expert local annotations. Each detected region in an image is deemed to be true positive (TP) if it overlaps with at least 50% (but not exceeding more than 150%) of the area manually marked by experts; else it is a false positive (FP). False negative (FN) is a region marked by expert that is undetected by the model. Area Under Curve (AUC) of SN vs PPV plot is also used as a measure of performance.

3.4 Experiments and Results

3.4.1 Crowdsourced data

The average time taken by subjects to mark ROI for 100 images was around 90 minutes. The task was conducted in two sessions of 50 images each. Hence, a total of 1100 markings were obtained in a span of two days. The annotation performance is presented as a box plot in Fig. 3.4 for the 3 classes or zonal labels. The mean performance accuracy is 70%. The obtained class-wise performance of TPR/FPR of 89.6%/6.9% for Normal/class0, 80.7%/11.29% for class1 and 77.69%/10.7% for class2. These indicate that the crowd is good at correctly identifying normal images and detects HE in zone 1 (very large) more accurately than zone 2 (size of Optic disc) suggesting a bias towards the larger zone. Since lesions in zone 2 require immediate referral, urging subjects to scrutinize this zone may be advisable.

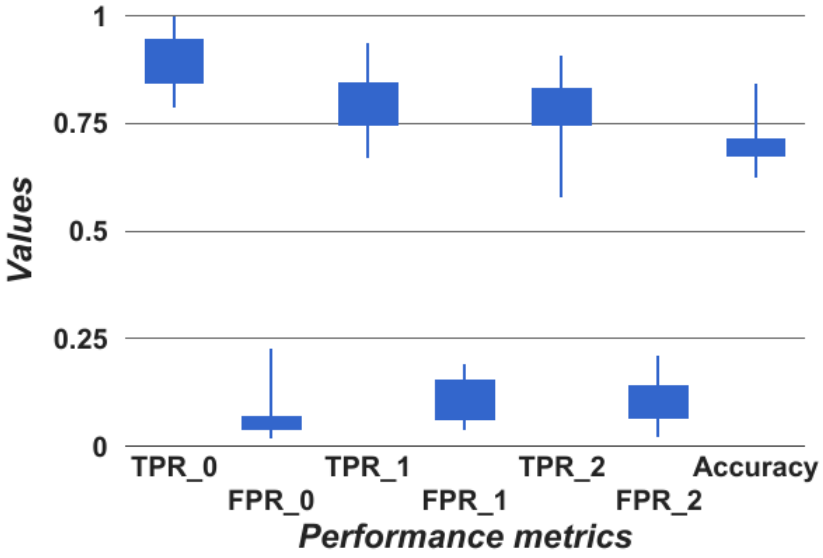


Figure 3.4: Box plot of crowd annotation performance metrics

3.4.2 Aggregation of labels

The impact of the terms in Eq.3.1 is studied by setting β_i-0/1. The obtained TPR and FPR are listed in Table3.1. With the baseline as majority voting, considering only image-level performance for RF, results in a 10% improvement in accuracy while addition of local performance boosts this to 22%. This is noteworthy as local performance is known only for 6% of the images given to crowd. Experience does not seem to be beneficial for this experiment as accuracy suffers when performance *and* experience are considered. This may be due to the fact that crowd is made of students and hence experience is really not meaningful.

3.4.3 DNN for hard exudate detection

Training a DNN with small set of expert annotated data and augmenting it with the standard approaches as well as crowdsourced data was studied as follows. Various models were trained using: i) only expert (E), ii) expert and augmented data (iii) expert and crowdsourced annotations (C) merged using I+L, iv) E, C (DNN), augmented data derived from E+C and finally v) E,C (I+L) and augmented data derived from E+C. Sample results of HE detection are shown in Fig. 3.5 for v.

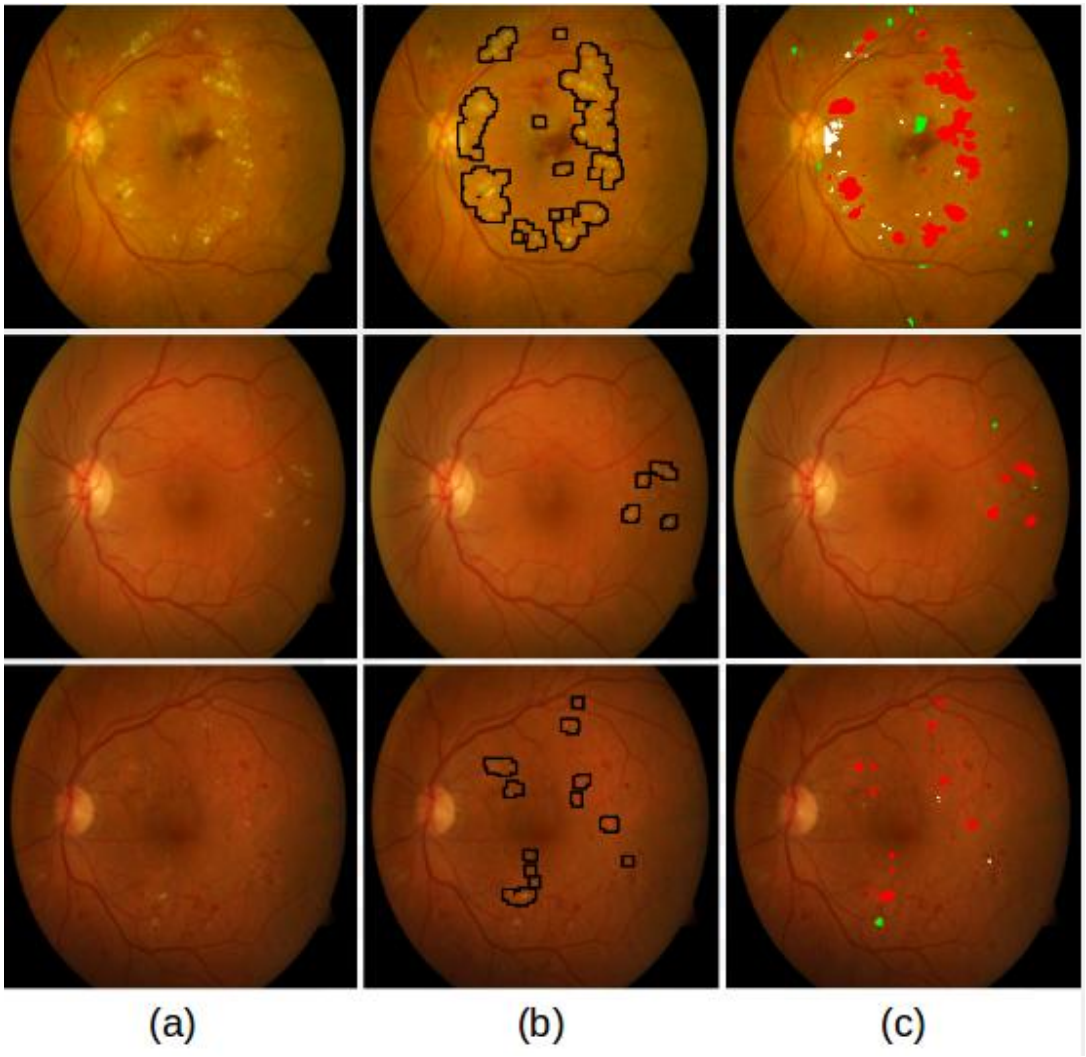


Figure 3.5: (a) Sample images (b) ground truth marked by experts (c) DNN output for HE detection. Color coding : true positive (TP) - red, false positive (FP) - green and false negative (FN) - white (compared to local expert annotations)

