A solution to overcome the sparsity issue of annotated data in medical domain

by

Pujitha Appan K., Jayanthi Sivaswamy

in

CAAI Transactions on Intelligence Technology

Report No: IIIT/TR/2018/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
September 2018

IET Journals
The Institution of Engineering and Technology

A solution to overcome the sparsity issue of annotated data in medical domain

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Appan K. Pujitha[1] *Jayanthi Sivaswamy*[2]

[1,2] *Center for Visual Information Technology, IIIT Hyderabad, India.*
** E-mail: pujitha.ak@research.iiit.ac.in and jsivaswamy@iiit.ac.in*

Abstract: Annotations are critical for machine learning and developing Computer Aided Diagnosis (CAD) algorithms. Good performance of CAD is critical to their adoption, which generally rely on training with a wide variety of annotated data. However, a vast amount of medical data is either unlabeled or annotated only at the image-level. This poses a problem for exploring data driven approaches like deep learning for CAD. Data augmentation is a popular solution in addressing this need but has limitations in adding real variability in the data. In this paper, we propose a novel crowd sourcing and synthetic image generation for training deep neural net-based lesion detection. The noisy nature of crowdsourced annotations is overcome by i) assigning a reliability factor for crowd subjects based on their performance and experience and ii) requiring region of interest markings rather than pixel-level markings from the crowd. A generative adversarial network-based solution is proposed to generate synthetic images with lesions to control the overall severity level of the disease. We demonstrate the reliability of the crowdsourced annotations and synthetic images, independently and also by presenting a solution for training the DNN with data drawn from a heterogeneous mixture of annotations, namely, very limited number of pixel-level markings by experts, crowdsourced ROI markings and synthetically generated data. Experimental results obtained for hard exudate detection from color fundus images show that training with processed/refined crowdsourced data/ synthetic images is effective as detection performance in terms of sensitivity improves by 25%/27% over training with just expert-markings.

1 Introduction

The latest paradigm shift of machine learning towards Deep Learning (DL) is spurred by its success on many longstanding computer vision tasks. This has motivated exploration of DL in wide ranging medical applications from disease detection [1] to segmentation [2]. Since DL is a data driven framework, its 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. Data augmentation (via geometric transformations) for robust training is a popular solution adopted by the computer vision community. However, this has limitations in the medical domain as it does not introduce any real variability that is essential for robust learning of abnormalities, normal anatomy etc. We examine this problem of sparsity of annotated data and explore 2 different avenues for solutions: (i) crowdsourcing and (ii) synthesis.

Crowdsourcing has been shown to be reliable [3–5] and useful to train classifiers [6]. Annotations have been crowd-sourced from fundus images, endoscopy and MRI of brain [3–5] for image level classification and reference correspondence. They have also been used to segment a surgical instrument from Laparoscopic images [6]. An active learning framework with crowdsourcing serves to reduce the burden on the crowd as it allows only low confident samples predicted by a model to be given to the crowd. Atlas forests were updated in [7] based on crowd refined annotations (on instrument boundary) to generate a new atlas. A convolutional neural network (CNN) was trained in [8] and the crowdsourced mitosis candidates (in a patch of size 33×33) were merged with an aggregation layer for updating the model. Crowd annotations are inherently noisy and hence merging them to derive a single ground truth (GT) is a key issue. Methods ranging from simple Majority Voting (MV) [6] to a stochastic modeling using Expectation Maximization [4] and introducing an aggregation layer in a CNN [8] have been proposed in literature.

A second avenue that is free of human annotation is the synthesis route. Image synthesis has been attempted to generate digital brain phantoms [9] and whole retinal images [10] using complex modeling. These have been aimed at aiding denoising, reconstruction and segmentation solutions. Recently, simulation of brain tumors in MR images [11] has also been explored to aid CAD algorithm development.

With the advent of DL, modeling of complex structures and synthesizing images has become easier with a class of neural networks called generative adversarial networks or GAN [12]. 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 [13], predicting CT from MRI images [14], synthesize normal retinal images from vessel mask [15], segmenting anatomical structures such as vessels [16] and optic disc/cup [17].

In this paper, we take up the problem of DR lesion detection from color fundus images and explore the use of the aforementioned 2 avenues to aid the development of robust CAD solution. Our contribution is three fold:

- We consider crowdsourcing as an independent (of 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.
- We propose a GAN for generating images with pathologies in a *controlled manner*.
- Finally, we illustrate how a heterogeneous mixture of annotations derived from experts, crowd and through synthesis can address the data sparsity problem .

The rest of the paper is organized as follows: §2 describes the method of collecting crowd annotations, aggregating the annotations, generation of synthetic images and developing a DL solution for hard exudate (HE) detection. §3 describes the datasets used,

implementation details and evaluation metrics used in the assessment of the proposed solution. §4 describes the experiments and results, finally conclusion is given in §5.

2 Methods

As a part of the pre-processing step, given retinal images are corrected for non-uniform illumination using luminosity and contrast normalization [18].

2.1 Crowdsourcing annotations

2.1.1 Subjects and tasks: A total of 11 engineering students volunteered to be 'crowd' members/subjects. Four of these were familiar with fundus images (L_k) while the rest were not familiar with any medical images (L_{nk}). The task given to the crowd subjects was twofold: i) determine whether the given image is normal/abnormal and ii) if abnormal, mark the ROI containing HE. A free hand annotation tool (Paint.Net [*]) was provided for the task. Fig. 1 shows a screenshot of the annotation tool.

2.1.2 Materials: 100 images were given to each subject. Of the 100, 6 images were from DIARETDB1 [19] which provides ROI markings from 4 experts. The remaining 94 (70 with HE and 24 with no lesions) were from MESSIDOR [20] which provides annotations only at the image-level. Thus, each of the 100 images that was selected for crowdsourcing has a label (normal/abnormal) and only 6 have additional information about locations of abnormalities. A sample image with HE is shown in Fig. 2 along with relevant landmarks.

