Computers in Biology and Medicine 136 (2021) 104725

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep Bayesian baseline for segmenting diabetic retinopathy lesions: Advances and challenges

Azat Garifullin [a,*], Lasse Lensu [a], Hannu Uusitalo [b,c]

[a] Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, LUT University, P.O. Box 20, 53851, Lappeenranta, Finland
[b] Department of Ophthalmology, ARVO F313, 33014, Tampere University, Finland
[c] Tays Eye Center, Tampere University Hospital, P.O. Box 2000, 33520, Tampere, Finland

ARTICLE INFO

Keywords:
Bayesian deep learning
Diabetic retinopathy
Lesion segmentation
Microaneurysm
Hard exudate
Soft exudate
Haemorrhage

ABSTRACT

Early diagnosis of retinopathy is essential for preventing retinal complications and visual impairment due to diabetes. For the detection of retinopathy lesions from retinal images, several automatic approaches based on deep neural networks have been developed in the recent years. Most of the proposed methods produce point estimates of pixels belonging to the lesion areas and give no or little information on the uncertainty of method predictions. However, the latter can be essential in the examination of the medical condition of the patient when the goal is early detection of abnormalities. This work extends the recent research with a Bayesian framework by considering the parameters of a convolutional neural network as random variables and utilizing stochastic variational dropout based approximation for uncertainty quantification. The framework includes an extended validation procedure and it allows analyzing lesion segmentation distributions, model calibration and prediction uncertainties. Also the challenges related to the deep probabilistic model and uncertainty quantification are presented. The proposed method achieves area under precision-recall curve of 0.84 for hard exudates, 0.641 for soft exudates, 0.593 for haemorrhages, and 0.484 for microaneurysms on IDRiD dataset.

1. Introduction

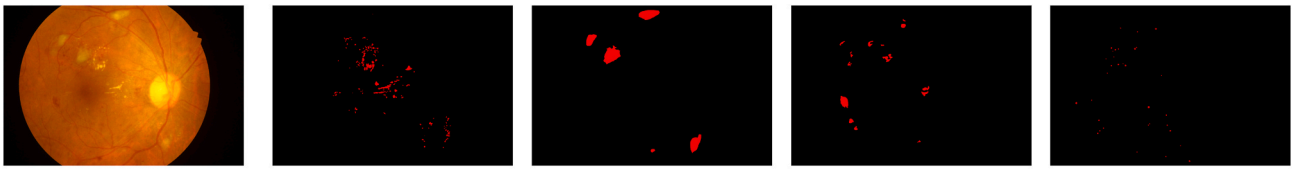
Diabetic retinopathy (DR) is the most common complication of diabetes mellitus and can lead to a vision loss if not treated properly [1]. Screening of the condition and early detection of retinal abnormalities is essential and consists of examining retinal images for diabetic lesions. In the early stages of the disease, these lesions are small, typically have low contrast and sometimes difficult to detect for humans. The core of the screening problem is, however, the amount of images that need to be analyzed. Thus, automatic retinal image analysis methods are required. One way to build an assisting system is to train an end-to-end classifier that processes an input image and yields a diabetic retinopathy grade [2]. These systems are often criticized for being black-boxes producing results that are difficult to interpret [3]. As an alternative, one can train a segmentation model that processes the input image and produces a segmentation map where each element represents the probability of being a lesion. This way the diagnosis can be inferred from the segmentation maps by counting the detected lesions.

In recent years, the field of DR lesion segmentation has advanced with the introduction of new retinal image datasets making it possible to accelerate research in related computer vision methods [4]. One of the most widely used benchmarks is Indian diabetic retinopathy image dataset (IDRiD) dataset providing high-quality ground truth masks for hard exudates, soft exudates, haemorrhages and microaneurysms. Porwal et al. [5] published the results of the IDRiD challenge held in 2018. The best performing algorithms were represented by deep convolutional architectures such as U-Net [6], dense fully-convolutional network (Dense-FCN) [7] and Mask-RCNN [8] or their variants. It should be noted that the data is very unbalanced and achieving high sensitivity was a challenge for many algorithms. To overcome this issue, the authors used balanced cross-entropy [9] and dice loss [10]. Due to the high resolution of the images, the models were trained in a patchwise manner.

Guo et al. [11] proposed a multi-scale feature fusion method to handle issues with small lesion detection. Binary cross-entropy (BCE) loss with balancing coefficients was used to improve the sensitivity. The model was trained with full images resized to 1440×960 pixels without any further preprocessing. Yan et al. [12] proposed mutually local-global U-Net mitigating the problems of patchwise training which does not capture the global context. The proposed architecture consists

* Corresponding author.
E-mail addresses: azat.garifullin@lut.fi (A. Garifullin), lasse.lensu@lut.fi (L. Lensu), hannu.uusitalo@tuni.fi (H. Uusitalo).

https://doi.org/10.1016/j.compbiomed.2021.104725
Received 28 June 2021; Received in revised form 29 July 2021; Accepted 30 July 2021
Available online 6 August 2021



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

Fig. 1. (a) An example of IDRiD image with ground truth masks for (b) hard exudates, (c) soft exudates, (d) haemorrhages, and (e) microaneurysms.

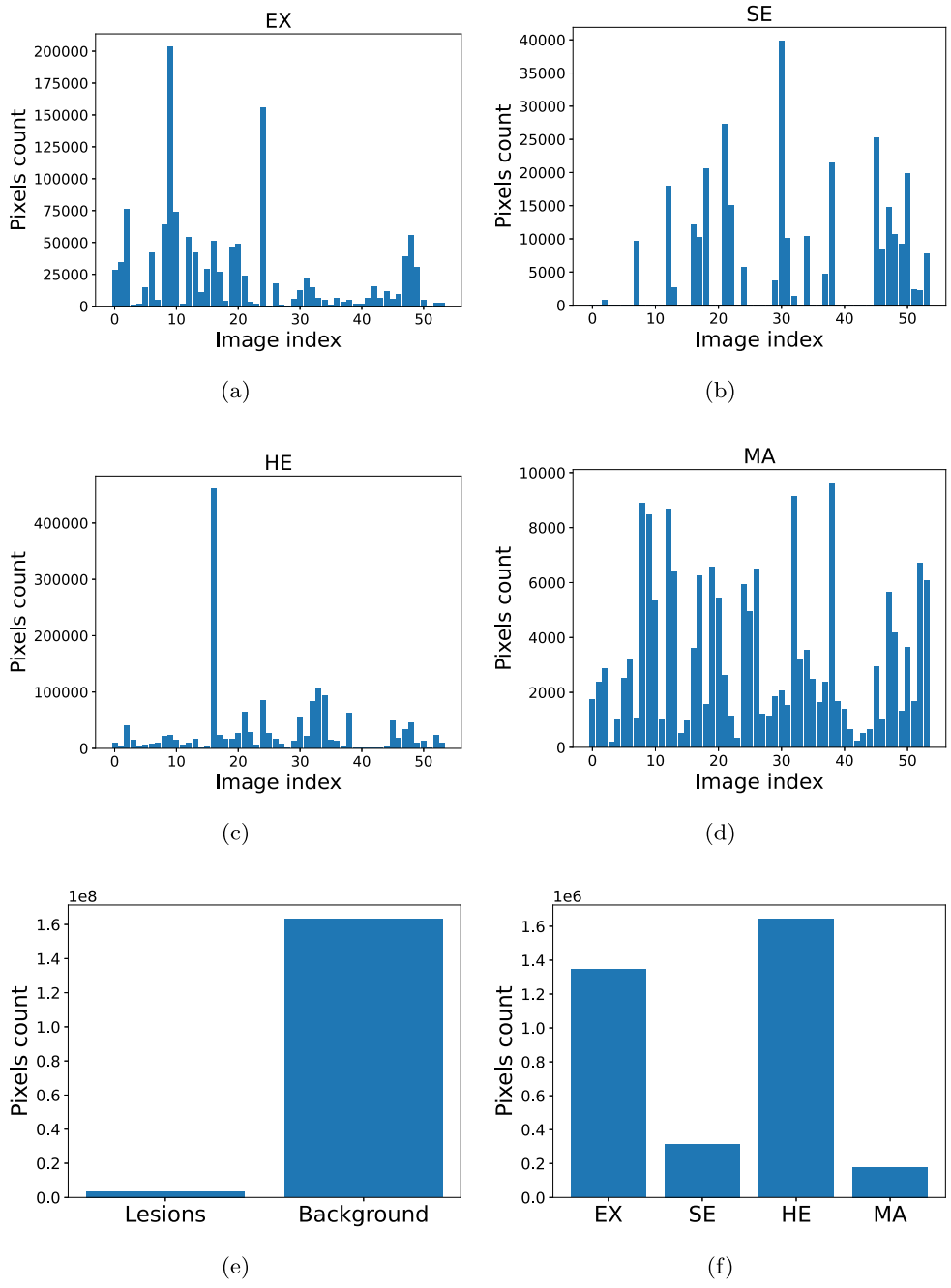


Fig. 2. Statistics of lesions in the train set. The number of positive pixels per image for (a) hard exudates (EX), (b) soft exudates (SE), (c) haemorrhages (HE), and (d) microaneurysms (MA). (e) The number of pixels for the lesions and the background. (f) The number of positive pixels for each lesion for the whole dataset.