The assessment is based on SN, PPV and the AUC values which are reported in Table. 3.2. The change in PPV values are shown in the table by fixing SN value at approximately 90%. The model

over-fits on data trained only on expert annotations within few epochs. Data augmentation improves the AUC and PPV by about 2%, whereas, crowdsourcing improves them by over 11%. Finally, when the annotations (expert and crowd) are augmented and added to the training set the improvement in AUC and PPV are a healthy 24.5% and 50%, respectively. Setting PPV to 70% results in SN values ranging from 70% to 96%; which is a similar level of improvement (Fig. 3.6) as that of PPV. The proposed training strategy is thus very effective in improving the detection performance. The sensitivity versus PPV plots are shown for the different trained models in Fig. 3.6.

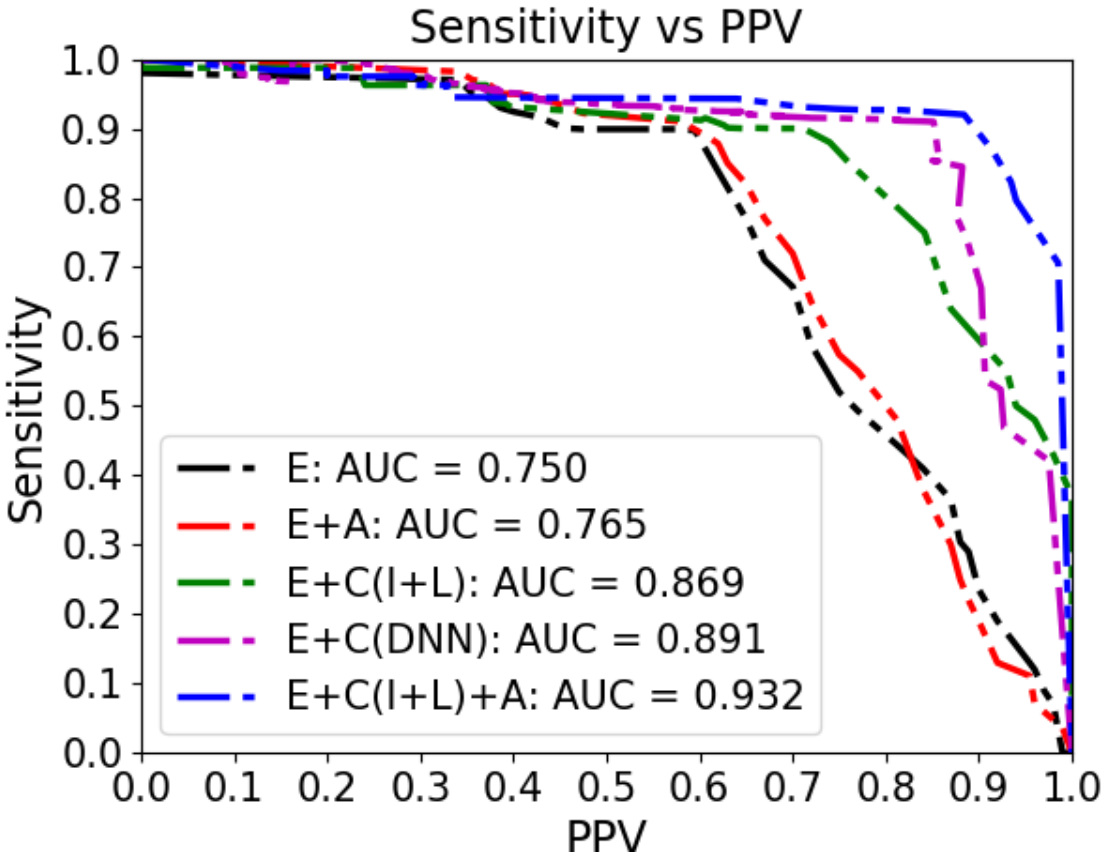


Figure 3.6: Performance of Deep Neural Net for hard exudate detection

The final model trained on 308 abnormal images was also tested on 40 normal images from DIARETDB1. No abnormalities were detected in 35 images, while the average FP per image for 5 images was 2.6. Comparison with the existing approaches for HE detection is difficult as the validation datasets and the number of images vary. Nevertheless, for completeness, we report them next. An unsupervised approach [43] reports a SN of 90.2% and PPV of 96.8% based on ROI detection while [39] [49] report pixel based classification with SN ranging from 70-78% and PPV ranging from 75-78%.

3.5 Concluding Remarks

Crowdsourcing is an alternative source of annotation, but can be effective only with introduction of measures to improve the reliability of annotations. The proposed RF concept allows good and experienced annotators from the crowd to have higher weights in the final, weighted-sum based merging of annotations. The results show that including a small (6% of total set to be annotated) set of images with expert annotations and using commonly available image level annotations can improve the reliability of crowd annotation. This improvement enables the crowd annotation to be considered on par with that of experts for training a DNN-based CAD system. Training with a heterogeneous set of data (expert, crowd) together with data augmentation have significant impact on the detection performance (in terms of AUC) of a CAD by at least 25%. Hence, crowdsourcing, after steps taken to improve its reliability, can be an alternative form of data augmentation. There are some limitations to our study. It is limited to only hard exudate detection, further experimentation can be done to train and evaluate on other abnormalities. The image level annotation that is used in our case study has a coarse spatial encoding whereas there are scenarios where images are labeled only as normal/abnormal. In such a scenario, the set used to assess local level performance of subjects may have to be enlarged. This can effectively reduce the size of crowd annotations that can be obtained if annotation load is held constant.

Chapter 4

Retinal image synthesis for CAD development

4.1 Introduction

Generation of synthetic medical data is aimed at addressing a range of needs. Early examples are generating digital brain phantoms [10] and synthesizing a whole retinal image [33] using complex modeling. These were aimed at aiding the development of algorithms for denoising, reconstruction and segmentation. Recently simulation of brain tumors in MR images [37] has also been explored to aid CAD algorithm development. With the advent of deep learning, modeling of complex structures and synthesizing images has become easier with a class of neural networks called generative adversarial networks or GAN [17].

GAN is an architecture composed of two networks, namely, a generator and a discriminator. Functionally, the generator synthesizes images from noise while the discriminator differentiates between real and synthetic images. GAN have recently been explored for a variety of applications: detection of brain lesions [40], predicting CT from MRI images [36], synthesizing normal retinal images from vessel mask [11], segmenting anatomical structures such as vessels [18] and optic disc/cup [44].

