Azat Garifullin

DEEP BAYESIAN APPROACH TO EYE FUNDUS IMAGE SEGMENTATION

ACTA UNIVERSITATIS LAPPEENRANTAENSIS 1003

Azat Garifullin

DEEP BAYESIAN APPROACH TO EYE FUNDUS IMAGE SEGMENTATION

Dissertation for the degree of Doctor of Science (Technology) to be presented with due permission for public examination and criticism in the Auditorium 1314 at Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland on the 9th of December, 2021, at noon.

Acta Universitatis
Lappeenrantaensis 1003

Supervisor Professor Lasse Lensu
LUT School of Engineering Science
Lappeenranta-Lahti University of Technology LUT
Finland

Reviewers Assistant Professor Juho Kannala
Department of Computer Science
Aalto University
Finland

Professor Emanuele Trucco
School of Science and Engineering
University of Dundee
Scotland, UK

Opponents Professor Jussi Tohka
Faculty of Health Sciences
University of Eastern Finland
Finland

Professor Emanuele Trucco
School of Science and Engineering
University of Dundee
Scotland, UK

ISBN 978-952-335-761-7
ISBN 978-952-335-762-4 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Lappeenranta-Lahti University of Technology LUT
LUT University Press 2021

Abstract

Azat Garifullin
Deep Bayesian Approach to Eye Fundus Image Segmentation
Lappeenranta 2021
64 pages
Acta Universitatis Lappeenrantaensis 1003
Diss. Lappeenranta-Lahti University of Technology LUT
ISBN 978-952-335-761-7
ISBN 978-952-335-762-4 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491

Eye diseases cause different retinal abnormalities that can be detected and diagnosed by examining eye fundus images. Due to the rapidly growing amount of data, there is a need for methods that are able to produce meaningful image representations and analysis results helping medical doctors to make correct diagnoses. Recent advances in deep learning have enabled very promising approaches for solving a variety of tasks related to automatic fundus image analysis. However, there is growing concern about the reliability of these methods and possible issues exist regarding their utilization in risk-sensitive scenarios.

This study extends the current research by studying fundus image segmentation from a deep Bayesian perspective that permits model parameters and their outputs to be treated as random variables. The treatment makes it possible to estimate how uncertain the model is about its predictions. The study focuses on subproblems including the segmentation of the retinal vasculature, optic disc, macula and diabetic retinopathy lesions. Considering the probabilistic nature of the chosen methods, validation procedures need to be augmented in order to evaluate not only the segmentation results but also the estimated uncertainties.

The experimental results show that the proposed Bayesian baselines for fundus image segmentation yield a performance that is comparable to the existing state-of-the-art approaches. The produced uncertainty estimates provide meaningful information about possible problems during the inference. However, the uncertainty validation results suggest that predicting misclassifications using uncertainty in a straightforward manner is limited. The results of additional experiments using weight averaging techniques and spectral image data are provided. This work also discusses the problems encountered when applying Bayesian methods to fundus image segmentation.

Keywords: Bayesian deep learning, fundus imaging, image segmentation, diabetic retinopathy, lesion segmentation, vasculature, optic disc, macula

Acknowledgements

This work was carried out at Computer Vision and Pattern Recognition laboratory at Lappeenranta–Lahti University of Technology LUT, Finland, between 2018 and 2021. I am grateful to LUT Doctoral School for funding the research.

I would like to express my deepest gratitude to my supervisor Prof. Lasse Lensu for the guidance during this research. I would like to thank Prof. Hannu Uusitalo for valuable advices and discussion especially during the journal publication process.

My sincere thanks are due to Dr. Pauli Fält, Prof. Markku Hauta-Kasari, Prof. Hannu Uusitalo et al. for the development of spectral retinal imaging and the collection of spectral retinal images for the DiaRetDB2 dataset utilized in this work.

I thank my honored pre-examiners and opponents Prof. Juho Kannala, Prof. Emanuele Trucco and Prof. Jussi Tohka for their valuable participation in the dissertation process.

To the friends and family.

