

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

Appan K. Pujitha
 Center for Visual Information Technology
 International Institute of Information Technology (IIIT)
 Hyderabad, India.
 Email: pujitha.ak@research.iiit.ac.in

Jayanthi Sivaswamy
 Center for Visual Information Technology
 International Institute of Information Technology (IIIT)
 Hyderabad, India.
 Email: jsivaswamy@iiit.ac.in

Abstract—Annotations are critical for machine learning and developing Computer Aided Detection (CAD) algorithms. However, a majority of medical data is either unlabeled or annotated only at the image-level. This poses a problem specifically for employing deep learning based approaches for CAD development as they require large amounts of annotated data for training. Data augmentation is a popular solution to address this need. We explore crowdsourcing as a solution for training a deep neural network (DNN) for lesion detection. Our solution employs a strategy to 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 present 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 and crowdsourced ROI markings. Experimental results obtained for hard exudate detection from color fundus images show that training with processed/refined crowdsourced data is effective as detection performance improves by 25% over training with just expert-markings and by 11% over training with annotation derived using majority voting among the crowd.

Keywords—Crowdsourcing; Reliability Factor; Deep Neural Net.

I. 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].

¹ The DL 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. 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.

¹This work was supported by the Dept. of Electronics and Information Technology, Govt. of India under Grant: DeitY/R&D/TDC/13(8)/2013

Crowdsourcing has been considered as a solution to address the issue of sparsity of annotated data. It has been shown to be reliable [3]–[5] and useful to train classifiers [6]. In [3]–[5], annotations were crowd-sourced from fundus images, endoscopy and MRI of brain, while in [6], 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 [7] [8]. 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 fed back to update the model. Atlas forests are used in [7] 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 [8] 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) [6] to a stochastic modeling of the crowdsourced information using Expectation Maximization [4] and introducing an aggregation layer in a CNN [8].

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 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

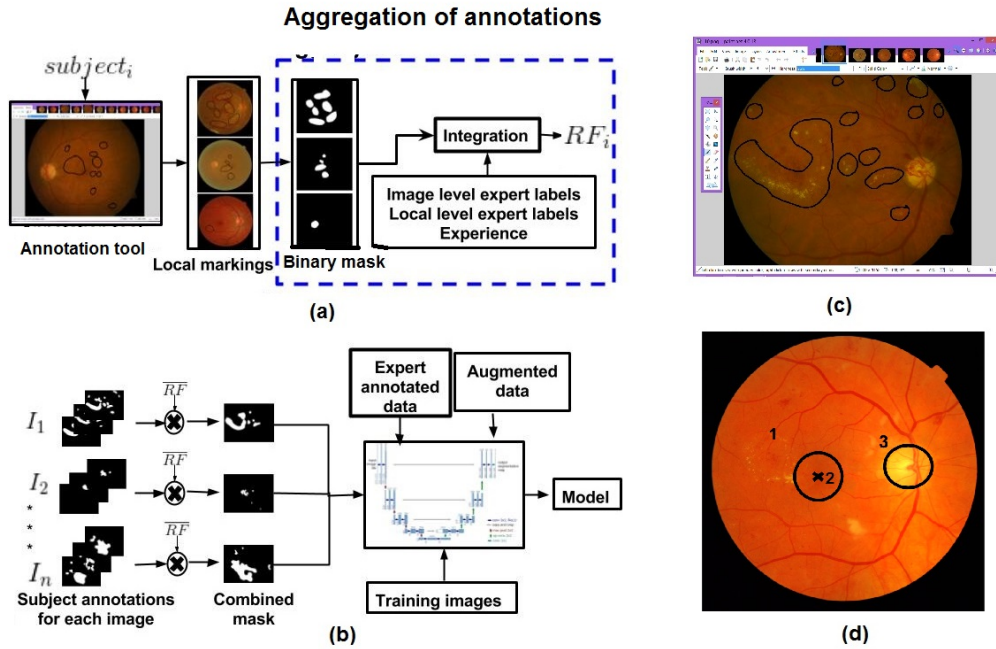


Figure 1: Scheme for (a) RF computation for each subject (b) Aggregation of annotations using RF and training U-Net with heterogeneous mixture of annotations; (c) Screenshot of annotation tool. Lesions area marked with black boundary by a subject (d) Fundus image with labeled regions: 1 and 2 are zones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

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%.

II. 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. 1 (a,b).

A. Collection of crowd Annotation

The subjects of the crowd are given a free hand annotation tool (Paint.Net ²) for the task. Fig. 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

²<http://www.getpaint.net/download.html>

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 [9] which provides ROI markings from 4 experts; 94 images were from MESSIDOR [10] 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. 1(d).

B. 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 2 factors: experience of the subject and their performance. The former can be obtained with explicit queries. The assessment of the latter 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 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 :

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

where $S_j \in [0, 2]$ are scores determined based on the factors mentioned above and described in detail next; β_i

$\epsilon \in [0, 1]$ 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 [10]): 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 Positive Rate (TPR) and False Positive Rate (FPR) (Eq. 6) 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:

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

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:

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

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 :

$$H_j = \sum_{i=1}^{11} RF(i)I_{ji} \quad (4)$$

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:

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

The obtained heat map is finally binarised by thresholding.