We propose a GAN for generating images with pathologies in a *controlled manner* and illustrate how the generated synthetic images can be used to address the data sparsity problem which hampers the development of robust CAD solutions for abnormality detection. We choose *staging* of diabetic retinopathy (DR) from given color retinal images as a case study. The ETDRS standard for staging of DR is based on the number and location of hard exudates (HE)/ haemorrhages (HM) [50]. However, very few images are publicly available with local markings of HE/ HM. Recent deep learning-based methods [46], [51] overcome this problem by sampling a large public dataset (with only image-level annotations) to get local annotations for a much smaller subset of images which are abnormal. These annotations are privately held and hence such measures are not beneficial to a wide community for building a robust CAD solution.

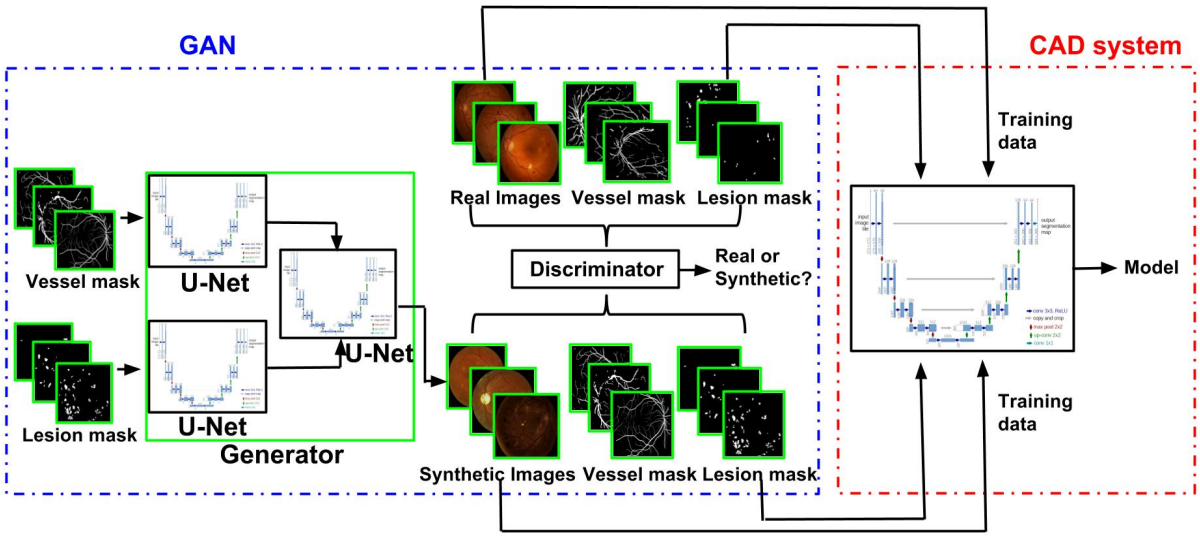


Figure 4.1: Proposed end-to-end pipeline for generation of abnormal retinal images and developing a CAD system for detection of haemorrhages.

4.2 Method

The proposed method consists of three modules: (i) pre-processing, (ii) synthetic image generation (with HE and HM) and (iii) CAD for HE and HM detection. Two seperate GAN and CAD models are trained for generating and detecting HE and HM. Below we describe the common GAN and CADH architecture. As a part of the pre-processing step, given retinal images are corrected for non-uniform illumination using luminosity and contrast normalization [23].

4.2.1 GAN for Synthesis of Retinal Images with Pathologies

Generating *normal* retinal images from vessel mask has been attempted earlier [11] with a single U-net for the generator and a 5-layer convolutional neural network for the discriminator. Our interest is in generating images with hard exudates (HE)/ haemmorages (HM) towards synthesis of exemplars for different stages of DR. HM are often indistinguishable from vessel fragments and therefore the input to the generator has to enable distinguishing between these both structures. Further, exemplar generation requires gaining control of the locations, size and density of HM. Hence, we propose a GAN architecture (shown in Fig.4.1) with a generator consisting of two parallel networks: one with a vessel mask as input and another with a lesion mask as input. The output of the networks, based on the U-net architectures, are merged and fed to a third U-net architecture which generates the whole retinal image with lesions. The generator thus maps from vessel (v_i) and lesion (l_i) masks to a retinal image (r_i) using a mapping function. A 5-layer convolutional neural network as in [11] is used for the discriminator to distinguish between the real and synthetic sets of images, with each set consisting of vessel and lesion masks along with retinal images.

The GAN learns a model as follows: there are two network architectures in GAN, one is the generator and the other is the discriminator. The generator maps a latent space to the desired data distribution (unsupervised) while the disciminator learn in a supervised manner. The discriminator iteratively reduces its misclassification error by more accurately classifying the real and synthetic images while the generator aims to deceive the discriminator by producing more realistic images. This is known as the zero-sum game. The overall loss function that is to be optimized is chosen as a weighted combination of 3 loss functions: L_{adv}, L_{SSIM} and L_1 as defined below in eq.4.1-4.4 to produce sharp and realistic images. (here, G and D, represent generator and discriminator respectively). (i) The adversarial loss function L_{adv} is defined as

$$\begin{aligned} L_{adv}(G,D) = \mathbb{E}_{(v,l),r\sim p_{data}((v,l),r)}[log(D((v,l),r))] \\ + \mathbb{E}_{v,l\sim p_{data}(v,l)}[log(1-D((v,l),G(v,l)))] \end{aligned} \tag{4.1}$$

where $\mathbb{E}_{(v,l),r\sim p_{data}}$ represents the expectation of the log-likelihood of the pair $((v,l),r)$ being sampled from the underlying probability distribution of real pairs $p_{data}((v,l),r)$, while $p_{data}(v,l)$ is the distribution of real vessel and lesion masks.

(ii) The Structure Similarity (SSIM) [48] index is useful in quantitatively measuring the structural similarity between two images $(r, G(v,l))$. It also has been shown to perform well for reconstruction and generation of visually pleasing images.

$$SSIM(p) = \frac{2\mu_r\mu_{G(v,l)} + C_1}{\mu_r^2 + \mu_{G(v,l)}^2 + C_1} \cdot \frac{2\sigma_{rG(v,l)} + C_2}{\sigma_r^2 + \sigma_{G(v,l)}^2 + C_2} \tag{4.2}$$

where (μ_r, $\mu_{G(v,l)}$) and (σ_r,$\sigma_{G(v,l)}$) are the means and standard deviation computed over patch centered on pixel p, C_1 and C_2 are constants. The loss L_{SSIM} can be computed as:

$$L_{SSIM} = 1 - \frac{1}{N}\sum_{p\epsilon P} SSIM(\tilde{p}) \tag{4.3}$$

where $\tilde{p}$ is the center pixel of a patch P in the image I.

(iii) The loss function L_1 is used mainly to reduce artifacts and blurring and is defined as

$$L_1 = \mathbb{E}_{(v,l),r\sim p_{data}((v,l),r)}(\|r - G(v,l)\|_1) \tag{4.4}$$

The overall loss function to be minimized is taken to be

$$L(G,D) = L_{adv} + \lambda_1 L_1 + \lambda_2 L_{SSIM} \tag{4.5}$$

where λ_1 and λ_2 control the contribution of the L_1 and L_{SSIM} loss functions respectively.