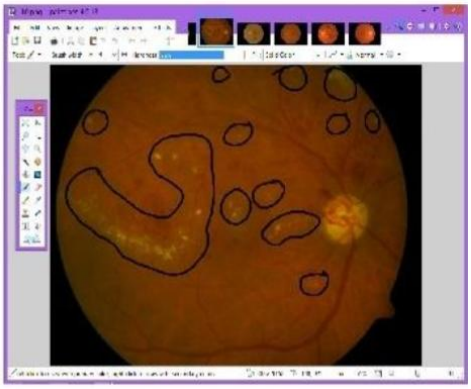


Fig. 1: Screenshot of annotation tool. Lesions area marked with black boundary by a subject

2.1.3 Processing Crowd Annotations: Our crowdsourcing exercise produced 11 annotations for each of the 100 images. These need to be integrated to derive a single annotation for every image. Merging the annotations via simple majority voting (MV) is likely to produce noisy annotation. Hence, a more elaborate procedure was designed for merging and GT derivation. This began with assigning a reliability factor (RF) to every subject i, shown in Fig. 3. Ideally, RF should rely on 2 factors: experience and performance of a subject. Experience can be determined via explicit queries. Performance needs to be assessed preferably by benchmarking against experts. A scheme was designed to reward a subject for good performance at both local ROI level (based on performance on the 6 images from DIARETDB1) and image-level and image-level (based on known labels of all 100 images). A score is given for each of these factors and the final RF is computed as a weighted sum of these scores. The

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

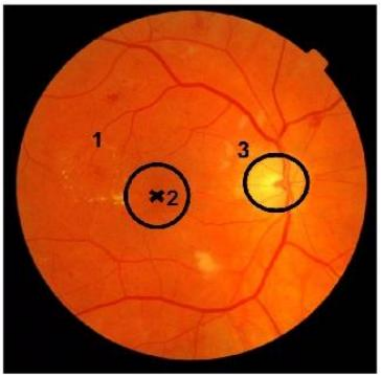


Fig. 2: Fundus image with labeled regions: 1 and 2 are zones of interest centered on macula and 3 is the optic disc.

reliability factor RF for the i^{th} subject was defined as :

$$RF(i) = \beta_1 S_1(i) + \beta_2 S_2(i) + \beta_3 S_3(i), \quad (1)$$

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

Scoring performance at an image-level: The crowd annotation for an image is binary (normal or abnormal) which is unlike the expert annotation for MESSIDOR. The latter encodes the location of HE (standard grading [20]): 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 S_1 to a subject not only based on correct labeling of normal images but rewarding them when their ROI is in the correct zone. S_1 is designed to be based on the True Positive Rate (TPR) and False Positive Rate (FPR) **(Eq. 12)** for each subject. The ROI location of an i^{th} subject is compared with the zonal labels (j) from MESSIDOR and the score $S_1(i)$ is calculated as follows:

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

Scoring 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:

$$S_2(i) = TPR(i) - FPR(i) + 1, \quad (3)$$

Scoring 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.

$$S_3(i) = 1(unfamiliar)\, or\, 2(familiar), \quad (4)$$

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

$$H_j = \sum_{i=1}^{11} RF(i) I_{ji}, \quad (5)$$

Here, I_{ji} is the annotated mask for the j^{th} image by the i^{th} subject. The majority voting based merging is when $RF(i) = 1, \forall i$ in the above equation.H map is finally binarised by thresholding.

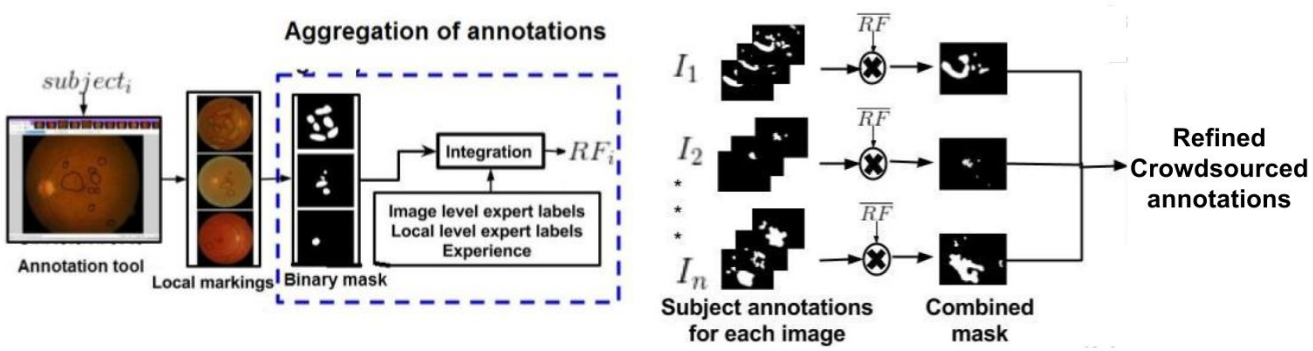


Fig. 3: Scheme for (a) RF computation for each subject (b) Aggregation of annotations using RF