Azat Garifullin
December 2021
Lappeenranta, Finland

a	a scalar
$\mathbf{a}$	a vector
$\mathbf{A}$	a matrix
$\mathcal{D}$	a dataset
N	the number of samples in a dataset
$\mathbf{x}$	an input image
$\mathbf{p}$	a ground truth segmentation map
$\boldsymbol{\theta}$	parameters of a model
$\hat{\mathbf{p}}$	estimated segmentation map
sigmoid	sigmoid activation function
y	a logit
$\exp(a)$	exponential function applied to a scalar
$\arg\max_x g(x)$	argument x at which function g takes its maximum
$p(\mathbf{a})$	probability density function of $\mathbf{a}$
$p(\mathbf{a} \mid \mathbf{b})$	conditional probability density function of $\mathbf{a}$ given $\mathbf{b}$
$\prod_{i=m}^{n}$	product over i from m to n
$\int g(x)dx$	integral of function g with respect to x
$\log x$	natural logarithm of x
$\sum_{i=m}^{n}$	sum over i from m to n
$\arg\min_x g(x)$	argument x at which function g takes its minimum
$\mathcal{L}$	a loss function
$\mathcal{R}$	a regularization term
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
D	the number of parameters of a model
$\hat{\mathbf{y}}$	an estimated vector of logits
$\boldsymbol{\sigma}$	estimated standard deviations of logits
$\mathbf{A} \odot \mathbf{B}$	element-wise (Hadamard) product of $\mathbf{A}$ and $\mathbf{B}$
$a \sim p_a$	a random variable a distributed according to p_a
N_A	the number of aleatoric samples
$\mathbf{I}$	identity matrix
$q_{\boldsymbol{\theta}}$	approximated posterior distribution of $\boldsymbol{\theta}$
$\boldsymbol{\omega}$	variational parameters
$\mathbf{M}_D$	dropout mask

D_{KL}	Kullback-Leibler divergence
$\mathcal{L}_{\mathrm{VI}}$	variational minimization objective
$\mathcal{R}_{\mathrm{VI}}$	variational regularization term
N_E	the number of epistemic samples
$\boldsymbol{\theta}_{\mathrm{SWA}}$	parameters of a model estimated using stochastic weight averaging
$\boldsymbol{\Sigma}_{\mathrm{SWAG}}$	a covariance matrix of parameters of a model estimated using stochastic weight averaging Gaussian
$\mathbb{V}_p$	a variance under distribution p
$\mathbb{E}_p$	an expectation under distribution p
U_A	aleatoric uncertainty
U_E	epistemic uncertainty
U_T	total uncertainty
SE	sensitivity
TP	true positives
FN	false negatives
PPV	positive predictive value
FP	false positives
SP	specificity
TN	true negatives
IoU	intersection over union
$A \cup B$	union of sets A and B
$A \cap B$	intersection of sets A and B
$F1$	F1 score
ECE	expected calibration error

AUC	area under the curve
AV	artery-vein
AVR	arteriole-to-venule ratio
BCE	binary cross-entropy
BN	batch normalization
CAM	class activation map
CDR	cup-to-disc ratio
DCB	dense convolutional block
Dense-FCN	dense fully-convolutional network
DiaRetDB1	DiaRetDB1 diabetic retinopathy database
DiaRetDB2	DiaRetDB2 diabetic retinopathy database

DRIONS-DB	digital retinal images for optic nerve segmentation database
DRIVE	digital retinal images for vessel extraction dataset
ECE	expected calibration error
FCN	fully-convolutional network
FOV	field-of-view
HRF	high resolution fundus image database
IDRiD	Indian diabetic retinopathy image dataset
IoU	intersection over union
MC-Dropout	Monte-Carlo dropout
MCMC	Monte-Carlo Markov chain
MESSIDOR	methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology
OCT	optical coherence tomography
PPV	positive predictive value
PR	precision-recall characteristic
ReLU	rectified linear units
RGB	red, green, blue
RIM-ONE	open retinal image database for optic nerve evaluation
RITE	retinal images vessel tree extraction
RNFL	retinal nerve fiber layer
ROC	receiver operating characteristic
SE	sensitivity
SGD	stochastic gradient descent
SP	specificity
STARE	structured analysis of the retina
SWA	stochastic weight averaging
SWAG	stochastic weight averaging Gaussian

List of publications

This dissertation is based on the following peer-reviewed articles. The rights have been granted by publishers to include the papers in dissertation.

I. Garifullin A., Kööbi O., Ylitepsa P., Ådjers P., Hauta-Kasari M., Uusitalo. H, Lensu L. (2018). Hyperspectral image segmentation of retinal vasculature, optic disc and macula. Conference article, Digital Image Computing: Techniques and Applications (DICTA), pp. 1-5.

II. Garifullin A., Lensu L., Uusitalo H. (2020). On the uncertainty of retinal artery-vein classification with dense fully-Convolutional neural networks. Conference article, Advanced Concepts for Intelligent Vision Systems (ACIVS), pp. 87-98.

III. Lindén, M., Garifullin, A., Lensu L. (2020). Weight averaging impact on the uncertainty of retinal artery-venous segmentation. Conference article, Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis (UNSURE), pp. 52-60.