4.2.2 CAD for HE/ HM Detection

We chose the U-Net [41] to build a CAD solution for detection of HE (referred to as CADH). This is used to demonstrate that the synthetic images (generated by our proposed GAN) are a reliable resource

in training the U-net. The U-net architecture consists of a contracting and an expansive path. The contracting path is similar to a typical CNN architecture, whereas in the expanding path, max-pooling is replaced by up-sampling. There are skip connections between contracting and expanding paths to ensure localization. The U-net is modified in terms of the number of filters at each convolutional layer and the loss function. The number of filters at each stage is reduced to half to simplify computations. The loss is modified to account for the misclassification of lesions. The U-net architecture provides the segmentation of HE. The segmented HE are counted and the image is classified into the respective grade accordingly (as given in section 3.1: training data for CADH).

4.3 Implementation and evaluation details

4.3.1 Datasets

Both GAN and CADH were trained on *pathological* images. These are drawn from DRiDB [38] (31-HE, 31-HM) and a locally sourced dataset denoted as LoD (53-HE, 58-HM). Testing of CADH was done at i) lesion level on 42/ 40 pathological images for HE/ HM from DIARETDB1 [25] and **ii) at a stage-level on 308 abnormal images + 892 normal images (without HE/ HM) from MESSIDOR [12].**

Lesion markings are available for DIARETDB1 from four experts, while for DRIDB and LoD it is from one expert. The consensus of 3 experts was considered to derive a binary mask for DIARETDB1. The ground truth of all the three datasets were overlapped with the respective images and thresholded to get a pixel-level lesion mask. The vessel masks, whenever unavailable were derived using method in 4.3.1.1. Images from all datasets were cropped and resized 512x512 before feeding them to GAN or CADH.

4.3.1.1 Training Data for GAN

Training of the GAN requires both lesion and vessel masks. The lesion masks for the training data are available from experts, but vessels masks are available only for DRiDB. It is tedious and time consuming task to mark the vessels in each of the retinal images. Hence, vessel masks were derived using method [32], which has proved to perform relatively well for vessel segmentation even in the presence of pathologies.

4.3.1.2 Training Data for CADH

For training the CADH, a heterogeneous mixture of data were combined, namely, expert annotated data, synthetic data and augmented data. The DRiDB and LoD datasets were sources of expert annotated data. Augmented data was derived by applying random transformations to the images. This

included random rotation between -25^0 to 25^0, random translation in vertical / horizontal directions in the range of 50 pixels, and random horizontal / vertical flips. Finally, the synthetic retinal images were generated using GAN as follows. The vessel and lesion masks were taken randomly from LoD and DRiDB. The lesion masks were modified using the same random transformations such as flipping the lesions sector wise, flipping horizontally and vertically, rotations and translations. Retinal images containing HE are graded using zone-based label, based on the location of HE (standard grading [12]): 0 indicating a normal image, 1 if the lesions are outside a circular region (of diameter equal to optic disc) surrounding the macula and 2 if they are inside this circular region. Images containing HM are graded with severity levels as in [12]: grade 0/1 (no HM), grade 2 (1-5 HM) and grade 3 (more than 5 HM). The lesions masks were derived to provide exemplars for each level using these rules. The number of lesions in each category were maintained by masking out few lesions or adding new lesions from another lesion mask randomly. Fig. 4.2 and Fig. 4.3 show samples of the vessel, lesion masks and generated synthetic images containing HE (zone 1 and 2) and HM (grade 2 and 3) respectively.

4.3.2 Computing Details

The models were implemented in Python using Keras with Theano as backend and trained on a NVIDIA GTX 970 GPU, 4GB RAM. Training was done with random initialized weights for 2000 epochs by minimizing the loss functions described in Section 4.2.1 using Adam optimizer. For model parameters, learning rate was initialized to 2×10^{-4} for GAN and 1×10^{-5} for CADH. A batch size of 4 was considered for both cases and other parameters were left at default values. Class weights were outlined as the inverse ratio of the number of positive samples to negative samples and modified empirically.

4.3.3 Evaluation Metrics

The synthetically generated images were evaluated quantitatively and qualitatively (two sample synthetic images are shown in Fig. 4.4). The mean and standard deviation of the Q_v score described in [28] was computed over all images (42/40 abnormal (HE/HM)) in DIARETDB1.

The performance of CADH was evaluated using Sensitivity (SN) and Positive Predictive Value (PPV) which are defined as follows: $SN = \frac{TP}{TP+FN}$ and $PPV = \frac{TP}{TP+FP}$. To evaluate against the given local annotations by experts, the pixel wise classification was converted to region wise detection by applying connected component analysis and requiring at least 50% (but not exceeding more than 150%) overlap with manually marked regions to identify true positive detections (TP); else it is false positive (FP). If a region is marked by the expert but was not detected by the model then it is a False negative (FN). The area under the SN vs PPV curve (AUC) is also taken as a measure of performance.