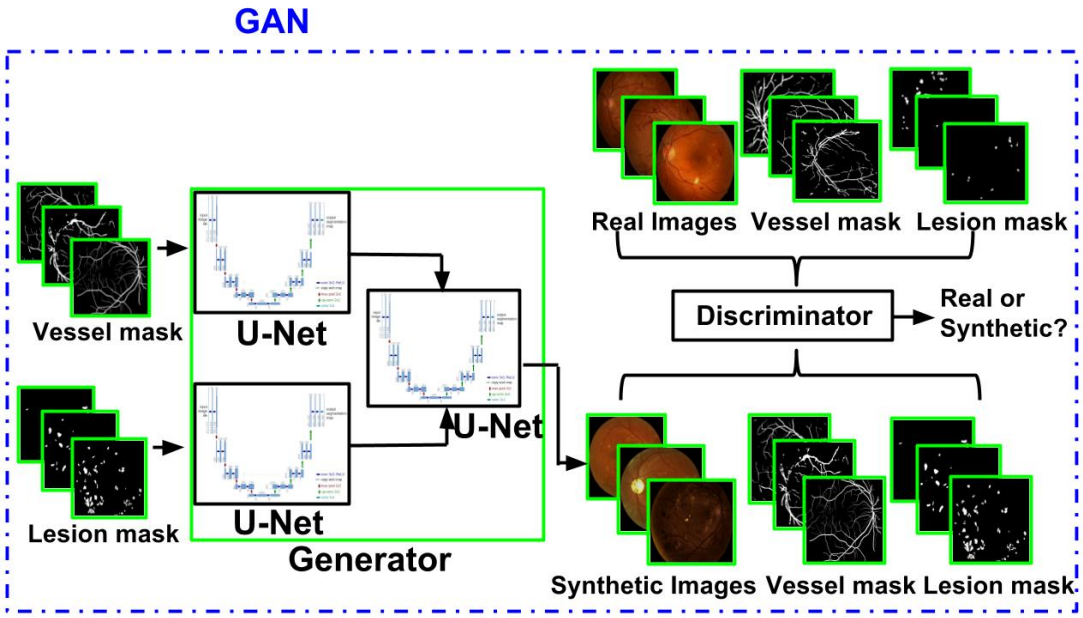


Fig. 4: Proposed GAN architecture for generation of abnormal retinal images.

2.1.4 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-level and image-level markings from the expert. In this strategy, the weights (β) are assigned by the DNN to different factors based on the performance of the subjects.

In this approach, we chose U-net to train the models. Let C_i be a subject and I_{ij} be the local annotation (image level annotation) given by the subject i on image j. Here, $i\epsilon\ \{1, 2, \cdots, 11\}$ as there are 11 subjects and $j\epsilon\ \{1, 2, \cdots, 76\}$ as 70 abnormal images from MESSIDOR and 6 images from DIARETDB1 are considered for training. Each U-net (U_i) is trained to detect hard exudates using the above 76 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 for each model by comparing against the local ground truth marked by experts. Further, the pixel level annotations are converted to image level annotations to asses performance at image level. The RF for each subject is given based on these values as:

$$RF(i) = \frac{SN(i) + PPV(i)}{2}, \tag{6}$$

2.2 GAN for Synthesis of Retinal Images with Pathologies

A second route we explore for deriving annotations is synthesis using a GAN made of a discriminator and a generator. A GAN learns a model as follows: 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. GAN-based synthesis of *Normal* retinal images has been demonstrated in [15] (from a vessel mask) with a single U-net for the generator and a 5-layer convolutional neural network for the discriminator. 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. The number of filters at each stage are reduced to half to simplify computations.

Our interest is in synthesis of images *with HE* to serve as exemplars for different stages of DR, which is based on the locations and

density of HE. Hence, we designed a GAN architecture (shown in Fig.4) 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 [15] 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 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 eqns.7-10 to produce sharp and realistic images.

(i) The adversarial loss function L_{adv} is defined as

$$L_{adv}(G,D) = \mathbb{E}_{(v,l),r\sim p_{data}((v,l),r)}[log(D((v,l),r))] + \mathbb{E}_{v,l\sim p_{data}(v,l)}[log(1-D((v,l),G(v,l)))], \quad (7)$$

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) [21] 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.

$$SSIM(p) = \frac{2\mu_r\mu_{G(v,l)} + C_1}{\mu_r^2 + \mu_{G(v,l)}^2 + C_1} \cdot \frac{2\sigma_{rG(v,l)} + C_2}{\sigma_r^2 + \sigma_{G(v,l)}^2 + C_2}, \quad (8)$$

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:

$$L_{SSIM} = 1 - \frac{1}{N}\sum_{p\epsilon P} SSIM(\tilde{p}), \quad (9)$$

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

$$L_1 = \mathbb{E}_{(v,l),r\sim p_{data}((v,l),r)}(\|r - G(v,l)\|_1), \quad (10)$$

The overall loss function to be minimized is taken to be

$$L(G,D) = L_{adv} + \lambda_1 L_1 + \lambda_2 L_{SSIM}, \quad (11)$$

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

2.3 CAD Solution for Hard Exudate Detection

The U-Net [22] was chosen to build a CAD solution for HE detection. This solution will be referred to as CADH. A standard architecture was chosen as the aim is to demonstrate that the crowdsourced annotations and synthetic images (generated by our proposed GAN) are a reliable resources in training even a basic U-net.

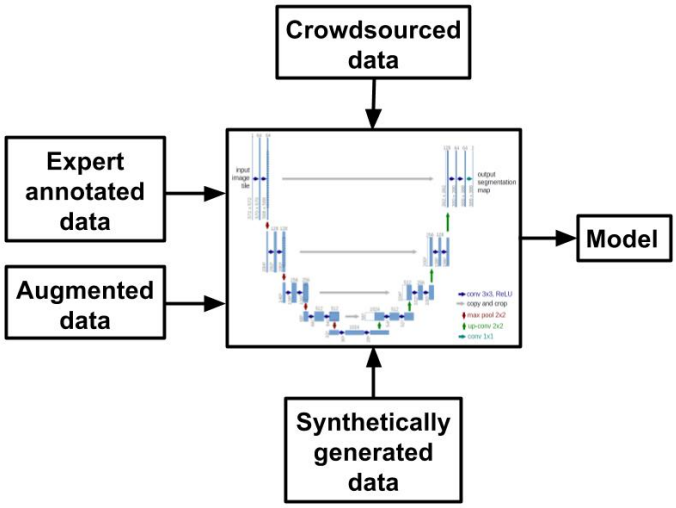


Fig. 5: Different sources of training data for CADH.

2.3.1 Preprocessing: Fundus images suffer from non-uniform illumination due to image acquisitions, camera limitations etc. This is corrected using luminosity and contrast normalization [18]. 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.

2.3.2 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 equal to crowdsourced images/ Synthetic images.