C. Deep Neural Network

We chose the U-Net [11] 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 [12]. 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° to 25° , 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.

III. IMPLEMENTATION AND EVALUATION DETAILS

Datasets: Four public datasets, (DRiDB [13], DMED [14], 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 256×256 before feeding to the DNN.

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.

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 0.5×10^{-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.

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),

Table I: 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 %

TPR, FPR and accuracy were calculated as follows:

$$TPR_i = \frac{N_{ii}}{\sum_{j=0}^2 N_{ij}} \quad (6)$$

$$FPR_i = \frac{\sum_{j=0, j \neq i}^2 N_{ij}}{\sum_{j=0, j \neq i}^2 N_{ij} + \sum_{k=0, k \neq i}^2 \sum_{j=0, j \neq i}^2 N_{jk}}$$

$$Accuracy = \frac{\sum_{i=0}^2 N_{ii}}{\sum_{i=0}^2 \sum_{j=0}^2 N_{ij}} \quad (7)$$

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: $SN = \frac{TP}{TP+FN}$ and $PPV = \frac{TP}{TP+FP}$. 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.

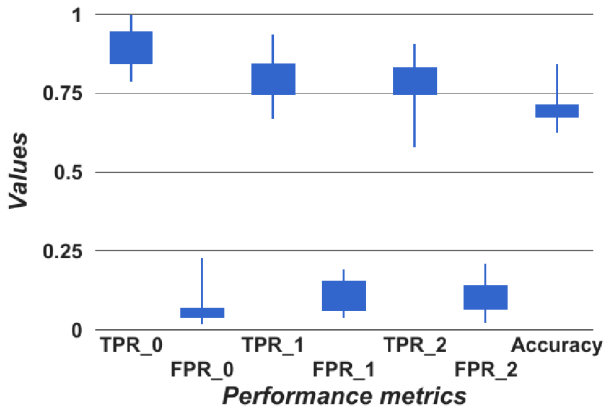


Figure 2: Box plot of crowd annotation performance metrics

A. Experiments and Results

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. 2 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.

Aggregation of labels: The impact of the terms in Eq.1 is studied by setting $\beta_i=0/1$. The obtained TPR and FPR are listed in TableI. 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.

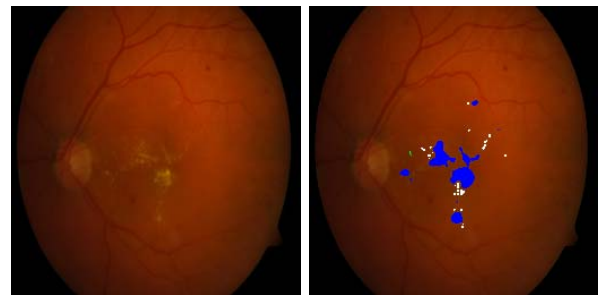


Figure 4: (a) A sample image and (b) DNN output for HE detection. Color code: blue-TP, green-FP, white-FN

Deep Neural Net: 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) and finally iv) E,C and augmented data derived from E+C. Sample results of

Table II: 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	0.839
Expert + Crowd (I) + Augmentation (308)	90	85	0.879
Expert + Crowd (I+L) + Augmentation (308)	90.1	90.4	0.932

HE detection are shown in Fig. 4 for iv. The difference between expert annotation and U-net detection are shown in terms of TP, FP and FN. Further results with expert annotations are shown in Fig. 6.

The assessment is based on SN, PPV and the AUC values which are reported in Table. II. 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. 5) 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. 5.

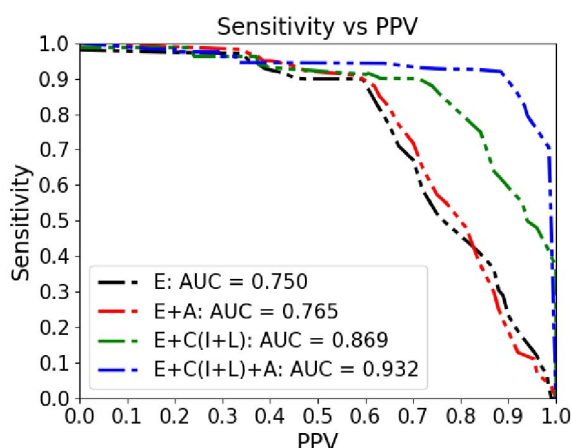


Figure 5: Performance of Deep Neural Net for hard exudate detection

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 [15] reports a SN of 90.2% and

PPV of 96.8% based on ROI detection while [16] [17] report pixel based classification with SN ranging from 70-78% and PPV ranging from 75-78%.

IV. 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. The study was done only for 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. Future investigation can be into this aspect and on techniques to find the optimal set of weights for RF computation.