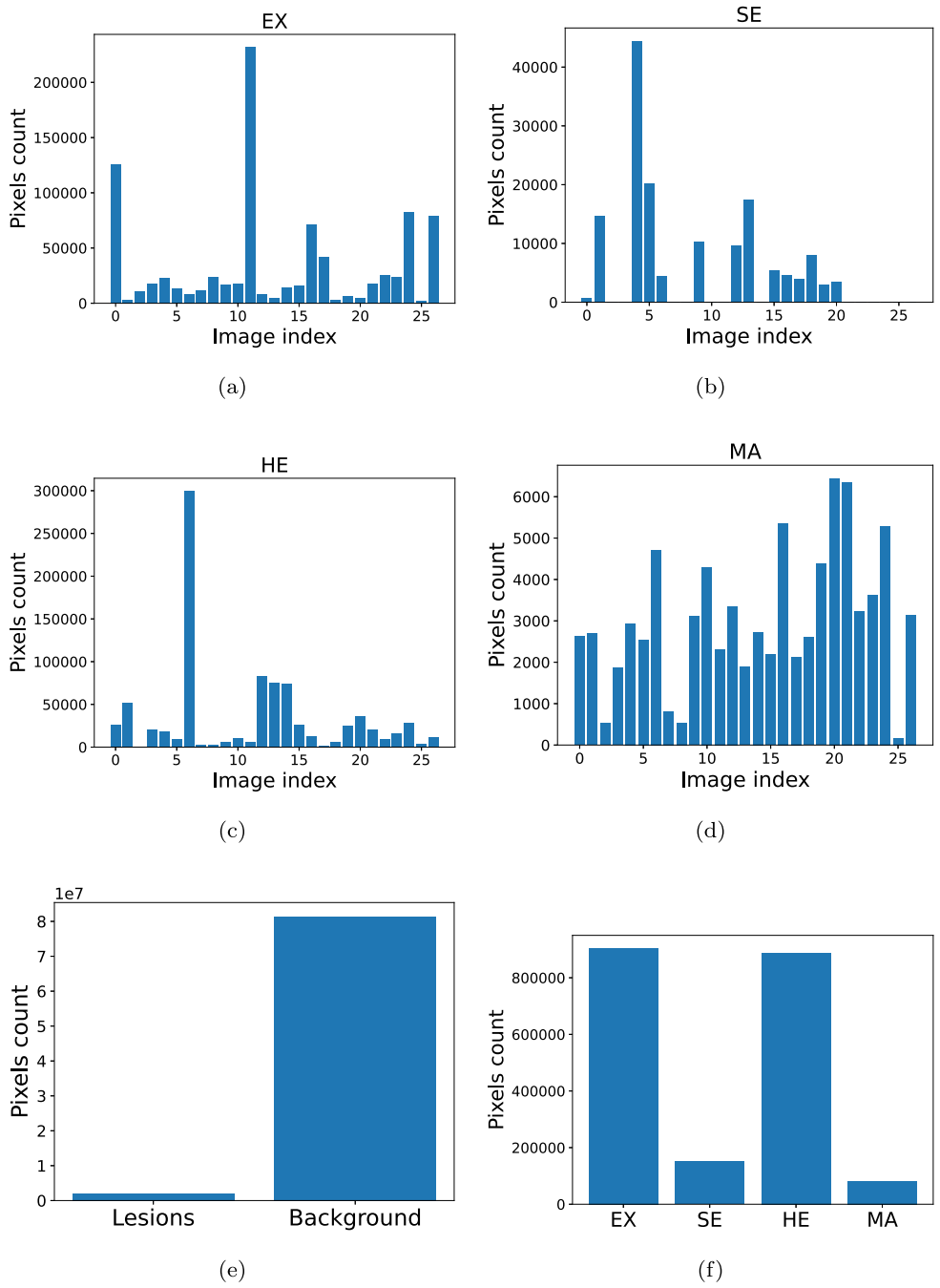


Fig. 3. Statistics of lesions in the test set. The number of positive pixels per image for (a) hard exudates (EX), (b) soft exudates (SE), (c) haemorrhages (HE), and (d) microaneurysms (MA). (e) The number of pixels for the lesions and the background. (f) The number of positive pixels for each lesion for the whole dataset.

of two U-Nets (global and local) that share the last layers of their decoders. Both networks are jointly trained minimizing weighted cross-entropy loss to deal with the class imbalance.

The aforementioned approaches consider only point estimates of the trained models and produced results. Thus, the question of reliability of a trained model arises. In this work, the problem is addressed by using Bayesian deep learning modeling a distribution over the learned parameters of the model and produces the segmentation results in a form of posterior predictive distribution. Recently, Bayesian deep learning models have started finding their applications in the area of retinal image analysis. Leibig et al. [13] evaluated dropout based uncertainty measures and demonstrated improved diagnostic performance using uncertainty-informed decisions. Filos et al. [14] proposed a new benchmark for deep Bayesian models with application to DR diagnosis also assessing the robustness of the models to out-of-distribution examples and distribution shift.

This work extends the preceding research with Bayesian DR lesion segmentation. To the best of authors' knowledge, this is the first work discussing the Bayesian approach for DR lesion segmentation. The aim is to establish a baseline that would inspire future research on the topic. The contributions of this work can be highlighted as follows:

1. The introduction of a novel Bayesian baseline for DR lesion segmentation allowing the analysis of segmentation distributions.

2. An assessment and analysis of model calibration and prediction uncertainties.
3. The presentation of an extended validation procedure for DR lesion segmentation task beyond the point estimates.

The rest of the paper is organized as follows: Section 2 describes the utilized dataset and gives the information about class imbalance and the statistics of labels, and Section 3 explains the Bayesian image segmentation setup, utilized data sampling approach and training details. Section 4 explains the evaluation protocol and presents the performance metrics together with the visualizations of the inferred results. Section 5 discusses faced issues and directions for future research. The results of the work are summarized in Section 6.

2. IDRiD dataset

The IDRiD dataset is a common benchmark for the diabetic retinopathy lesion segmentation [5]. It contains 54 train and 27 test images of resolution 4288×2848 with segmentation masks aiming to be spatially accurate for four lesion types: hard exudates, soft exudates, haemorrhages, and microaneurysms. An example image from the dataset is shown in Fig. 1.

The class imbalance can be visualized as a bar graph with the number of positive pixels for lesions for each image separately as well as for the whole dataset. The calculated statistics for the train and test sets are presented in Fig. 2 and Fig. 3.

3. Bayesian lesion segmentation

3.1. Background

The classical approaches give only point estimates for the class label probabilities and the model parameters are considered to be deterministic. In order to capture imperfect data labeling and image noise, the model outputs and learned parameters can be considered as random variables. The first approach captures the heteroscedastic aleatoric uncertainty that depends on the input data, whereas the second represents the epistemic uncertainty that models a distribution of the learned parameters. Here, a brief explanation for the lesion segmentation task is given below. More detailed explanations for the uncertainties can be found in Refs. [15,16].

Let f be a model, with parameters $\boldsymbol{\theta}$, that maps an input image $\mathbf{x}$ to a map of logits $\widehat{\mathbf{y}}$, accompanied by a map standard deviations $\boldsymbol{\sigma}$ of the logits:

$$[\widehat{\mathbf{y}}, \boldsymbol{\sigma}] = f(\mathbf{x}, \boldsymbol{\theta}). \tag{1}$$

Then, the probabilities of the class labels can be calculated as follows:

$$\widehat{\mathbf{p}} = \mathrm{sigmoid}(\widehat{\mathbf{y}} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\odot$ stands for the Hadamard product and $\boldsymbol{\varepsilon}$ are sampled during inference.

Epistemic uncertainty can be captured by considering the model parameters to be a random variable and making use of the following posterior predictive:

$$p(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}, \tag{3}$$

where $\mathcal{D}$ denotes a dataset of input-output pairs.

Typically, the parameter's posterior $p(\boldsymbol{\theta}|\mathcal{D})$ for complex models such as deep neural networks is intractable and variational approximations are used [16]. The posterior in (3) can be replaced by a simpler distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ with variational parameters $\boldsymbol{\omega}$. In this work, Monte-Carlo dropout [16] is used as a framework to perform stochastic variational inference. The relation between the true and approximate posteriors is given by

$$\boldsymbol{\omega} = \boldsymbol{\theta} \odot \mathbf{M}_D, \tag{4}$$

where $\mathbf{M}_D$ is a dropout mask that randomly sets the model weights to zero.

The training procedure can then be formulated as the minimization of the Kullback-Leibler divergence D_{KL} between the true posterior and the approximation. This is equivalent to minimizing the negative variational lower bound [16]:

$$\mathscr{L}_{\mathrm{VI}}(\boldsymbol{\omega}) = \int q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \log p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega} - D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \| p(\boldsymbol{\omega})), \tag{5}$$

where $\mathbf{X}, \mathbf{Y}$ represent the inputs and outputs of the model, respectively, and $p(\boldsymbol{\omega})$ is the prior for the variational parameters $\boldsymbol{\omega}$. The expectation in the first part of (5) is typically approximated using Monte-Carlo integration [16]. In this work, it is approximated using one sample from the variational distribution. Therefore, the optimization objective becomes

$$\mathcal{L}_{\mathrm{MCD}}(\boldsymbol{\omega}) = \sum_{i=0}^{N-1} \mathcal{L}(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\omega}) + \mathcal{R}(\boldsymbol{\omega}), \tag{6}$$

where i is an index of the training example and N is the total number of samples in the training set. $\mathcal{R}$ is a regularization term that depends on the form of a prior distribution over the parameters of the model. In this case, the prior is a normal distribution corresponding to L_2 weight decay. The loss function chosen for this work is binary cross-entropy and it is summed over the aleatoric samples:

$$\mathcal{L}(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}) = \sum_{i=0}^{N-1} \sum_{j=0}^{N_A-1} \mathcal{L}_{\mathrm{BCE}}\left(\mathbf{y}_{ij}|\mathbf{x}_i, \boldsymbol{\omega}\right), \tag{7}$$

where N_A is a number of aleatoric samples.