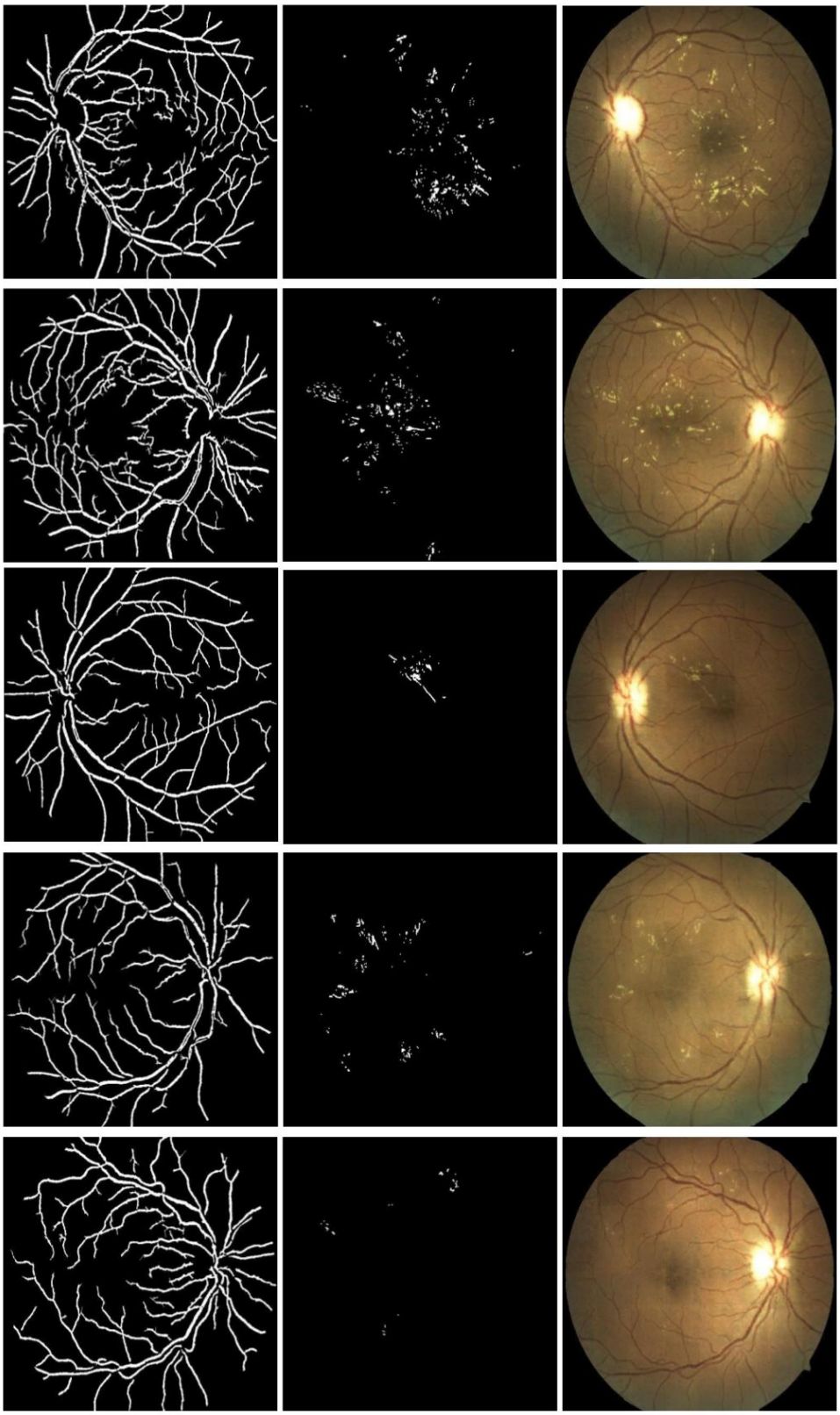


Figure 4.2: From left to right: vessel mask, lesion mask, synthetic image (for HE). From top to bottom: first two sample images fall under zone 1 and the last three images fall under zone 2.

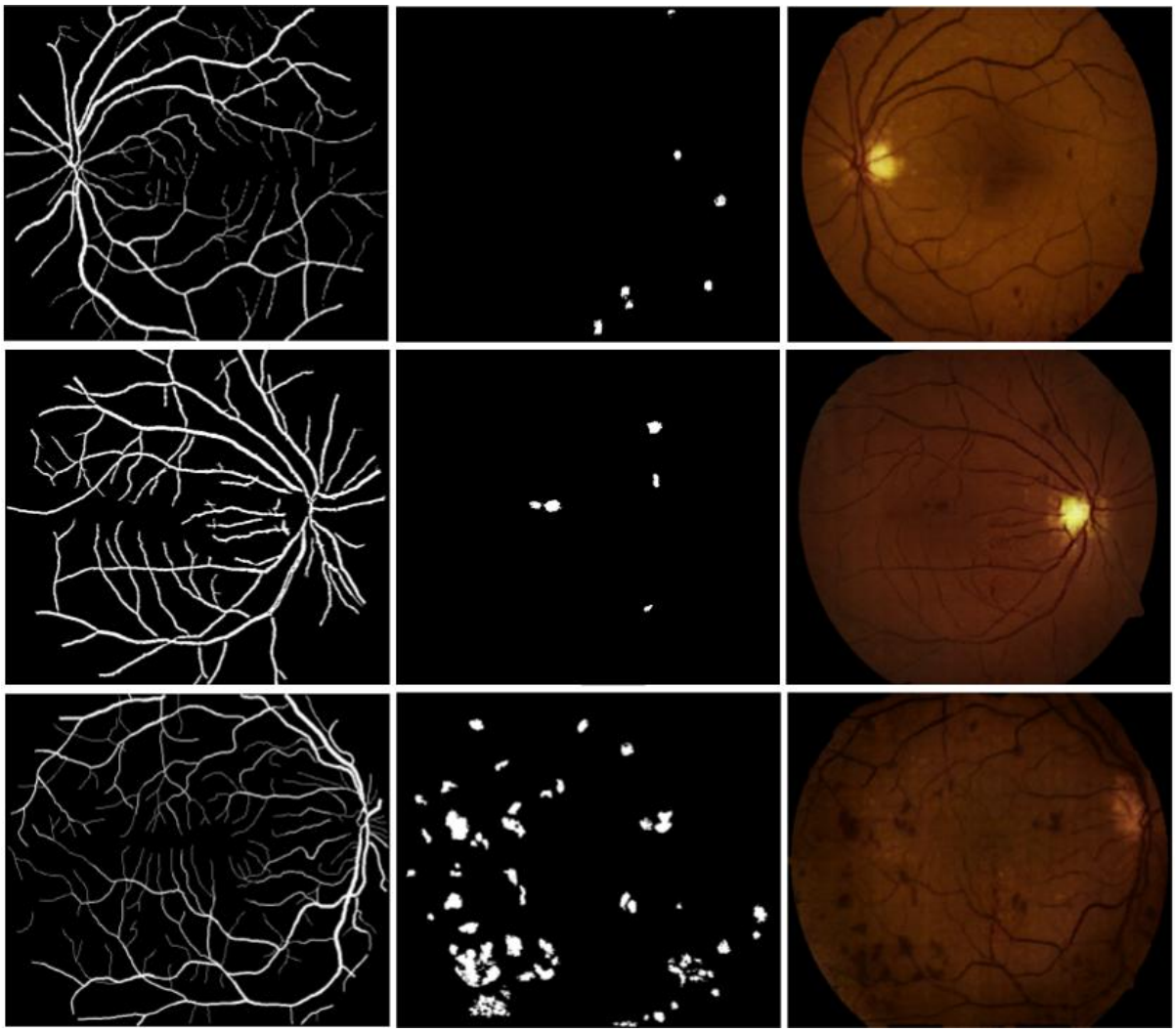


Figure 4.3: From left to right: vessel mask, lesion mask, synthetic image (for HM). From top to bottom: first two sample images fall grade 2 and the last image falls under grade 3.

4.4 Experiments and Results

4.4.1 Synthetic Image Generation (GAN)

Fig.4.4 shows two sample synthetic retinal images generated by the proposed GAN model and the corresponding real image. The first two columns show the vessel and lesion masks given as input to the GAN. Third and fourth columns show the synthetic and the corresponding real images, respectively. The synthetic images appear realistic yet differ from the real images in terms of background color, texture and illumination. Lesion locations are roughly similar but sizes are different as lesion masks are not results of exact segmentations of lesions.