3 Implementation details

3.1 Datasets

Four public datasets, namely, (DRiDB [23], DMED [24], MESSIDOR and DIARETDB1) were used in this work. DMED has pixel level annotations (from 1 expert) whereas DIARETDB1 (4 experts) and DRiDB (1 expert) have ROI markings. The consensus marking of 3 experts was used to derive a binary mask in the case of DIARETDB1. The obtained binary mask was overlapped on the image and thresholded to get pixel level lesion mask. Images from all datasets were cropped and resized 512x512 before feeding them to GAN or CADH.

3.2 DNN for aggregation of crowd annotations

The training of each U-net consisted of 70 abnormal images from MESSIDOR and 6 abnormal images from DIARETDB1 given to the crowd for annotation. After augmentation it accounts to a total of 152 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 DNN for HE detection

Testing of CADH was done with DIARETDB1 (42 images). Training of CADH was done with the following datasets: expert annotations of DRIDB (31 images) and DMED (53 images); crowdsourced annotations of MESSIDOR (70 images). Additionally, annotations from synthetic images generated from GAN were also used which is described next.

3.3.1 Training Data for GAN: Training of the GAN requires both lesion and vessel masks. The lesion mask for the training data

Table 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 %

Table 2 Statistical significance value for label aggregation

Baseline	Combination	p-value
MV	I	0.3404
MV	I+L	0.0527
MV	I+L+E	0.4127
MV	I+E	0.4345

Table 3 Statistical significance value for different combination pairs of label aggregati

Baseline	Combination	p-value
I	I+L	0.067
I	I+L+E	0.7072
I	I+E	0.70
I+E	I+L	0.0358
I+L+E	I+E	0.442
I+L+E	I+L	0.0595

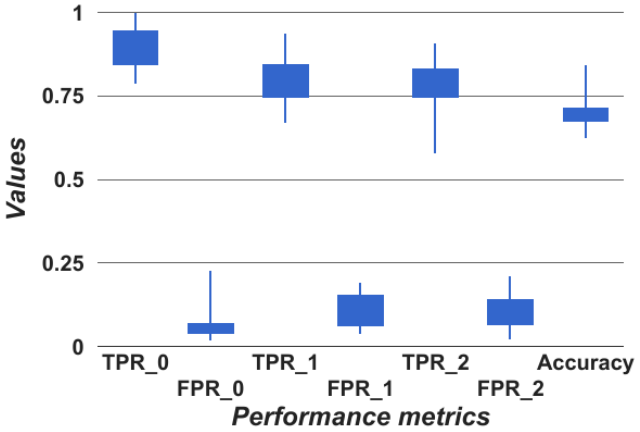


Fig. 6: Crowd annotation performance

are available from experts but vessels mask are available only for DRiDB. It is tedious and time consuming to mark the vessels in each of the retinal images. Hence, vessel masks were derived using a vessel segmentation method [25] which has proved to be robust to pathologies.

The synthetic retinal images were generated using GAN as follows. The required vessel and lesion masks were sourced from images selected randomly from DMED 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 with severity levels as in [20]. The lesions masks were derived to provide exemplars for each level using these rules. The position of lesions in each category were maintained by masking out few lesions or adding new lesions from another lesion mask randomly.

3.4 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 function 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 the 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 Evaluation metrics

Several experiments were conducted to assess the relative merits of crowdsourcing and synthesis of annotated data for training CADH. The merit was determined based on the HE detection performance. ***Crowd Annotations:*** Crowdsourced annotations was assessed 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:

$$TPR_i = \frac{N_{ii}}{\sum_{j=0}^{2} N_{ij}},$$
$$FPR_i = \frac{\sum_{j=0,j\neq i}^{2} Nij}{\sum_{j=0,j\neq i}^{2} Nij + \sum_{k=0,\neq i}^{2} \sum_{j=0,\neq i}^{2} N_{jk}}, \quad (12)$$

$$Accuracy = \frac{\sum_{i=0}^{2} N_{ii}}{\sum_{i=0}^{2} \sum_{j=0}^{2} N_{ij}}, \quad (13)$$

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

Aggregation of crowdsourced annotations: The label aggregation is assessed using TPR, FPR in respective zones and also accuracy. The statistical significance of the aggregated labels against baseline is also calculated by using p-value.

Assessment of generated synthetic images: The synthesized images were evaluated quantitatively and qualitatively (5 sample results are shown in Fig. **10**). The mean and standard deviation of the Q_v score described in [26] was computed over all images (42 abnormal) in DIARETDB1.

Assessment of CADH: 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.

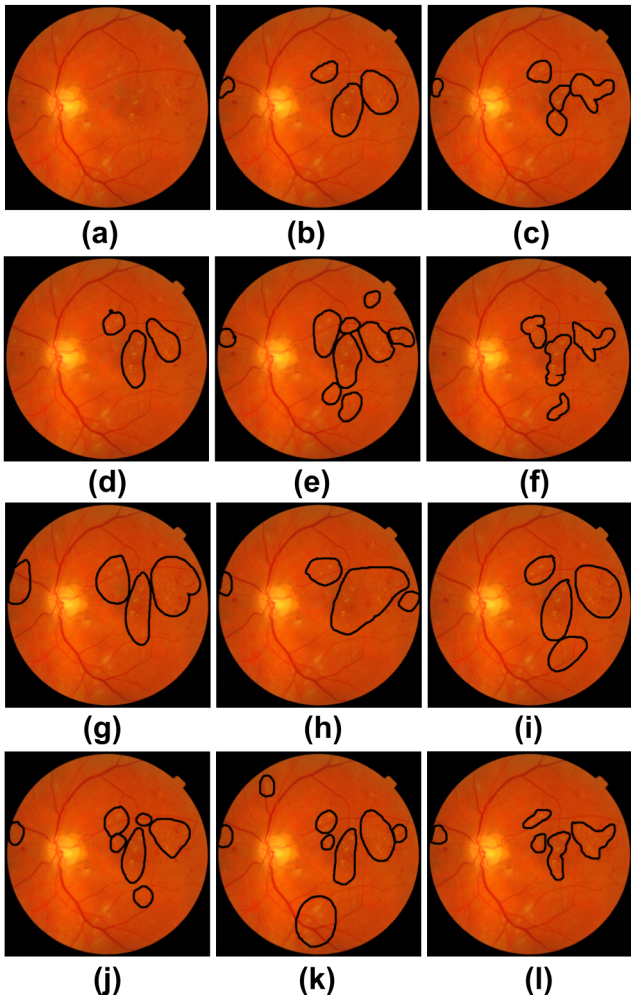


Fig. 7: Sample image and crowd annotations. (a) original image (b to l) markings by 11 subjects overlaid on the image.