The training scheme described above does not take into account class imbalance. In this work, a straightforward oversampling scheme based on class frequencies statistics is used and it is described in the next section.

3.2. Oversampling

One way to handle class imbalance is to perform oversampling of the underrepresented classes. Here, three-stage sampling is performed:

1. Positive samples are selected with π^+ probability and negative samples are selected with $1 - \pi^+$ probability.
2. An image of the selected class is sampled with the probability p_i^{image} proportional to the logarithm of the pixel count of the given class, that is,

$$p_i^{\mathrm{image}} = \frac{\log \max(N_i^{\mathrm{image}}, 1)}{\sum_j \log \max(N_j^{\mathrm{image}}, 1)}, \tag{8}$$

where N_i^{image} is the number of positive pixels for the class of interest in the image with index i.

3. The final step is to select an image patch containing pixels of the class of interest. In order to select such a patch, we follow a scheme similar to the previous stage. The image is divided into a set of overlapping patches and the patch is selected with probability

$$p_i^{\mathrm{patch}} = \frac{\log \max(M_i^{\mathrm{patch}}, 1)}{\sum_j \log \max(M_j^{\mathrm{patch}}, 1)}, \tag{9}$$

where M_i^{patch} is the number of positive pixels for the class of interest in the patch with index i.

The log scale here is used in order to increase the diversity of chosen

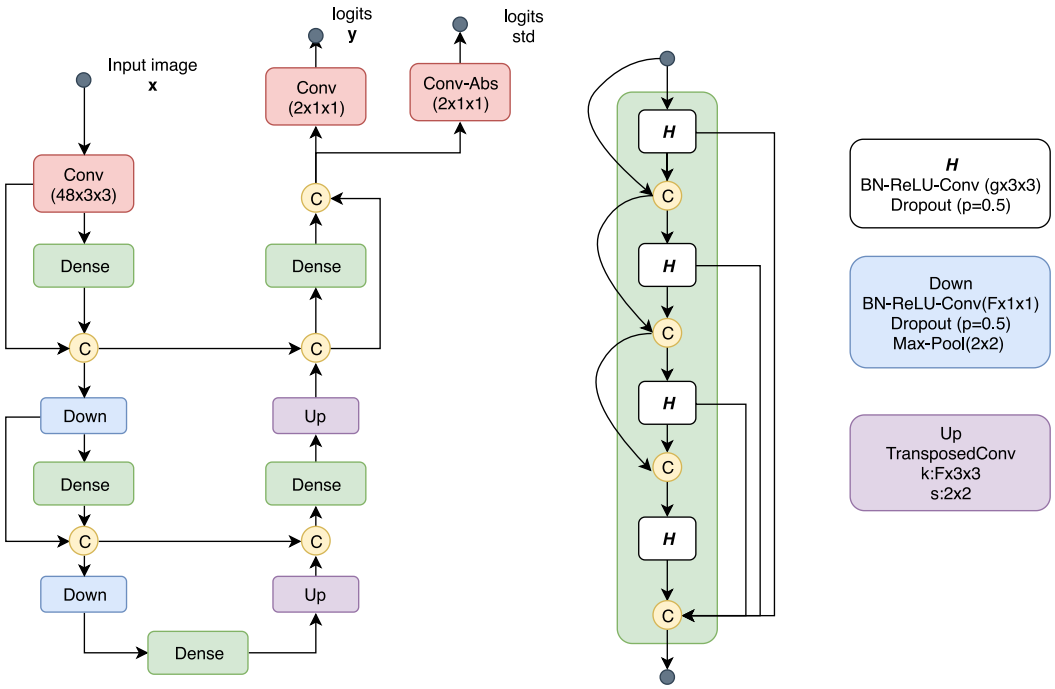


Fig. 4. The Dense-FCN architecture: *Dense* stands for a dense convolutional block; *C* is a tensor concatenation; *H* is a block consisting of batch normalization (BN), rectified linear unit (ReLU) and a convolutional layer with growth rate *g*; *Down* is a transition-down block with *F* output feature maps; *Up* is a transition up with *F* output feature maps and 2×2 stride; *logits std* denotes standard deviations of logits.

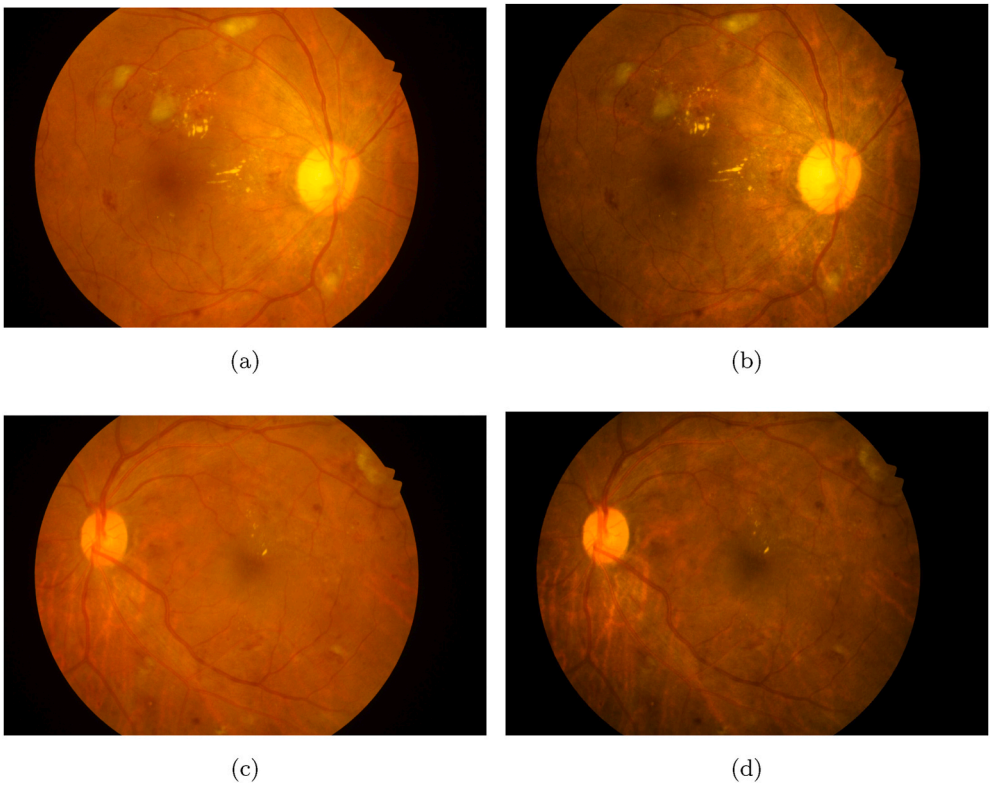


Fig. 5. Two examples of the original (left column) and enhanced (right column) images.

samples. π^+ is a tunable hyperparameter and should be chosen depending on the class imbalance in a particular case. In this work, $\pi^+ = 0.5$ is used as experimentally it has been found that this value provides the best results.

3.3. Architecture

The architecture utilized in this work is a Dense-FCN [17]. It has been shown that Dense-FCNs have less parameters and may outperform other fully-convolutional network (FCN) architectures in a variety of different segmentation tasks [17]. Here we adapt the Dense-FCN architecture for the lesion segmentation task.

The main building block of Dense-FCNs is a dense convolutional block (DCB) where the input of each layer is a concatenation of the outputs of the previous layers. The block consists of repeating batch normalization (BN), rectified linear unit (ReLU), convolution and dropout $p = 0.5$ layers resulting in g feature maps (growth rate).

The main concept of Dense-FCNs is similar to other encoder-decoder architectures in the sense that the input is first compressed to a hidden representation by the downsampling part. Thereafter the segmentation masks are recovered by an upsampling part. The downsampling part consists of DCBs and downsampling transitions with skip connections to the upsampling part. The upsampling part consists of DCBs and upsampling transitions. An example of two blocks in downsampling and upsampling paths of a Dense-FCN is shown in Fig. 4.

The total number of trainable parameters is 9319778. The architectural parameters used are are as follows:

- The growth rate for all DCBs: $g = 16$.
- The downsampling path consists of DCBs with depths $D_{\text{down}} = [4, 5, 7, 10, 12, 15]$.
- The upsampling also consists of five DCBs with depths $D_{\text{up}} = [12, 10, 7, 5, 4]$.
- The first and last convolution layers are the same as in Fig. 4.