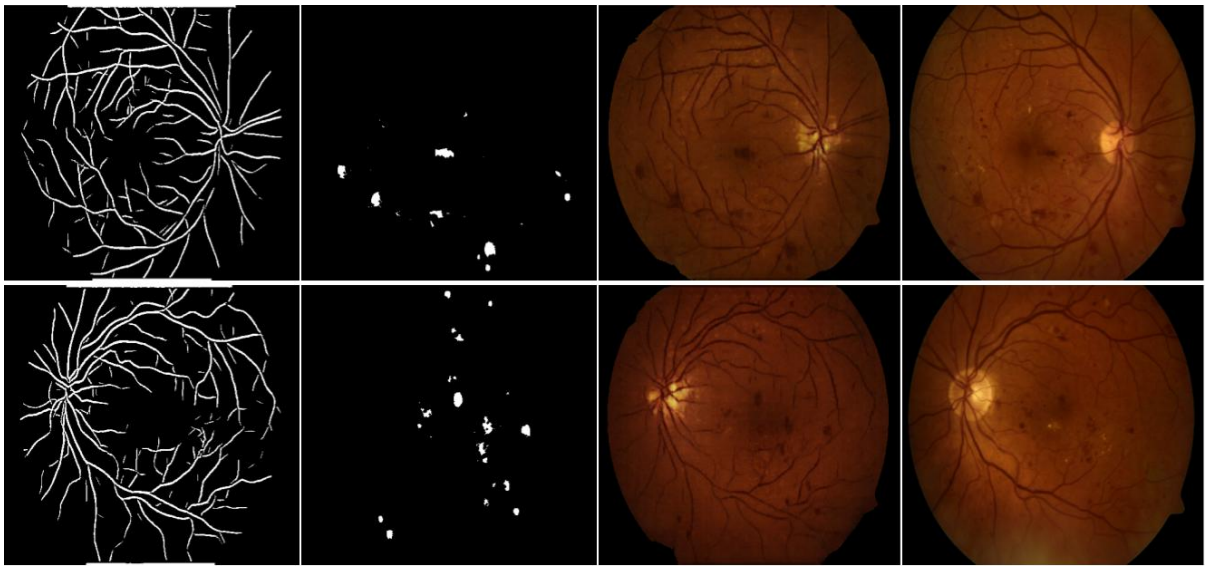


Figure 4.4: Results of GAN-based image synthesis. From left to right: vessel mask, lesion mask, synthetic image and corresponding real image.

The mean/ standard deviation of Q_v computed over all images with HE (HM) in DIARETDB1 is 0.074/0.017 (0.0516/ 0.0144) and over all the synthetic images generated from vessel and lesion mask from DIARETDB1 is 0.082/0.02 (0.0675/0.0239). The Q_v score is higher for images of greater quality, this indicates synthetic images are considered better as they contain less noise.

4.4.2 CAD for HE/ HM Detection (CADH)

The utility of the synthetic data for CAD development was tested by training 4 different CADH models by varying the training set content. Denoting the set of real images with expert annotations as E and the set of synthetic images generated by GAN with the corresponding lesion masks as S, the training set variations considered are: (i) only E, (ii) E with data augmentation (E+A), (iii) E and S, (iv) E, S with data augmentation (E+S+A). The computed SN at a fixed PPV and AUC values for these variants

Table 4.1: HE detection performance with different training regimes.

Trained data (No. of images)	**SN(%)**	**PPV(%)**	**AUC**
Expert (84)	55.8	78	0.75
Expert (84) + Augmentation (70)	60	78	0.765
Expert (84) + Synthetic (70)	70	78.2	0.869
Expert (84) + Synthetic (70) + Augmentation (154)	90	77.5	0.894
Expert (84) + Synthetic (140) + Augmentation (224)	94.7	78	0.94

are reported in Table. 4.1 and in Table. 4.2 for HE and HM respectively . The SN vs PPV curve is shown in Fig. 4.5for HE and in Fig. 4.6 for HM.

The tabulated results indicate that addition of synthetic data (E+S) boosts SN (HE/HM) by (25.4/25.6)% and (16/12.3)% over E and E+A, respectively. The full set of E+S+A yields the best performance with an improvement (over E) in SN by (60/37.7)% and AUC by (15.7/23)%. This establishes the effectiveness of synthetic data in general and in CAD development. Increasing the number of synthetic images improved the performance (row 5). In order to assess if synthetically derived data has artifacts, the E+S+A variant was tested on an exclusive set of separately generated synthetic images using vessel and lesion masks of DIARETDB1. The obtained results (row 6) shows a minor degradation over that for real images (row 4), implying the generated data is free of artifacts.

Table 4.2: HM detection performance with different training regimes.

Trained data (Number of images)	**SN (%)**	**PPV (%)**	**AUC**
Expert (89)	63.1	79.4	0.690
Expert (89) + Augmentation (89)	70.6	79.6	0.742
Expert (89) + Synthetic (71)	79.3	79.6	0.829
Expert (89) + Synthetic (71)+ Augmentation (160)	86.9	79.8	0.851
Expert (89) + Synthetic (141)+ Augmentation (230)	92.7	79.4	0.905
Expert (89)+ Synthetic (71)+ Augmentation (160) *	84.2	80	0.834

* detection performance on only synthetic images.

Benchmarking the detetion performance of HE and HM: Most recent approaches for HE detection report at the image-level (normal or has HE) rather than at a local level. The exception is [39] where a deep learning based approach is reported to have an F_1 score of 0.78 with SN and PPV of 78% each on 50 images from DRIDB dataset. In order to benchmark HM with a recent fast CNN method [46] which aims at HE/no HE classification, the model trained on E+S(140)+A was also tested on the MESSIDOR dataset which provides severity grades for each image. Our solution has a SN/ SP of 92/ 94.4% for

grade 2 and 89.4/ 90.1% for grade 3. [46] reports SN/ SP of 91.9/ 91.4% for a binary (not grade-wise) classification which is less compared to 94/ 91.7% obtained for CADH. This indicates that the model trained with synthetic data (can be generated in abundance) is better than that trained with expert annotated data (which is difficult to obtain).

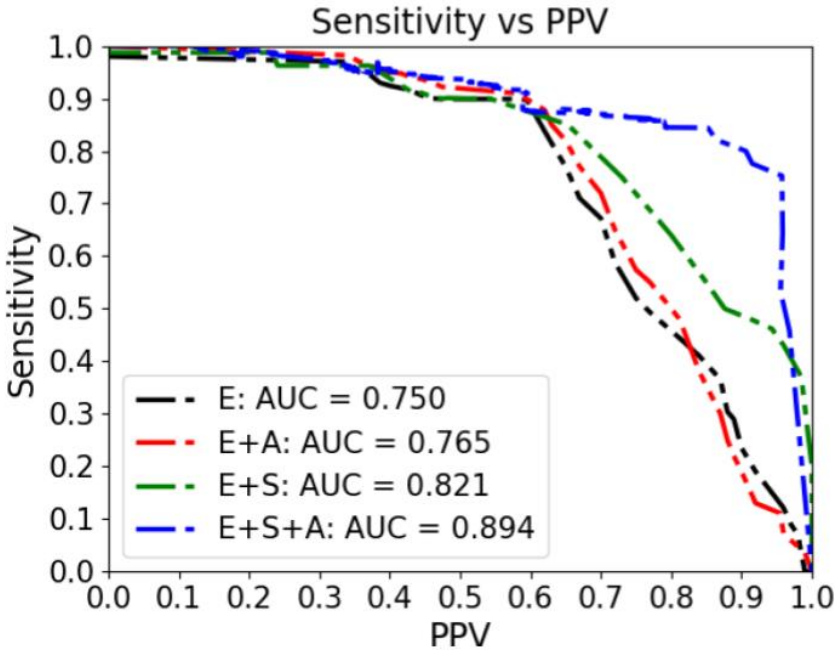


Figure 4.5: SN vs PPV curve for CADH.