REFERENCES

- [1] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." *JAMA*, pp. 2402-2410, Dec 2016.
- [2] A. Brébisson *et al.*, "Deep neural networks for anatomical brain segmentation." *CoRR*, vol. abs/1502.02445, Jun 2015.
- [3] 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*, p. 8(8), Aug 2013.

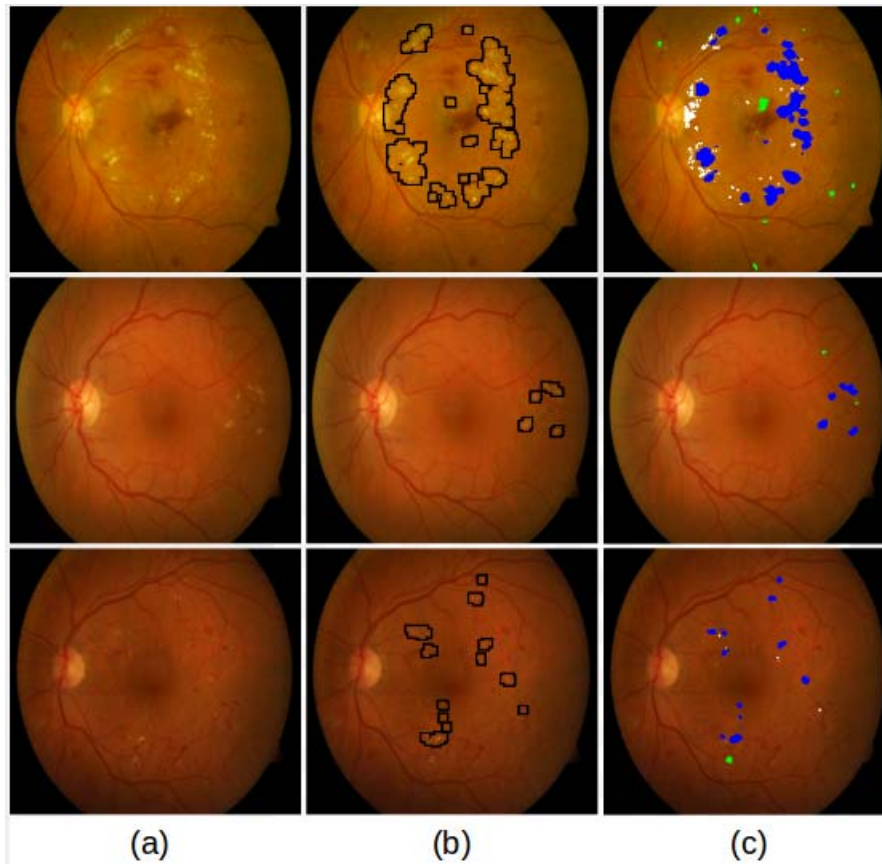


Figure 6: (a) Sample images (b) Expert annotation (c) DNN output for HE detection. Color code: blue-TP, green-FP, white-FN

- [4] L. Maier-Hein *et al.*, “Crowdsourcing for reference correspondence generation in endoscopic images.” in *MICCAI*, Sept 2014, pp. 349–356.
- [5] M. Ganz *et al.*, “Crowdsourcing for error detection in cortical surface delineations.” *Int J CARS*, pp. 12–161, Jan 2017.
- [6] L. Maier-Hein *et al.*, “Can masses of non-experts train highly accurate image classifiers?” in *MICCAI*, Jan 2014, pp. 438–445.
- [7] L. Maier-Hein *et al.*, “Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence.” in *MICCAI*, Oct 2016, pp. 616–623.
- [8] S. Albarqouni *et al.*, “Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images.” *IEEE TMI*, vol. 35, pp. 1313–1321, May 2016.
- [9] V. Kalesnykiene *et al.*, “Diaretdb1 diabetic retinopathy database and evaluation protocol.” 2007.
- [10] E. Decencire *et al.*, “Feedback on a publicly distributed database: the messidor database.” *Image Analysis & Stereology*, vol. 33, pp. 231–234, aug 2014.
- [11] O. Ronneberger *et al.*, “U-net: Convolutional networks for biomedical image segmentation.” *CoRR*, vol. abs/1505.04597, May 2015.
- [12] G. D. Joshi *et al.*, “Colour retinal image enhancement based on domain knowledge.” in *ICVGIP*, Dec 2008, pp. 591–598.
- [13] P. Prentai *et al.*, “Diabetic retinopathy image database (dridb): A new database for diabetic retinopathy screening programs research.” in *ISPA*, 2013, pp. 704–709.
- [14] L. Giancardo *et al.*, “Exudate-based diabetic macular edema detection in fundus images using publicly available datasets.” *Medical image analysis*, vol. 16, pp. 216–226, Jan 2012.
- [15] C. Sanchez *et al.*, “Retinal image analysis based on mixture models to detect hard exudates.” *Med Image Anal*, vol. 13(4), pp. 650–8, Aug 2009.
- [16] P. Prentai *et al.*, “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.
- [17] D. Welfer *et al.*, “A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images.” *Comput Med Imaging Graph*, vol. 34(3), pp. 228–35, Apr 2010.