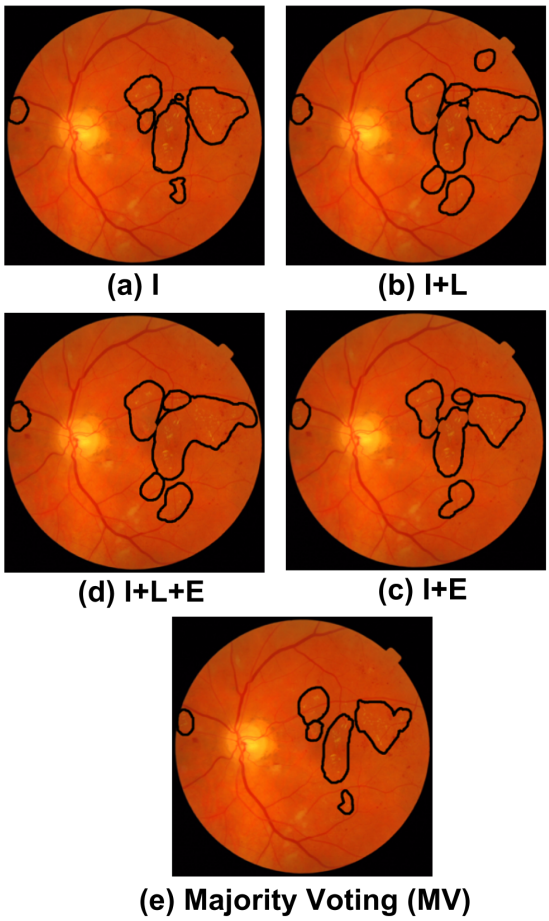


Fig. 8: The result of aggregation of the subject annotations considering different factors: I - Image level performance, L - Local level performance and E - Experience of the subject. Majority Voting is taken as baseline when none of the above information is available

5 Results

5.1 Crowdsourced annotations

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. Sample retinal image and respective eleven crowd annotations is shown in Fig. 7. The annotation performance is presented as a box plot in Fig. 6 for the 3 zonal labels/classes. The mean accuracy obtained was 70% and the class-wise TPR/FPR figures were: 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 the large zone 1 more accurately than much smaller -+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.

Aggregation of labels: The impact of the factors that contribute to the proposed RF (see Eq.1) was studied by setting β_i-0/1. The obtained TPR and FPR are listed in Table. 1. The baseline is taken as majority voting in the following discussion. When only image-level performance is considered for the RF, there is a 10% improvement in accuracy which is boosted to 22% with the addition of performance at the local level. This is noteworthy as performance at local level is known only for 6% of the images given to the crowd. The aggregation result of different annotations considering different factors is shown in Fig. 8. The aggregation of annotations using DNN gave an overall accuracy of 97% with $TPR_0 - 100\%, FPR_0 - 12\%, TPR_1 - 100\%, FPR_1 - 17\%, TPR_2 - 92\%$ and $FPR_2 - 0\%$.

The statistical significance of different aggregations as compared to the baseline MV is shown in Table. 2. This shows that I+L is statistically significant compared to other combinations. The p-value is also reported for different pair of combinations to estimate the importance of each factor (Table. 3). We can infer that I+L can be considered as the alternate hypothesis, rejecting I+E and I+L+E. Experience does not seem to be beneficial for this experiment as performance suffers and also the statistical significance is less. This may be due to the fact that crowd was made of students and hence experience is really not meaningful.

5.2 Synthetic Image Generation (GAN)

Two sample retinal images (with HE) generated by the proposed GAN model are shown in Fig.10. The first two columns show the vessel and lesion masks given as input to the GAN. The next two columns show the synthesized and the corresponding real images. The synthesized 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.

The mean/ standard deviation of Q_v computed over all images (42) with pathologies in DIARETDB1 are 0.074/0.017 and over all the synthetic images generated from vessel and lesion mask from DIARETDB1 is 0.082/0.02. These values are nearly equal indicating synthetic and real images are similar.

Table 4 HE detection performance with crowdsourced data and synthetic data for training (E-Expert, A-Augmentation, S-Synthetic and C-Crowd)

Source of training data (No. of images) (T-total no. of images)	SN(%)	PPV(%)	AUC
E (84) (T-84)	90	60.3	0.750
E (84) + A (70) (T-154)	89.8	61.6	0.765
E (84) + S (70) (T-154)	90	64	0.821
E (84) + C (I+L) (70) (T-154)	90.1	71.5	0.869
E (84) + C (MV) (70) + A (154) (T-308)	90	84.6	0.839
E (84) + C (I) (70) + A (154) (T-308)	90	85	0.879
E (84) + S (70) + A (154) (T-308)	90	82	0.894
E (84) + C (I+L) (70) + A (154) (T-308)	90.1	90.4	0.932
E (84) + C (70) + S (70) + A (224) (T-448)	89.8	92.0	0.942
E (84) + C (70) + S (140) + A (294) (T-588)	90	92.8	0.95
E (84) + C (140) + S (140) + A (364) (T-728)	89.7	93.4	0.956

5.3 CADH for Hard Exudate Detection

The utility of crowd sourcing and synthesizing annotations for CAD development was tested separately with 4 CADH models derived by training with different training sets. Denoting the set of real images with expert annotations as E, the crowdsourced annotations as C (Experiment-1) and the set of synthetic images generated by GAN with the corresponding lesion masks as S (Experiment-2), the variants of the training set considered were : (i) only E, (ii) E with data augmentation (E+A), (iii) E and C (S), (iv) E, C (S) with data augmentation (E+(C (S))+A). For an SN of 90%, the computed PPV and AUC values are reported in Table. 4 for Experiment-1 and Experiment-2. The SN vs PPV curve is shown in Fig. 9(a) for Experiment-1 and Fig. 9(b) for Experiment-2. The DNN output for HE detection and the corresponding expert annotations is shown in Fig. 11.