3.4. Image preprocessing

It was noticed in the experimental part of the work that simple preprocessing proposed in Ref. [18] improves the results. The preprocessing is implemented in two steps:

1. Luminosity enhancement employs luminance gain matrix G that is applied in the red-green-blue (RGB) color space:

$$\mathbf{x}' = [\mathbf{G} \odot \mathbf{r} \quad \mathbf{G} \odot \mathbf{g} \quad \mathbf{G} \odot \mathbf{b}], \tag{10}$$

$$\mathbf{G}_i = \frac{\mathbf{V}_i'}{\max\{\mathbf{r}_i, \mathbf{g}_i, \mathbf{b}_i\}}, \tag{11}$$

where $\mathbf{r}$, $\mathbf{g}$ and $\mathbf{b}$ are red, green and blue image channels respectively, $\mathbf{x}'$ is an enhanced image, and $\mathbf{V}_i'$ is an enhanced luminance value at pixel with index i. The enhanced luminance value is calculated by converting the image to hue-saturation-value (HSV) color space and enhancing the luminance $\mathbf{V}$ using gamma enhancement. Here, we choose $\Gamma = 1/2.2$ as in the original work [18].

2. Contrast enhancement is performed using Contrast Limited Adaptive Histogram Equalization [19] algorithm with the clip limit 0.1 and the grid size 8×8.

In order to reduce requirements for computing resources, the images were resized to the resolution of 2144×1440 pixels. Two examples of the original and enhanced images are presented in Fig. 5.

3.5. Training details

The Dense-FCN was trained for 100 epochs with 500 steps per epoch on random patches 224×224 with the batch size equal to 6. The patches were generated with the overlap 192×192. Data augmentation by vertical and horizontal mirroring was applied. The parameter values were empirically tuned based on initial experiments with the IDRiD dataset.

The weights were initialized using HeNormal [20]. In addition to dropout, L_2 regularization with the weight decay factor 10^{-4} was used. As the optimizer, Adadelta [21] with the learning rate $l = 1$ and the decay rate $\rho = 0.95$ was used. The learning rate was adjusted according to the following schedule:

1. if $0 <= \text{epoch} < 50$, $l = 1$;
2. if $50 <= \text{epoch} < 70$, $l = 0.1$;
3. if $70 <= \text{epoch} < 85$, $l = 0.01$;
4. if $85 <= \text{epoch} < 100$, $l = 0.001$.

4. Experiments and results

4.1. Evaluation protocol

In [5], many authors processed images in a patchwise manner during the validation stage. In this work, it was noticed that with Bayesian neural networks this can lead to checkerboard artifacts that have a negative impact on the segmentation performance. Therefore, in the inference stage images are not divided into patches but are processed as full images. It is also worth to note that full-resolution processing is much faster and it takes approximately 14 min to process an image with 50 epistemic and 100 aleatoric samples. The input and output images have the resolution of 2144×1440 pixels.

In order to evaluate the segmentation performance, the following classification metrics are used:

- Sensitivity (SE) is used to assess the ability of the model to discover lesions:

$$SE = \frac{TP}{TP + FN}, \tag{12}$$

where TP and FN are the amounts of true positive and false negative pixels, respectively.

- Positive predictive value (PPV) is used in addition to sensitivity but taking into account false positives FP:

$$PPV = \frac{TP}{TP + FP}. \tag{13}$$

- Specificity (SP) is used to assess to ability of the model to correctly segment healthy pixels:

$$SP = \frac{TN}{TN + FP}, \tag{14}$$

where TN is the amount of true negative pixels.

- Area under receiver-operating-characteristic curve (ROC-AUC) is an integral metric regardless of the thresholding value. ROC-AUC is calculated under the area of the curve plotted as a true positive rate against false positive rate by varying the threshold.
- Area under precision-recall curve (PR-AUC) is another integral metric regardless of the thresholding value. PR-AUC more realistically represents the segmentation performance in comparison to the area under receiver operating characteristic ROC-AUC.

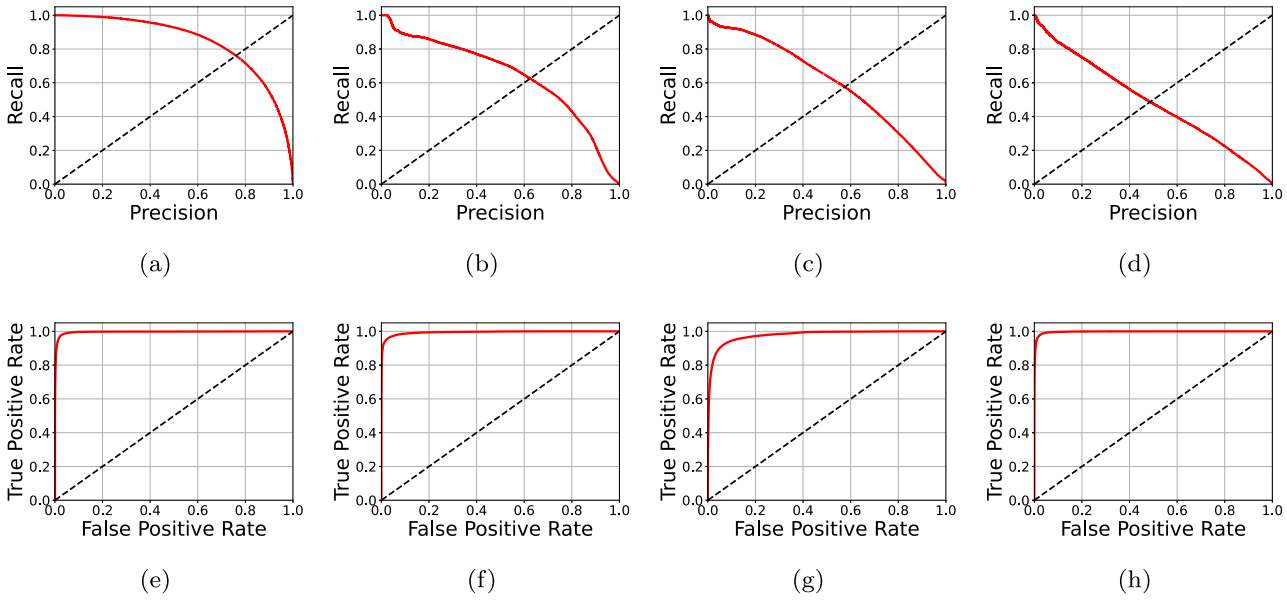


Fig. 6. Precision-recall and receiver operating characteristic curves for (a), (e) hard exudates; (b), (f) soft exudates; (c), (g) haemorrhages; (d), (h) microaneurysms.

- Expected calibration error (ECE) is used to assess a model's calibration [22]:

$$ECE = \mathbb{E}_{\widehat{p}}[|\mathbb{P}(\widehat{y} = y|\widehat{p} = \pi) - \pi|], \quad \pi \in [0, 1], \tag{15}$$

where $\widehat{p}$ is a confidence estimate of the predicted class $\widehat{y}$, y is a true label and π is a true probability.

Together with ECE, reliability diagrams are also presented. These reliability diagrams are graphs showing the expected accuracy against classification confidence, thereby representing calibration quality. In the case of perfect calibration, the graph is an identity function.

In the evaluation, sensitivity, specificity and positive predictive value are calculated by thresholding the output predictive mean with $T = 0.5$.

In the inference, the model parameters are sampled 100 times and the number of inferred aleatoric samples is $N_A = 100$. The final posterior predictive mean is calculated over all the predicted samples, and the aleatoric uncertainty U_A and epistemic uncertainty U_E of the outputs are calculated as in Ref. [23]:

$$U_A = \mathbb{E}_q\left[\mathbb{V}_{p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}[\mathbf{y}]\right], \tag{16}$$

$$U_E = \mathbb{V}_q\left[\mathbb{E}_{p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}[\mathbf{y}]\right], \tag{17}$$

$$U_T = U_A + U_E, \tag{18}$$

where $\mathbb{E}$ and $\mathbb{V}$ denote expectation and variance, respectively, and U_T is the total predictive uncertainty.

Apart from characterizing the total uncertainty, it is also important to evaluate the meaningfulness of the produced uncertainty maps. This is a more challenging task since only point estimates of ground truth labels are available. However, it is reasonable to assume that incorrectly segmented areas must have higher uncertainties. Mobiny et al. [24] proposed to use the uncertainty as a tool predict incorrect classification results by thresholding the output uncertainties. Camarasa et al. [25] analyzed different uncertainty measures for medical image segmentation and concluded that the averaged variance and averaged entropy perform equally well and are better than other metrics. In this work, the standard deviation is used. We follow the same approach and use the following:

1. Area under uncertainty precision-recall curve (PR-AUC) is used an integral metric to assess the quality of uncertainty estimates.
2. Uncertainty sensitivity (U-SE) is used to assess the ability of the uncertainty estimates to discover misclassifications.
3. Uncertainty specificity (U-SP) is used to assess the ability of the uncertainty estimates to correctly classify misclassifications.
4. Uncertainty expected calibration error (U-ECE) is also used to validate the uncertainty calibration.

U-SE and U-SP are calculated using the threshold which is half of the maximum uncertainty value.

To summarize, the extended validation approach consists of the analysis of the produced segmentation masks as well as comparison of the produced uncertainties and the misclassification maps.

4.2. Evaluation of segmentation results

The precision-recall (PR) and receiver operating characteristic (ROC) curves are shown in Fig. 6. It is clear that the ROC curves demonstrate close-to-optimal classification results due to large class imbalance. On the other hand, the PR curves represent the classification performance more realistically. The corresponding performance metrics are given in Table 1. Based on the figures and the table, it is clear that the easiest task is to segment the hard exudates, whereas the most difficult one is the

Table 1
Evaluation results of the baseline training scheme. The abbreviations of the evaluation metrics are explained in the text.