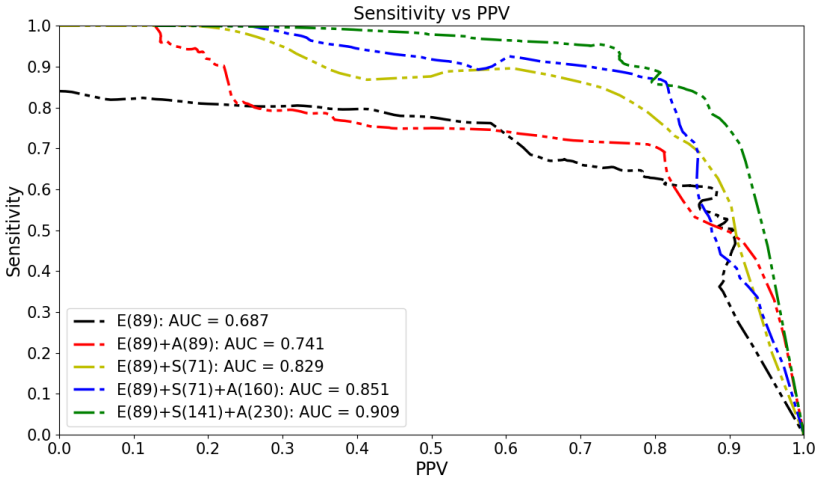


Figure 4.6: SN vs PPV curve for CADH.

4.5 Concluding Remarks

We proposed a novel solution to develop retinal images with HE/ HM using generative adversarial networks. The network is trained to generate the retinal image using vessel and lesion masks. Hence, we can develop retinal image with any type of severity, by providing the corresponding lesion mask. The synthetic abnormal images generated are shown to be realistic in the type of lesions produced and also the color, texture using the Q_v metric. These generated images are valuable in developing a CAD system which detects and localizes HE/ HM as addition of synthetic data had led to significant improvement in both SN and AUC. Our proposed approach can be extended to other image modalities and thus has a wide potential.

Chapter 5

Conclusion and future work

DR is a very complex and severe disease which can lead to permanent blindness. Regular screening and diagnosis is required to avoid permanent damage to the eye. For the diagnosis of large population we need accurate CAD systems. But as we have seen, it is difficult to achieve high accuracy due to the amount of availability of annotated data. In this thesis, we discussed the solution for the data sparsity problem in three different setups.

First we studied how to mitigate this problem by using unlabeled data in the training. Although NV detection is a difficult task as it is characterized by complex texture changes and automation system development is impeded by limited availability of annotated data, the proposed co-training method was seen to overcome these limitations. The proposed co-training semi-supervised framework along with the selection of independent features and ensuring spatial consistency has shown an improvement in AUC value from 0.95 to 0.98 which is significant. Consistent good performance of the method across datasets, at both patch and image levels, depicts its robustness to changes in resolution, illumination, tissue type and noisy conditions.

We have also seen crowdsourcing as an alternative source of annotation, but can be effective only with introduction of measures to improve the reliability of annotations. The proposed RF concept allows good and experienced annotators from the crowd to have higher weights in the final, weighted-sum based merging of annotations. The results show that including a small (6% of total set to be annotated) set of images with expert annotations and using commonly available image level annotations can improve the reliability of crowd annotation. This improvement enables the crowd annotation to be considered on par with that of experts for training a DNN-based CAD system. Training with a heterogeneous set of data (expert, crowd) together with data augmentation have significant impact on the detection performance (in terms of AUC) of a CAD by at least 25%.

Finally we have generated retinal images and their respective annotations using generative adversarial networks. The network was trained to generate the retinal image using vessel and lesion masks. Hence, we can develop retinal image with any type of severity, by providing the corresponding lesion mask. The synthetic abnormal images generated were shown to be realistic in the type of lesions produced and also the color, texture. These generated images are valuable in developing a CAD system

which detects and localizes haemorrhages as addition of synthetic data led to improvement in both SN and AUC by 17%.

We addressed the data sparsity problem by gradually lessening the human interaction. In the initial approach, the annotations for the training data were done only by experts. More training data was added in the process annotated by the model itself. This has a limitation on the number of patches that were predicted with high confidence by the model. In the second approach, the burden of annotation was shifted slightly to the crowd. The annotations collected from the crowd were improved using reliability factor. This reliability factor depends on the performance of the crowd on a dataset which had coarse spatial encoding. There are other scenarios where this spatial encoding is unavailable and the images are labelled as normal/ abnormal. To overcome the above limitation and to further reduce human interaction, we proposed the third method which generates the retinal images according to a given severity level. These generated images were also shown to be of similar quality as that of real images.

5.1 Future work

In the first part of the thesis, we proposed a co-training framework which uses unlabelled data for training. As a part of future work, we can experiment for the appropriate ratio of unlabelled data to labelled data for accurate classification. The algorithm performance can be improved by using advanced fusion algorithm. The unlabelled data which were chosen to add back to the training set, can be evaluated using active learning approach. Thus, ensuring the quality of unlabelled data added and improving the accuracy.

In the second part of the thesis, we have discussed about crowdsourcing to address the data sparsity issue. The crowd can further be guided towards viewing in a particular zone to improve accuracy and time efficiency. For example, in the case of hard exudate abnormality, the crowd can be asked to concentration on zone 2. Since, presence of HE in zone 2 indicates high severity of the abnormality. Further investigation can be done to train and evaluate on other abnormalities and also on techniques to find the optimal set of weights for RF computation.

In the last part of the thesis, we worked on generating HE/ HAE abnormalities in retinal images. This can be extended to synthesize other abnormalities in medical domain such as multiple sclerosis in brain MRI and glaucomatous cup in retinal image. The above method is limited to border artifacts and improper optic disc generation. The synthetic images can be improved further by improving the generator architecture such as progressive GAN [26]. This architecture proved to generate accurate minor details. Thus, future work can be done in the direction of incorporating this type of architecture to concentrate on minor details such as optic disc and vessel junctions.

The CAD system for HE/ HAE detection and generator in GAN architecture was based on U-net. Further study can be done to assess how the change in this network architecture affects the CAD performance and also the number of expert annotated data needed for training. For example, SegNet [7] can

be used which reduces the memory cost by reuse of pooling indices instead of transferring the entire feature map to the decoders.

Related Publications

1. Appan K. Pujitha, Gamalapati S. Jahnavi and Jayanthi Sivaswamy, "Detection of neovascularization in retinal images using semi-supervised learning." IEEE 14^{th} International Symposium on Biomedical Imaging (ISBI) - (2017).

2. Appan K. Pujitha and Jayanthi Sivaswamy, "Crowdsourced annotations with improved reliability, as an additional form of data augmentation." 4^{th} Asian Conference on Pattern Recognition (ACPR) - (2017).

3. Appan K. Pujitha and Jayanthi Sivaswamy, "Retinal image synthesis for CAD development." 15^{th} International Conference on Image Analysis and Recognition (ICIAR) - (2018).