Based on the figures in the table, we observe the following. Data augmentation helps improve the AUC and PPV by about 2% each, whereas, inclusion of C (S) helps improve AUC by 15.8 (9.5)% and PPV by 18.5 (6.7)%. Finally, when the expert and crowdsourced annotations are augmented and added to the training set, there is a further improvement in AUC by 24.5 (19)% and in PPV by 50 (37)%.

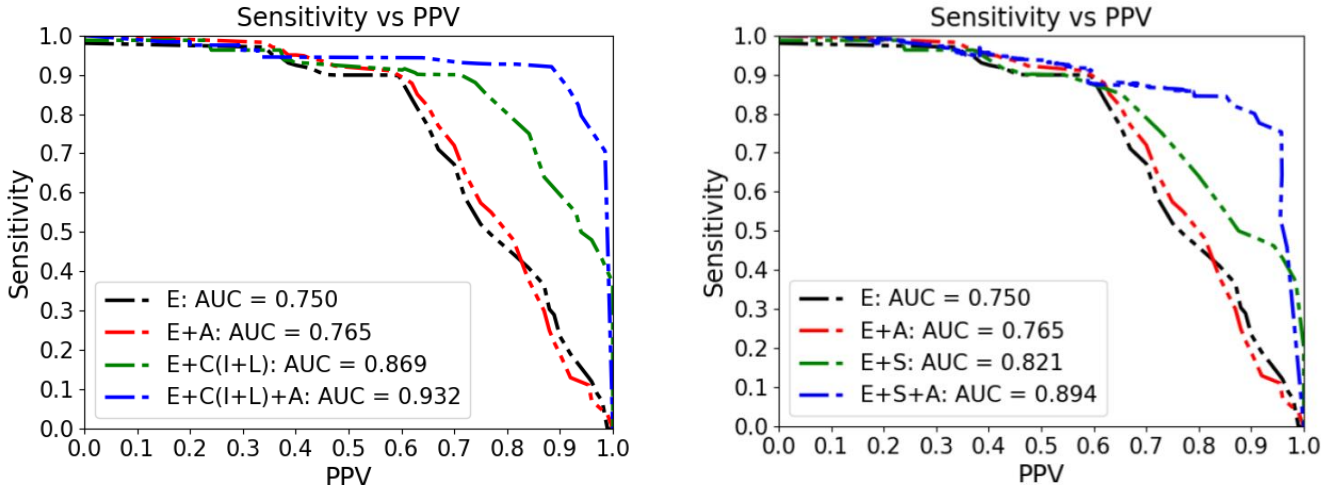
Setting PPV to 70% results in SN values ranging from 70% to 96% (Fig. 9); which is a 37% improvement in SN (similar level of improvement as that of PPV when SN is set constant at 90%). Thus, we conclude that C (S) are very effective in boosting the performance of CADH.

Most recent approaches for HE detection report at the image-level (normal or has HE) rather than at a local level. The exception is [27] 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.

6 Conclusion

In this paper, we have explored two options to address sparsity of annotated medical data which is critical for developing DL based CAD solutions. Crowdsourcing is an alternate source of annotation, but can be effective only if measures are taken to improve the reliability of annotations. The proposed RF concept aid weighted merging of annotations with good performance rewarded with a higher weight; it was shown that it is possible to assess the 'goodness' of a performance with very little cost (getting the crowd to annotate a small set of images previously annotated by experts). GAN-based synthesis is another alternative. A GAN solution was proposed to generate the retinal images with HE using vessel and lesion masks. This approach gives user a greater control as retinal images can be synthesized with any type of severity, by providing the corresponding lesion mask.

Our experiment results indicate that overall, the crowdsourced annotations and synthetic data (by themselves or in combination) are reliable for developing DL-based CAD solutions. Specifically,



(a) Experiments-1 (b) Experiments-2

Fig. 9: Performance of Deep Neural Net (SN vs PPV) for hard exudate detection

annotations via crowdsourcing data proved to be more effective than via synthesis (a PPV of 90% versus 82% for SN of 90%). However, crowdsourcing also involves manual work and hence comes at a higher cost. Combining the data from crowd and synthetic sources is a good compromise as it was seen to improve the performance (to 92% PPV). Synthetic data can be easily generated in abundance. The effect of changing the relative proportion of synthetic data in the training set can be seen from the results in the last two rows of Table 4. They suggest that increasing the proportion of synthetic data can boost the performance though the quantum of improvement appears small. The reasons for the same could be that the lesion masks were randomly chosen and hence the actual variability in data was not captured in training. A better scheme to ensure this variability could yield better improvement. Based on the encouraging results, future work can be directed towards exploring such solutions for other abnormality detection problems as well other modality images.

Acknowledgment

This work was supported by the Dept. of Electronics and Information Technology, Govt. of India under Grant: DeitY/R&D/TDC/13(8)/2013.

7 References

1 V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec 2016.

2 A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation." *CoRR*, vol. abs/1502.02445, Jun 2015.

3 D. Mitry, T. Peto, S. Hayat, J. E. Morgan, K. T. Khaw, and P. J. Foster, "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*, p. 8(8), Aug 2013.

4 L. Maier-Hein, S. Mersmann, D. Kondermann, C. Stock, H. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, A. l. Wekerle, S. Helfert, S. Bodenstedt, and S. Speidel, "Crowdsourcing for reference correspondence generation in endoscopic images." in *MICCAI, Boston, MA, USA*, Sept 2014, pp. 349–356.

5 M. Ganz, D. Kondermann, J. Andrulis, G. Moos Knudsen, and L. Maier-Hein, "Crowdsourcing for error detection in cortical surface delineations." *Int J CARS*, vol. 12, pp. 12–161, Jan 2017.