Label	PR-AUC	ROC-AUC	Sensitivity	PPV	Specificity	ECE
Hard exudates	0.842	0.995	0.767	0.753	0.997	0.090
Soft exudates	0.641	0.993	0.639	0.611	0.999	0.145
Haemorrhages	0.593	0.977	0.464	0.670	0.997	0.066
Microaneurysms	0.484	0.997	0.434	0.531	0.999	0.116

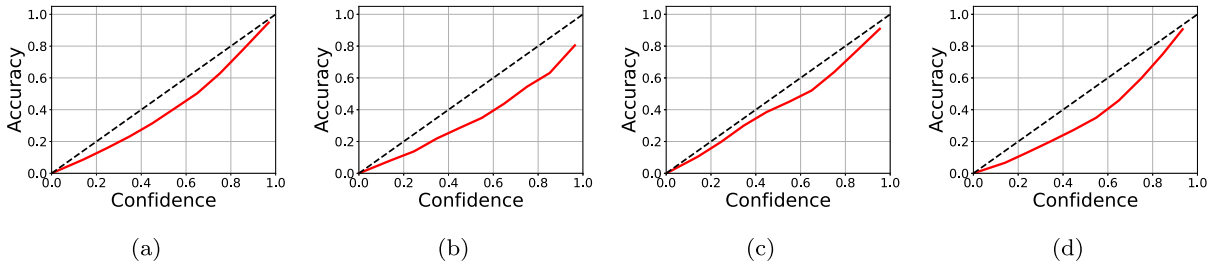


Fig. 7. Reliability diagram for (a) hard exudates; (b) soft exudates; (c) haemorrhages; (d) microaneurysms.

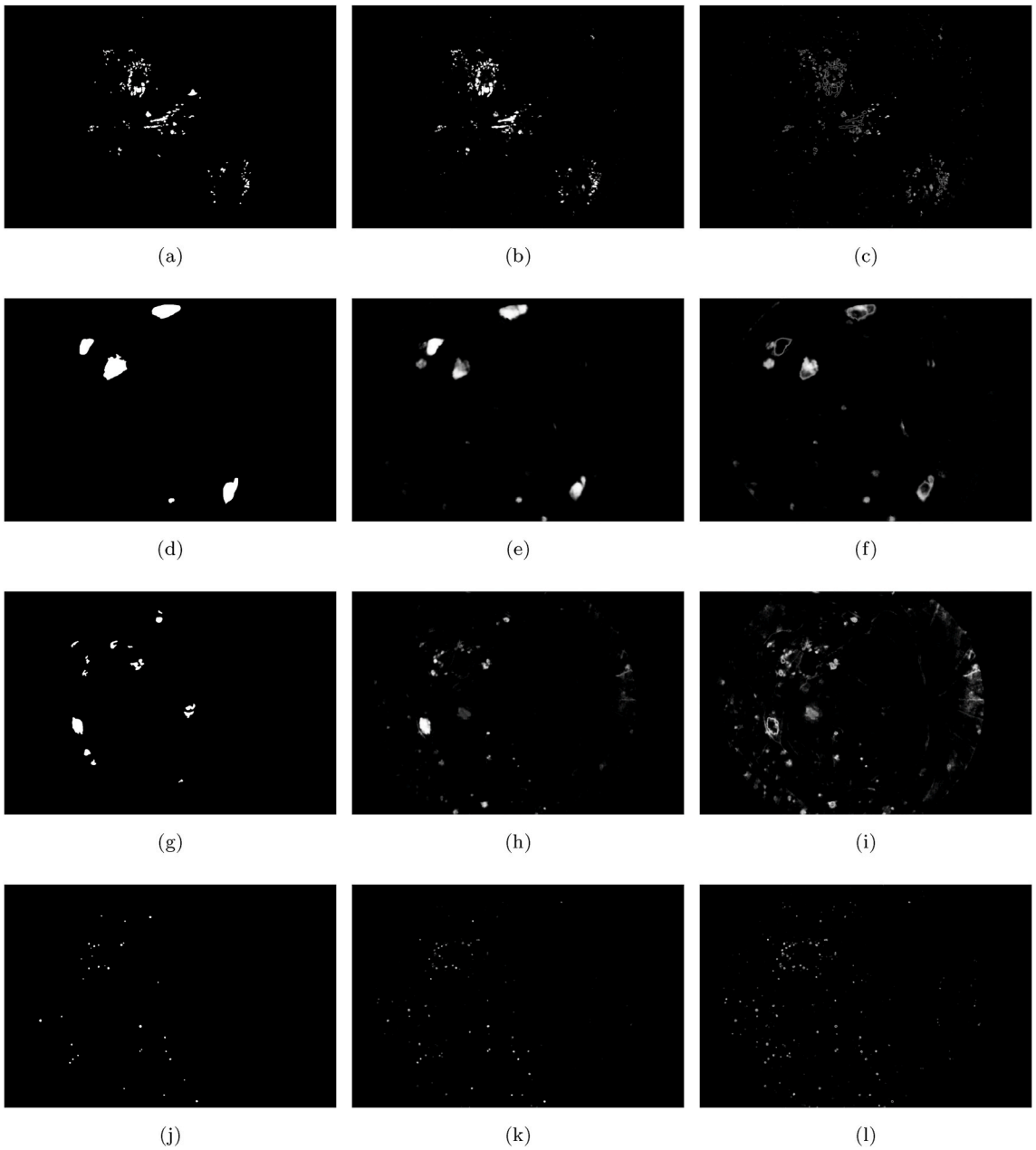


Fig. 8. Visualizations of inference results for input image 5b for lesions: (a), (b), (c) hard exudates; (d), (e), (f) soft exudates; (g), (h), (i) haemorrhages; (j), (k), (l) microaneurysms. The first column shows the ground truth masks, the second shows the mean inferred probabilities and the third shows epistemic uncertainty masks (standard deviations of probabilities).

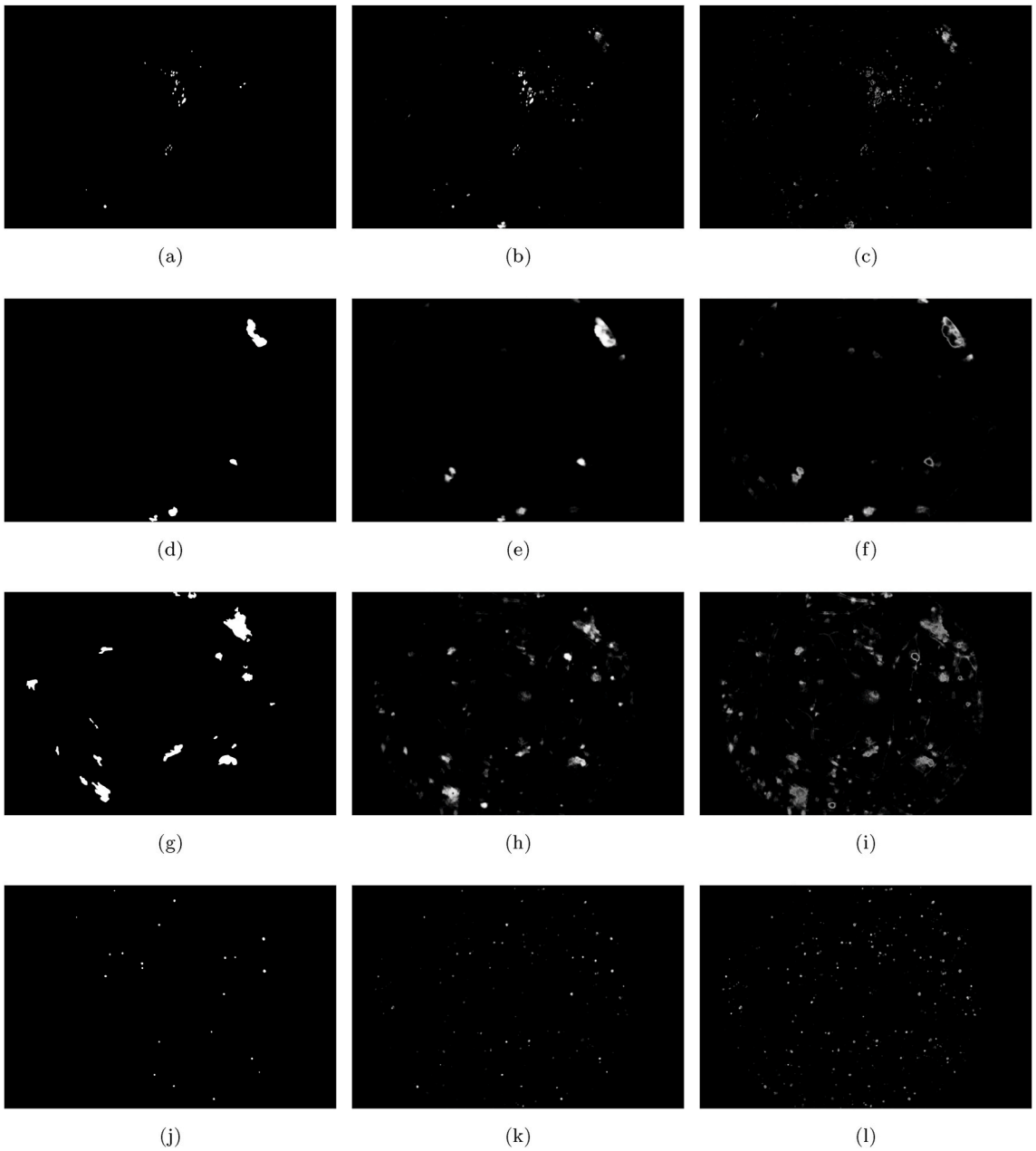


Fig. 9. Visualizations of inference results for input image 5d for lesions: (a), (b), (c) hard exudates; (d), (e), (f) soft exudates; (g), (h), (i) haemorrhages; (j), (j), (l) microaneurysms. The first column shows the ground truth masks, the second shows the mean inferred probabilities and the third shows epistemic uncertainty masks (standard deviations of probabilities).