4. Appan K. Pujitha and Jayanthi Sivaswamy, "A solution to overcome the sparsity issue of annotated data in medical domain." Special Issue of ACPR at CAAI Transactions on Intelligence Technology - (2018).

Bibliography

[1] Kaggle challenge diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection.

[2] U. Acharya et al. Computer based detection of diabetes retinopathy stages using digital fundus images. *Journal of Engineering in Medicine*, pages 502–512, Jun 2009.

[3] Agurto et al. Multiscale am-fm methods for diabetic retinopathy lesion detection. *IEEE Trans. Med. Imaging*, 29:502–512, Jun 2010.

[4] C. Agurto et al. A multiscale decomposition approach to detect abnormal vasculature in the optic disc. In *Computerized med. imaging and graphics*, pages 137–149, July 2015.

[5] M. Akram et al. Detection of neovascularization in retinal images using multivariate m-mediods based classifier. *Computerized Med. Imaging and Graphics*, 2013:346–357, Jun 2013.

[6] S. Albarqouni et al. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE TMI*, 35:1313–1321, May 2016.

[7] V. Badrinarayanan et al. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[8] A. Blum and T. Mitchell. Combining labelled and unlabelled data with co-training. In *Proc. Conf. on Computational learning theory*, pages 92–100, July 1998.

[9] A. Brébisson et al. Deep neural networks for anatomical brain segmentation. *CoRR*, abs/1502.02445, Jun 2015.

[10] D. L. Collins et al. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17 3:463–8, 1998.

[11] P. Costa et al. Adversarial synthesis of retinal images from vessel trees. In *Image Analysis and Recognition: 14th International Conference, ICIAR*, pages 516–523, 2017.

[12] E. Decencire et al. Feedback on a publicly distributed database: the messidor database. *Image Analysis & Stereology*, 33:231–234, aug 2014.

[13] A. F. Frangi and W. J. Niessen. Multiscale vessel enhancement filtering. In *MICCAI*, pages 130–137, October 1998.

[14] M. Ganz et al. Crowdsourcing for error detection in cortical surface delineations. *Int J CARS*, pages 12–161, Jan 2017.

[15] L. Giancardo et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Medical image analysis*, 16:216–226, Jan 2012.

[16] K. A. Goatman et al. Detection of new vessels on the optic disc using retinal photographs. *IEEE Trans. Med. Imaging*, 30:972–979, Apr 2011.

[17] Goodfellow et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[18] J. T. Guibas et al. Synthetic Medical Images from Dual Generative Adversarial Networks. *ArXiv e-prints*, Sept. 2017.

[19] V. Gulshan et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, pages 2402–2410, Dec 2016.

[20] G. Gupta et al. Computer-assisted identification of proliferative diabetic retinopathy in color retinal images. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5642–5645, Aug 2015.

[21] J. H.Friedman. Greedy function approximation: A gradient boosting machine. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 1189–1232, February 2002.

[22] G. D. Joshi et al. Colour retinal image enhancement based on domain knowledge. In *ICVGIP*, pages 591–598, Dec 2008.

[23] G. D. Joshi et al. Colour retinal image enhancement based on domain knowledge. In *Computer Vision, Graphics Image Processing, ICVGIP*, pages 591–598, Dec 2008.

[24] G. D. Joshi and J. Sivaswamy. Colour retinal image enhancement based on domain knowledge. In *Computer Vision, Graphics Image Processing, 2008. ICVGIP '08. Sixth Indian Conference on*, pages 591–598, Dec 2008.

[25] V. Kalesnykiene et al. Diaretdb1 diabetic retinopathy database and evaluation protocol, Jun 2007.

[26] T. Karras et al. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[27] T. Kauppi et al. Diaretdb0: Evaluation database and methodology for diabetic retinopathy algorithms. *Technical report*, 2006.

[28] T. Kohler et al. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. *International Symposium on Computer-Based Medical Systems, CBMS*, 00:95–100, 2013.

[29] L. Maier-Hein et al. Can masses of non-experts train highly accurate image classifiers? In *MICCAI*, pages 438–445, Jan 2014.

[30] L. Maier-Hein et al. Crowdsourcing for reference correspondence generation in endoscopic images. In *MICCAI*, pages 349–356, Sept 2014.

[31] L. Maier-Hein et al. Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In *MICCAI*, pages 616–623, Oct 2016.

[32] K. Maninis et al. Deep retinal image understanding. In *Medical Image Computing and Computer-Assisted Intervention, MICCAI*, pages 140–148, 2016.

[33] Menti et al. Automatic generation of synthetic retinal fundus images: Vascular network. In *Simulation and Synthesis in Medical Imaging: SASHIMI, Held in Conjunction with MICCAI*, pages 167–176, Oct 2016.

[34] D. Mitry et al. Crowdsourcing as a novel technique for retinal fundus photography classification: Analysis of images in the epic norfolk cohort on behalf of the ukbiobank eye and vision consortium. *PLoS ONE*, page 8(8), Aug 2013.

[35] J. Nayak et al. Automated identification of diabetic retinopathy stages using digital fundus images. *Jour. of Medical Systems*, 32:107–115, Oct 2008.

[36] D. Nie et al. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer-Assisted Intervention, MICCAI*, pages 417–425, Sept 2017.

[37] M. Prastawa et al. Simulation of brain tumors in mr images for evaluation of segmentation efficacy. *Medical Image Analysis*, 13:297 – 311, 2009.

[38] P. Prentai et al. Diabetic retinopathy image database(dridb): A new database for diabetic retinopathy screening programs research. In *ISPA*, pages 704–709, 2013.

[39] P. Prentai et al. Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion. *Comput Methods Programs Biomed*, 137:281–292, Oct 2016.

[40] M. Rezaei et al. Conditional adversarial network for semantic segmentation of brain tumor. *CoRR*, abs/1708.05227, Aug 2017.

[41] Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention, MICCAI*, pages 234–241, 2015.

[42] O. Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, May 2015.

[43] C. Sanchez et al. Retinal image analysis based on mixture models to detect hard exudates. *Med Image Anal*, 13(4):650–8, Aug 2009.

[44] Shankaranarayana et al. Joint optic disc and cup segmentation using fully convolutional and adversarial networks. In *Fetal, Infant and Ophthalmic Medical Image Analysis: OMIA Held in Conjunction with MICCAI*, pages 168–176, 2017.

[45] T. A. Syed and J. Sivaswamy. Latent factor model based classification for detecting abnormalities in retinal images. In *2015 Asian Conference on Pattern Recognition (ACPR)*, pages 411–415, Nov 2015.

[46] van Grinsven et al. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Trans. Med. Imaging*, 35:1273–1284, 2016.

[47] M. Vatanparast and A. Harati. A feasibility study on detection of neovascularization in retinal color images using texture. In *Computer and Knowledge Engineering, 2012 International Conf.*, pages 221–226, Oct 2012.

[48] Z. Wang et al. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[49] D. Welfer et al. A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images. *Comput Med Imaging Graph*, 34(3):228–35, Apr 2010.

[50] C. P. Wilkinson et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–82, 2003.

[51] Y. Yang et al. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. *CoRR*, abs/1705.00771, 2017.