6 L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel, "Can masses of non-experts train highly accurate image classifiers?" in *MICCAI, Boston, MA, USA*, Jan 2014, pp. 438–445.

7 L. Maier-Hein, T. Rob, J. Grohl, B. Glocker, S. Bodenstedt, C. Stock, E. Heim, M. Gotz, S. Wirkert, and H. Kenngott, "Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence." in *MICCAI, Athens (Greece)*, Oct 2016, pp. 616–623.

8 S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images." *IEEE TMI*, vol. 35, pp. 1313–1321, May 2016.

9 L. Collins, A. Zijdenbos, V. Kollokian, J. Sled, N. Kabani, C. Holmes, and A. Evans, "Design and construction of a realistic digital brain phantom," *IEEE Transactions on Medical Imaging*, vol. 17 3, pp. 463–8, 1998.

10 L. Bonaldi, E. Menti, L. Ballerini, A. Ruggeri, and E. Trucco, "Automatic generation of synthetic retinal fundus images: Vascular network," in *Simulation and Synthesis in Medical Imaging: SASHIMI, Held in Conjunction with MICCAI, Loughborough, UK*, Oct 2016, pp. 167–176.

11 M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. C. Courville, Y. Bengio, C. Pal, P. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks." *Med Image Anal*, vol. 35, pp. 18–31, 2017.

12 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2672–2680.

13 M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, and C. Meinel, "Conditional adversarial network for semantic segmentation of brain tumor," *CoRR*, vol. abs/1708.05227, Aug 2017.

14 D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *Medical Image Computing and Computer-Assisted Intervention, MICCAI, Quebec City, Canada*, Sept 2017, pp. 417–425.

15 P. Costa, A. Galdran, M. Meyer, A. Mendonca, and A. Campilho, "Adversarial synthesis of retinal images from vessel trees," in *Image Analysis and Recognition: 14th International Conference, ICIAR, Montreal, Canada*, 2017, pp. 516–523.

16 T. Virdi, J. T. Guibas, and P. S. Li, "Synthetic Medical Images from Dual Generative Adversarial Networks," *ArXiv e-prints*, Sep. 2017.

17 S. M. Shankaranarayana, K. Ram, K. Mitra, and M. Sivaprakasam, "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,QuÃl'bec City, Canadaur*, 2017, pp. 168–176.

18 G. Joshi and J. Sivaswamy, "Colour retinal image enhancement based on domain knowledge." in *ICVGIP, Bhubaneswar, India*, Dec 2008, pp. 591–598.

19 T. Kauppi, V. Kalesnykiene, J. K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kalviainen, and J. Pietila, "Diaretdb1 diabetic retinopathy database and evaluation protocol." 2007.

20 E. DecenciÃÍre, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J. C. Klein, "Feedback on a publicly distributed database: the messidor database." *Image Analysis & Stereology*, vol. 33, pp. 231–234, aug 2014.

21 Z. Wang, A. Conrad Bovik, H. Rahim Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

22 O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation." *CoRR*, vol. abs/1505.04597, May 2015.

23 P. Prentasic, S. Loncaric, Z. Vatavuk, G. Bencic, M. Subasic, T. PetkoviÄĞ, L. Dujmovic, M. Malenica Ravlic, N. Budimlija, and R. Tadic, "Diabetic retinopathy image database (dridb): A new database for diabetic retinopathy screening programs research." in *ISPA, Trieste, Italy*, 2013, pp. 704–709.

24 L. Giancardo, F. Meriaudeau, T. Karnowski, Y. Li, S. Garg, K. W Tobin, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets." *Medical image analysis*, vol. 16, pp. 216–226, Jan 2012.

25 K.-K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. Van Gool, "Deep retinal image understanding," in *MICCAI, Athens (Greece)*, 2016, pp. 140–148.

26 T. Kohler, A. Budai, M. Kraus, J. Odstrcilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," *International Symposium on Computer-Based Medical Systems, CBMS, Portugal*, vol. 00, pp. 95–100, 2013.

27 P. Prentasic and S. Loncaric, "Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion." *Comput Methods Programs Biomed*, vol. 137, pp. 281–292, Oct 2016.