Table 2
Evaluation results for the estimated uncertainty maps. The abbreviations of the evaluation metrics are explained in the text.

Label	U-PR-AUC	U-SE	U-PPV	U-SP	U-ECE
Hard exudates	0.336	0.031	0.566	0.999	0.104
Soft exudates	0.257	0.113	0.388	0.999	0.195
Haemorrhages	0.243	0.029	0.302	0.999	0.303
Microaneurysms	0.257	0.045	0.332	0.999	0.237

segmentation of microaneurysms. Low sensitivies are a common problem for the DR lesion segmentation task [5]. This can be explained by the relatively low contrast and size of lesions. Apart from the analysis of true positive classifications, it is also essential to have classifiers with high specificity. From Table 1 it is possible to see that specificities are very high for all types of lesions being close to one. Nevertheless, it can be easily achieved due to the class imbalance. PPVs, on the other hand, give more insights into the problem of false positive classifications comparing them to true positives. It is easy to notice that in the worst case scenario for microaneurysms there are almost as many falsely classified pixels of healthy tissues as correctly discovered pixels of microaneurysms. This fact gives additional motivation for analyzing the uncertainties.

The reliability diagrams are given in Fig. 7. It can be seen that the trained models are miscalibrated and the one for haemorrhages represents the best result. Guo et al. [22] have shown that deep neural networks are typically poorly calibrated and the authors proposed methods decreasing the degree of miscalibration. Guo et al. claimed that the ECE of approximately $0.01 - 0.02$ can be achieved for standard classification benchmark datasets and Dense architectures. In this work, no methods

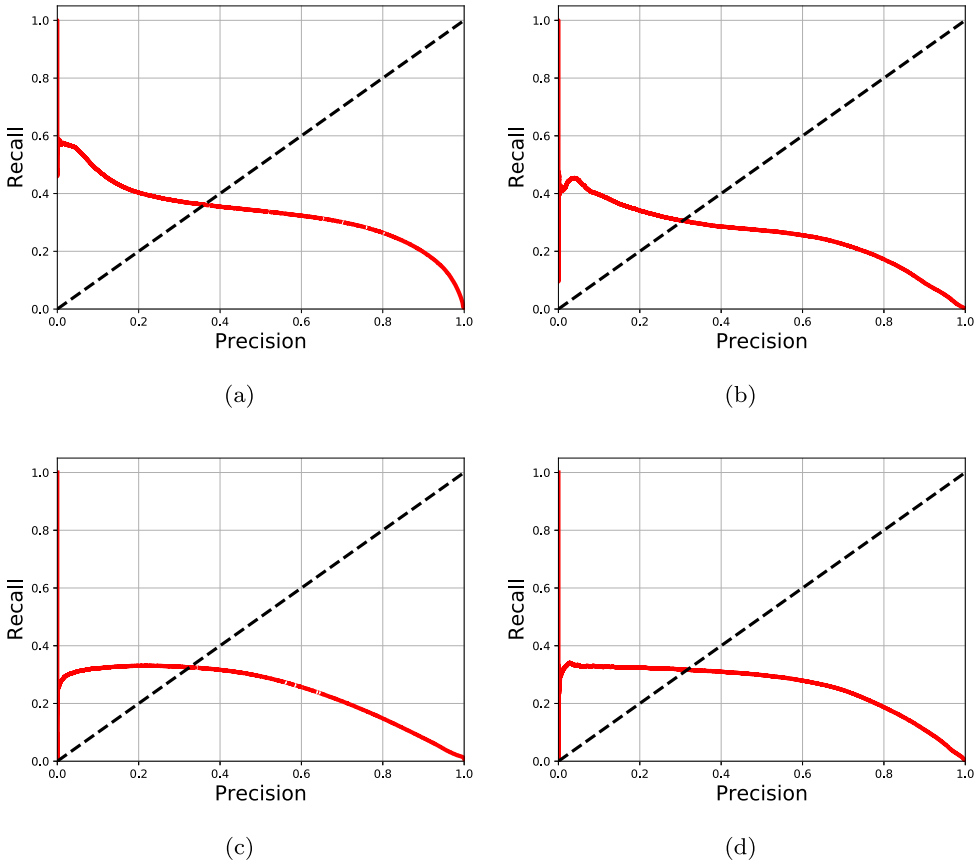


Fig. 10. Uncertainty precision-recall curves for (a) hard exudates; (b) soft exudates; (c) haemorrhages; (d) microaneurysms.

for improving the calibration were used and the reliability is assessed for the baseline model.

The segmentation results for two example images from the test set (shown in Fig. 5) are illustrated in Fig. 8 and Fig. 9. From the images, it is possible to observe visual similarities between the ground truth and mean inferred probability maps. Higher uncertainties are concentrated around the areas with high predicted confidence and false positive segmented pixels. A more detailed discussion about the inference results and the estimated uncertainties is given in the next section.

4.3. *Uncertainty quantification*

The PR curves and reliability diagrams are shown in Fig. 6 and the evaluation metrics are given in Table 2. From the results, it is clear that normalized uncertainties are not efficient predictors of misclassifications and have low sensitivities. It is worth to note that the evaluation procedure is straightforward and considers only soft uncertainties against hard misclassifications. Nevertheless, the uncertainties are not necessarily high only near the misclassification areas, but also near the areas of relatively low confidence as shown below. This can also explain the uncertainty miscalibrations. The uncertainty PR curves are given in Fig. 10 and the uncertainty reliability diagrams are presented in Fig. 11. From the reliability diagrams it is clear that the uncertainties are mostly underestimated, since the growing confidence values stop matching with the increasing accuracy values.

Inference results for hard exudates of the magnified example image are shown in Fig. 12. It is clear that the misclassifications and epistemic uncertainties are mostly concentrated around the edges of the lesions. This can be explained by unclear boundaries of the lesions. The aleatoric uncertainties acting as a learned loss attenuation are also higher around the borders. The boundary uncertainties are a general pattern for segmentation models and can be observed within a wide variety of tasks. It is also possible to see small yellow lesions being incorrectly classified as background which highlights the problems of detecting small-scale lesions. It is worth noting that there is a soft exudate left to the hard exudates cluster and the model is certain for not classifying it as a hard exudate.

Inference results for soft exudates of the magnified example image are shown in Fig. 13. The high boundary uncertainties are presented in this case as well. Soft exudates typically have low contrast, no texture, unclear edges and can be easily confused with the background. It is possible to see false positive detections of soft exudates in the lower left part of the image which is slightly more yellow comparing to the other background pixels. The soft exudate in the lower right part of the image has uneven contrast and the low-contrast part of the lesion is incorrectly classified as the background. In both cases, the model yielded non-maximum mean confidence and the incorrectly classified pixels also have high uncertainties.

In Fig. 14, the inference results for the haemorrhages of the magnified example image are presented. The lesion is surrounded by blood vessels and a part of the macula is presented in the magnified input image. The part with blood vessels to the left is incorrectly classified as a haemorrhage. It is also possible to see the model's confusion about the part with the macula. Epistemic uncertainty is in general higher near the areas with similar colors highlighting the surrounding blood vessels and macula.Inference results for microaneurysms of the magnified example image are given in Fig. 15. Microaneurysms are the smallest of all lesions and the epistemic uncertainty is high over the whole area of lesions. On the other hand, the aleatoric uncertainties are still higher near the edges. Being small-scale lesions with no textures, microaneurysms are confused with any red small spots, which is visible on the epistemic uncertainty maps.

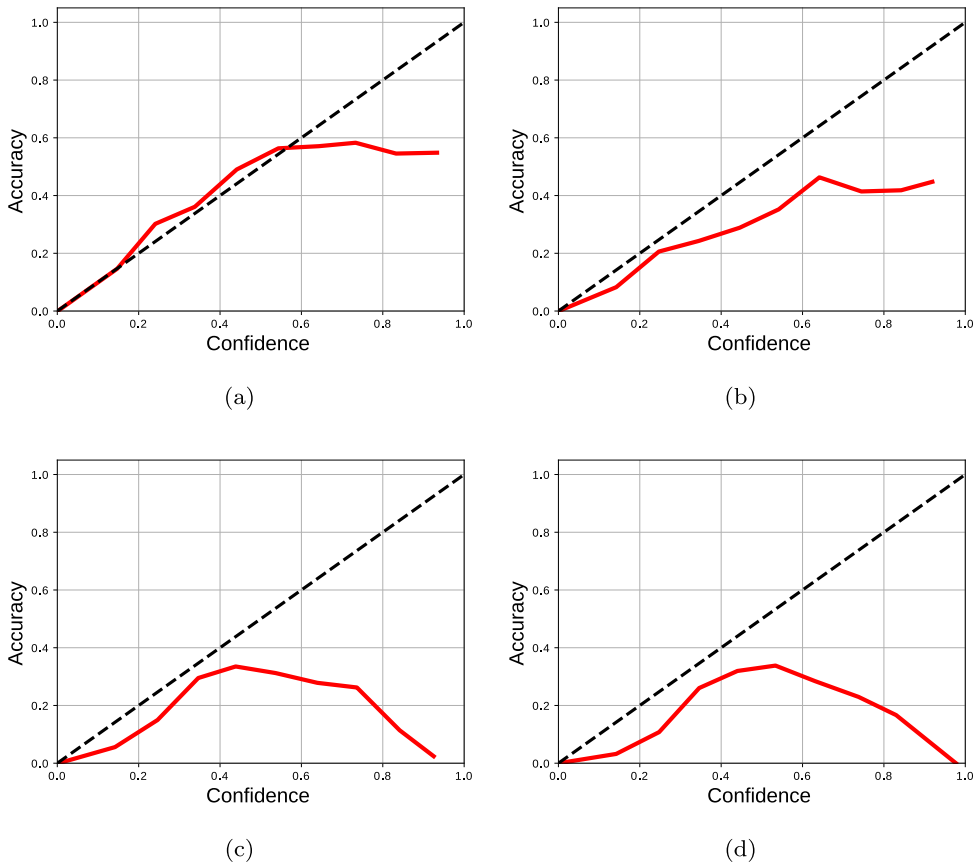


Fig. 11. Uncertainty reliability diagrams for (a) hard exudates; (b) soft exudates; (c) haemorrhages; (d) microaneurysms.