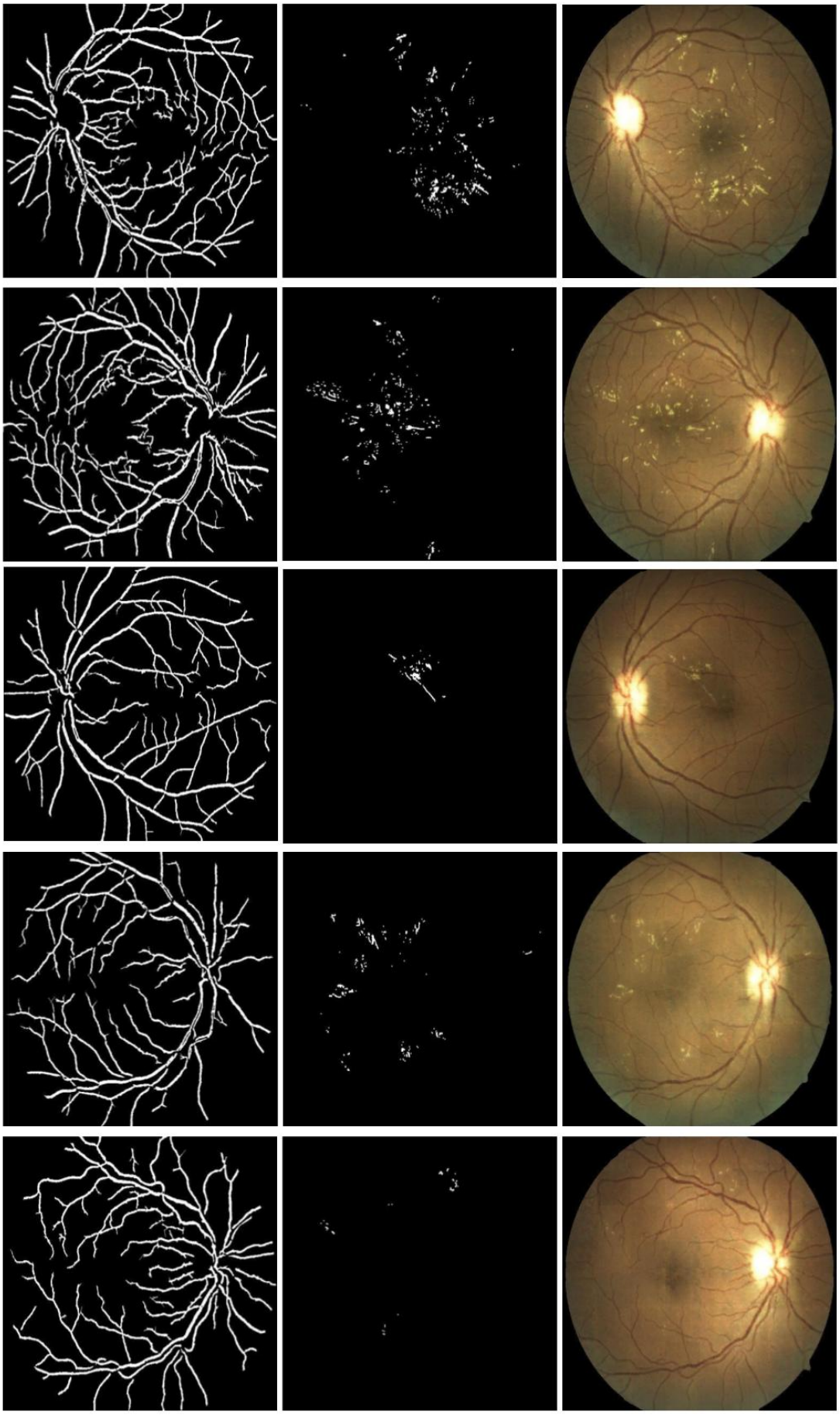


Fig. 10: Results of GAN-based image synthesis. From left to right: vessel mask, lesion mask, synthetic image. From top to bottom: first two sample images fall under zone 1 and the last three images fall under zone 2.

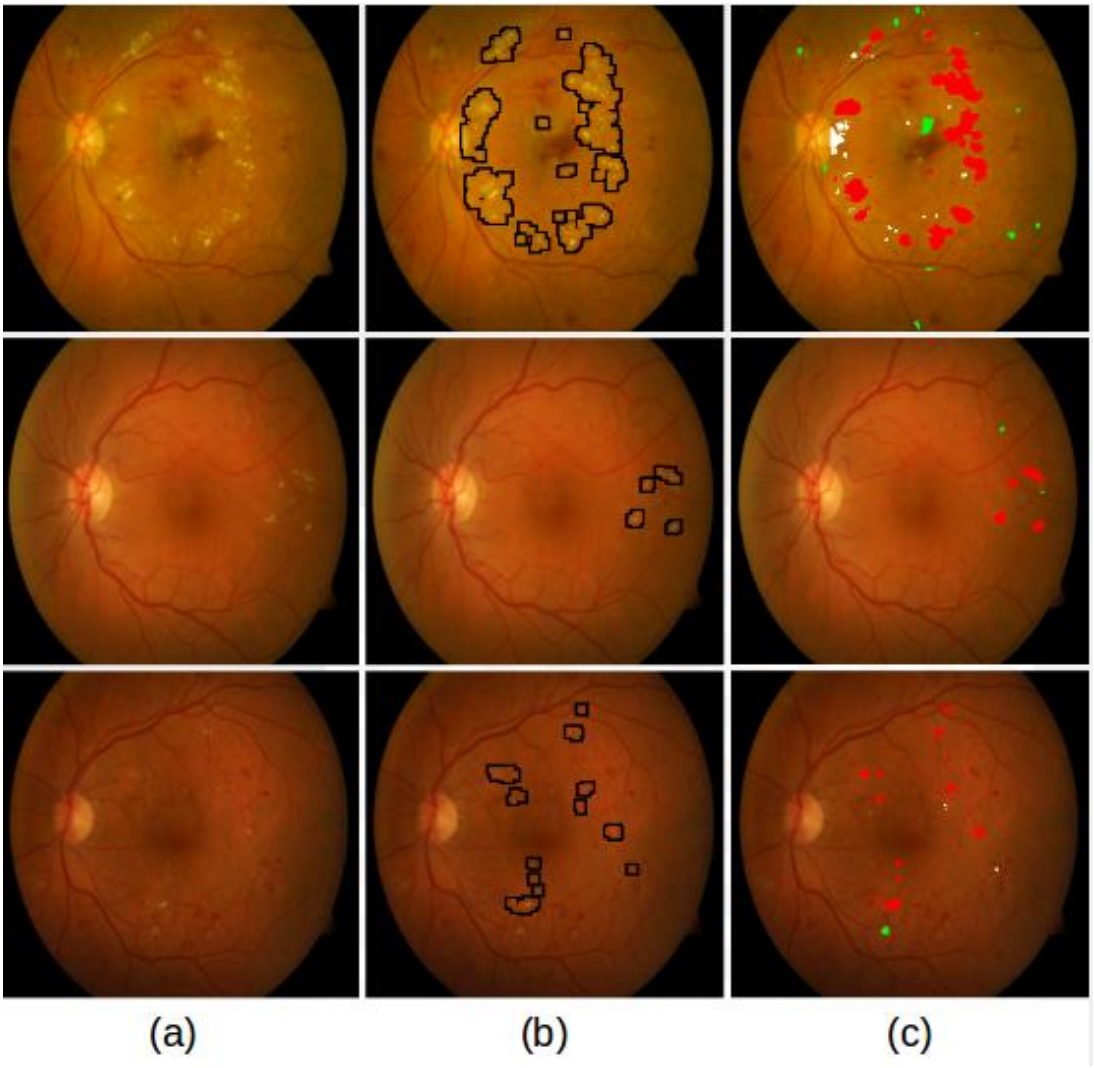


Fig. 11: (a) Sample images (b) ground truth marked by experts (c) DNN output for HE detection.