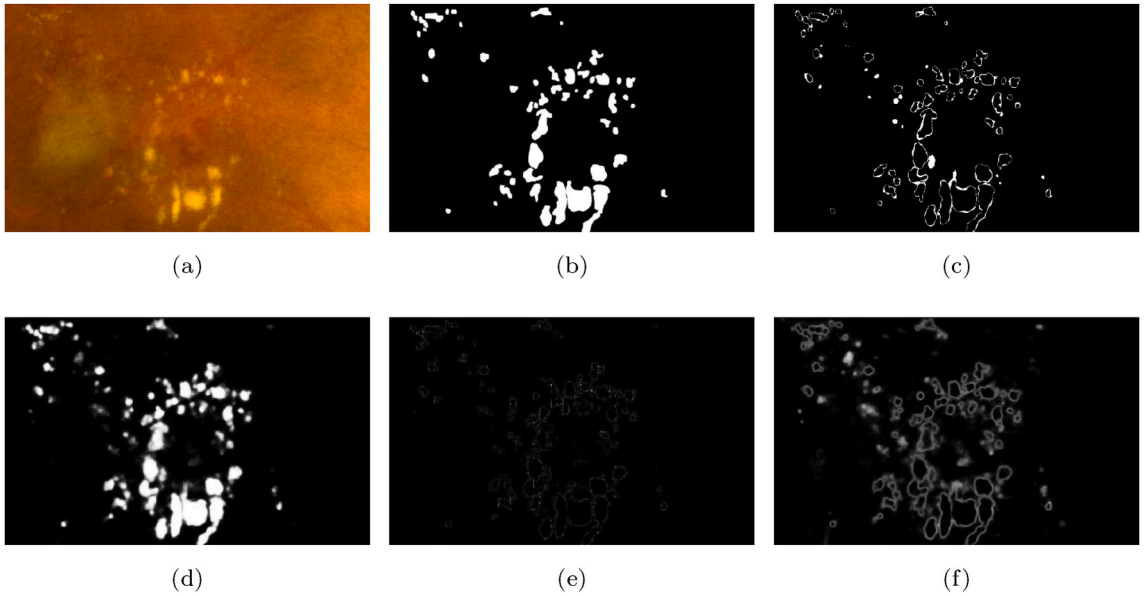


Fig. 12. Inference results for hard exudates with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

5. Discussion

The approach presented in this work shows classification performance comparable to previously reported methods [5]. The uncertainty maps can be used for the visual inspection and analysis of the performance. The estimated uncertainties and the produced confidence maps provide more information about the model's behaviour. Nevertheless, a few challenges remain and they are discussed in this section in addition to brief explanations of failed experiments.

One of the main issues in lesion detection is low sensitivity of the segmentation model. This problem is present in the related previous works [4,26] and also in this study. In medical image analysis and

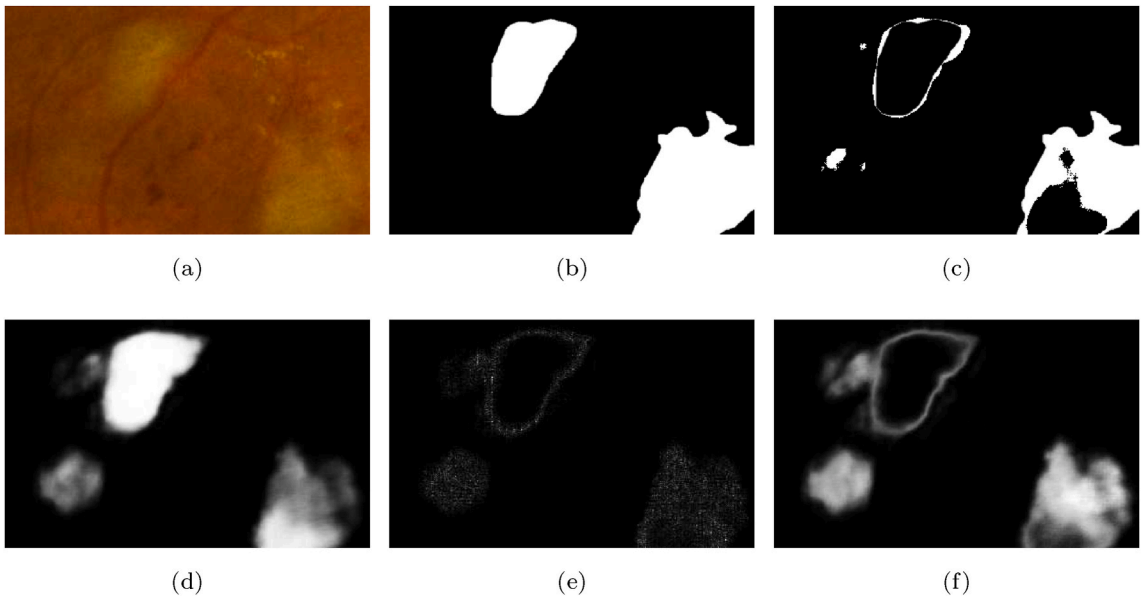


Fig. 13. Inference results for soft exudates with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

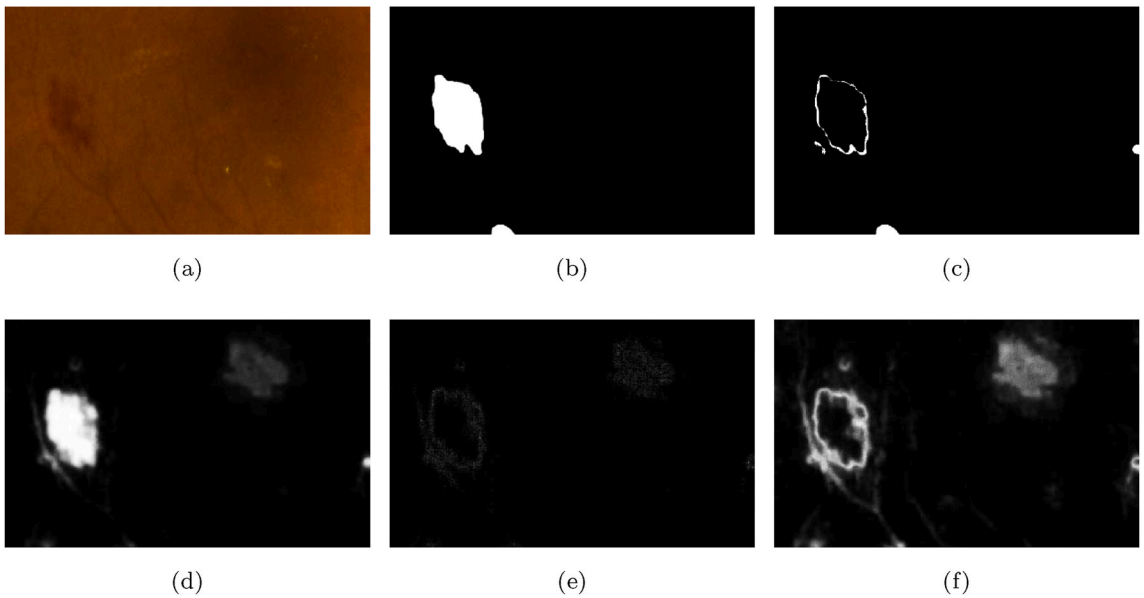


Fig. 14. Inference results for haemorrhages with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

segmentation, it is common to use custom heuristic loss functions [26] to improve sensitivity [27] or deal with lesion boundary issues [28]. We also experimented with other loss functions including focal loss [29], Tversky loss [27], generalized dice loss [28], and boundary loss [30]. Nevertheless, results outperforming the proposed baseline were not achieved. This negative outcome is likely due to omitting the tuning of loss functions' hyperparameters. These objectives are typically synthetic in the sense that they are formulated already in the form of loss functions and not as log-likelihoods. This means that they are not derived from specific distributions encoding the information about class imbalance. On the other hand, binary cross-entropy is derived as a negative logarithm of the Bernoulli likelihood. To study the issue with low sensitivity, more focused research is required to evaluate modern loss functions for medical image segmentation in the context of Bayesian deep learning and model calibration.

In this work, a straightforward scheme based on label statistics is used to balance the lesion and background data. A potentially more efficient approach would be to use Bayesian active learning [31] where uncertainty-based acquisition functions are used to select the training samples. Typically, these methods do not work well with unbalanced data which can be another topic for the future research.

Model and uncertainty calibration metrics are also subjects for further improvements. Apart from the classical calibration methods described in Ref. [22], alternative ways of improving the calibration exist. Thulasidasan et al. [32] proposed to use mix-up augmentation to improve the model calibration. Seo et al. [33] proposed single-shot calibration by regularizing the model with the uncertainty of the outputs. Laves et al. [34] considered the uncertainty calibration in the context of deep Bayesian regression and discovered that the predicted uncertainties are typically underestimated. The problem was solved

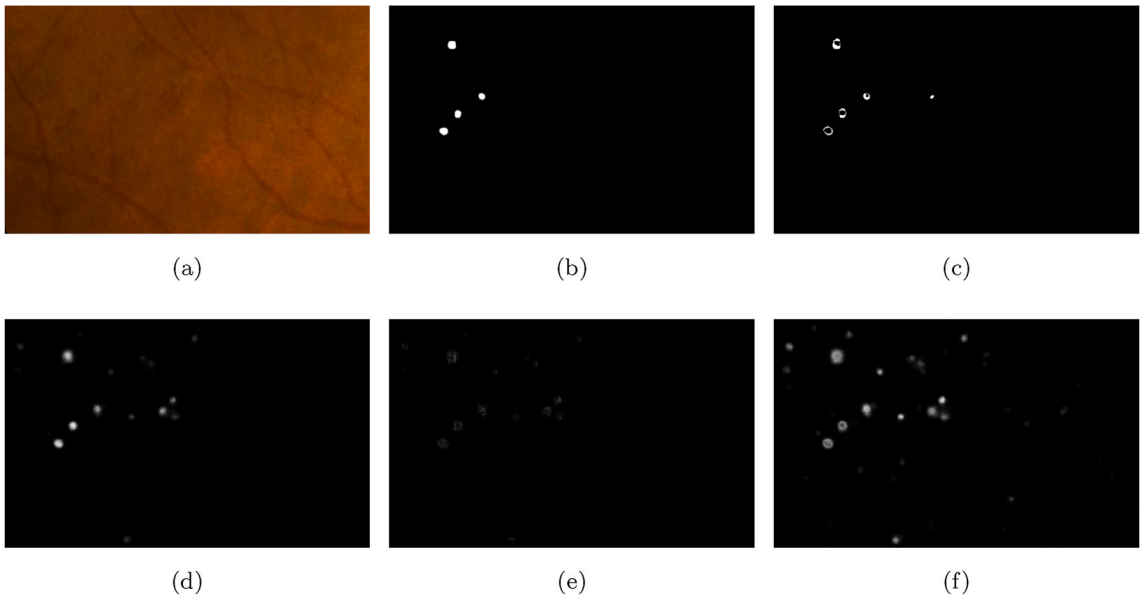


Fig. 15. Inference results for microaneurysms with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

using simple temperature scaling of aleatoric and epistemic uncertainty. During the development of this work, experiments with the uncertainty calibration using Platt scaling and isotonic regression were conducted. However, no improvements over the baseline were found. It is likely that a more systematic approach aiming to solve both calibration problems is required.

6. Conclusion

In this paper, a Bayesian baseline for the diabetic retinopathy lesion segmentation, allowing the analysis of segmentation distributions, model calibration and prediction uncertainties, is proposed. Also an extended validation approach consisting of the analysis of segmentation performance and the ability of uncertainty estimates to detect false classifications is provided. The presented results from the uncertainty quantification experiments show that the estimates are qualitatively similar to misclassification maps and can be used to assess issues in the lesion segmentation. Overall, the main challenges of the deep probabilistic model are the small-scale lesions, areas with low contrast and unclear boundaries. The color information is also essential for successful segmentation and healthy tissues can be confused with lesions when being of a similar color. Further research and development is required to make the predicted lesion segmentation uncertainties suitable for numeric quantification.

Declaration of competing interest

None of the authors have any conflict of interest.

Acknowledgement

This work has been supported by LUT Doctoral School. They were not involved in the study design, data collection and analysis, decision to publish, or preparation of this work. The computational resources for this work were provided by CSC – IT Center for Science, Finland. The authors wish to thank the authors of the open-access data utilized in this work.

References

[1] E. Reichel, D. Salz, Diabetic Retinopathy Screening, Springer International Publishing, Cham, 2015, pp. 25–38.
[2] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, M. Prunotto, Deep learning algorithm predicts diabetic retinopathy progression in individual patients, NPJ Dig. Med. 2 (1) (2019) 1–9.
[3] J. de la Torre, A. Valls, D. Puig, A deep learning interpretable classifier for diabetic retinopathy disease grading, Neurocomputing 396 (2020) 465–476, https://doi.org/10.1016/j.neucom.2018.07.102.
[4] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, H. Fu, Applications of deep learning in fundus images: a review, Med. Image Anal. 69 (2021) 101971, https://doi.org/10.1016/j.media.2021.101971.
[5] P. Porwal, S. Pachade, M. Kokare, G. Deshmukh, J. Son, W. Bae, L. Liu, J. Wang, X. Liu, L. Gao, et al., Idrid: diabetic retinopathy–segmentation and grading challenge, Med. Image Anal. 59 (2020) 101561.
[6] O. Ronneberger, P. Fischer, T. Brox, U-net, Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.
[7] S. Jegou, M. Drozdzal, D. Vázquez, A. Romero, Y. Bengio, The One Hundred Layers Tiramisu: Fully Convolutional Densenets for Semantic Segmentation, 2017, pp. 1175–1183, https://doi.org/10.1109/CVPRW.2017.156.
[8] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[9] S. Xie, Z. Tu, Holistically-nested edge detection, in: 2015 IEEE International Conference on Computer Vision, ICCV), 2015, pp. 1395–1403, https://doi.org/10.1109/ICCV.2015.164.
[10] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, The importance of skip connections in biomedical image segmentation, in: G. Carneiro, D. Mateus, L. Peter, A. Bradley, J.M.R.S. Tavares, V. Belagiannis, J.P. Papa, J.C. Nascimento, M. Loog, Z. Lu, J.S. Cardoso, J. Cornebise (Eds.), Deep Learning and Data Labeling for Medical Applications, Springer International Publishing, Cham, 2016, pp. 179–187.
[11] S. Guo, T. Li, H. Kang, N. Li, Y. Zhang, K. Wang, L-seg, An end-to-end unified framework for multi-lesion segmentation of fundus images, Neurocomputing 349 (2019) 52–63.
[12] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong, S. Cui, Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 597–600.
[13] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, S. Wahl, Leveraging uncertainty information from deep neural networks for disease detection, Sci. Rep. 7.
[14] A. Filos, S. Farquhar, A. N. Gomez, T. G. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, Y. Gal, A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks, arXiv preprint arXiv:1912.10481.
[15] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems, 2017, pp. 5574–5584.
[16] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.

[17] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE, 2017, pp. 1175–1183.
[18] M. Zhou, K. Jin, S. Wang, J. Ye, D. Qian, Color retinal image enhancement based on luminosity and contrast adjustment, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 65 (3) (2018) 521–527, https://doi.org/10.1109/TBME.2017.2700627.
[19] K. Zuiderveld, in: P.S. Heckbert (Ed.), Graphics Gems IV, Academic Press Professional, Inc., San Diego, CA, USA, 1994, pp. 474–485. Ch. Contrast Limited Adaptive Histogram Equalization.
[20] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV), 2015.
[21] M.D. Zeiler, ADADELTA: an adaptive learning rate method, Tech. rep., arXiv: 1212.5701 (Dec. 2012). URL, http://arxiv.org/abs/1212.5701.
[22] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.
[23] S. Hu, D. Worrall, S. Knegt, B. Veeling, H. Huisman, M. Welling, Supervised Uncertainty Quantification for Segmentation with Multiple Annotations, 01949, 1907. arXiv preprint arXiv.
[24] A. Mobiny, H. Nguyen, S. Moulik, N. Garg, C. Wu, Dropconnect Is Effective in Modeling Uncertainty of Bayesian Deep Networks, ArXiv abs/1906.04569.
[25] R. Camarasa, D. Bos, J. Hendrikse, P. Nederkoorn, E. Kooi, A. van der Lugt, M. de Bruijne, Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, Springer, 2020, pp. 32–41.
[26] M. Jun, Segmentation Loss Odyssey, arXiv preprint arXiv:2005.13449.
[27] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention u-net for lesion segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 683–687.
[28] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M.J. Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 240–248.
[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[30] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, I.B. Ayed, Boundary loss for highly unbalanced segmentation, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2019, pp. 285–296.
[31] A. Kirsch, J. van Amersfoort, Y. Gal, in: Batchbald: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning, NeurIPS, 2019.
[32] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, S. Michalak, On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks, arXiv preprint arXiv:1905.11001.
[33] S. Seo, P.H. Seo, B. Han, Learning for single-shot confidence calibration in deep neural networks through stochastic inferences, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9022–9030.
[34] M.-H. Laves, S. Ihler, J.F. Fast, L.A. Kahrs, T. Ortmaier, Well-calibrated regression uncertainty in medical imaging with deep learning, in: Medical Imaging with Deep Learning, PMLR, 2020, pp. 393–412.