IV. Garifullin A., Lensu L., Uusitalo H. (2021). Deep Bayesian baseline for segmenting diabetic retinopathy lesions: Advances and challenges. Computers in Biology and Medicine, pp. 104725, Volume 136.

Author's contribution

I prepared literature review, implemented the source code, conducted experiments and prepared the first article draft in papers I, II and IV. In paper III, MSc. Markus Lindén was the first author and I provided the initial source code and supervised further research and development.

Chapter I

Introduction

1.1 Background

Fundus photography is a useful tool offering possibilities for early screening of eye diseases and abnormal medical conditions. However, the screening requires trained personnel to perform the examination which can be time consuming and expensive especially due to the amount of data growing. In this situation computer aided screening tools can help to reduce the workload of the medical staff and to increase the efficiency of health care.

In the last two decades there has been a significant progress in automatic fundus image analysis supported by the development of benchmarks and state-of-the-art machine vision techniques [33, 36]. The majority of the modern approaches are based on deep neural networks and it indicates that deep neural networks are more effective than the classical methods [36].

Certain eye diseases can be diagnosed by solving a classification task where an input image is mapped to a disease label or probability of a presence of the disease [17, 53, 58]. Most of the works published on the problem are based on traditional deep learning approaches where uncertainties of the models and the outputs are not considered [36]. Taking uncertainties into account might be crucial for high-risk applications [32]. Leibig et al. [35] evaluated Bayesian deep learning uncertainty measures and showed improved decision making for the diagnostic performance of diabetic retinopathy. The uncertainty measures were used to decide whether a patient needs a further examination. Filos et al. [10] formalized the previous research as a benchmark for robustness of Bayesian deep learning and compared different Bayesian deep learning approaches. They showed that the new benchmark is more realistic compared to the previously used datasets as modern Bayesian deep models fail to provide reliable uncertainty estimates.

The alternative approach to screening is to assign each pixel of the input image a label describing a type of object to which the pixel belongs. The problem is called semantic segmentation and it is an area of active research. The typical problems include the segmentation of landmarks and different lesions [36]. The segmentation based approaches

can be more interpretable as they explicitly highlight the types of objects detected. Another advantage is that the segmentation based methods can better handle small objects, since the fundus images typically have higher resolution and are downscaled which causes an information loss. However, the scientific community has not sufficiently addressed the problem of reliability of the fundus segmentation methods.

Figure 1.1 shows an abstract scheme of such a computer-aided diagnosis system. The patient's eye is imaged using a fundus camera. Next, the resulting fundus images are processed by a Bayesian deep neural network which produces a probability distribution over the segmentation maps given the fundus images. The inferred distribution can be analyzed by a post-processing algorithm which produces an additional description of the patient's condition. The description can include the status or grades of certain diseases or biomarkers which can be used as indicators of different diseases. The fundus images, segmentation maps and patient's condition are provided to a clinician who can decide diagnosis and whether the patient's treatment plan needs to be revised.

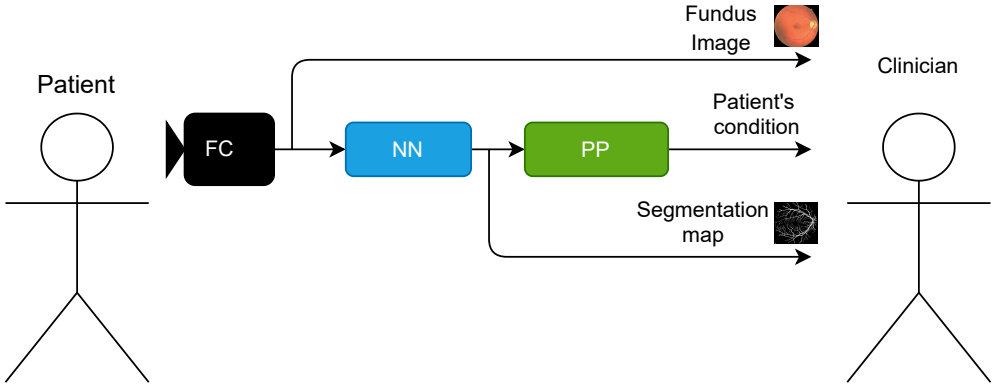


Figure 1.1: A scheme of a computer-aided diagnosis system: *FC* stands for a fundus camera producing fundus images; *NN* is a segmentation neural network producing a probability distribution of segmentation maps; *PP* is a post-processing algorithm characterizing the segmentation maps yielding a patient's condition.

1.2 Objectives

The goal of this work is to develop an uncertainty-aware Bayesian approach to the fundus image segmentation including both landmarks and diabetic retinopathy lesions. The considered landmarks are retinal arteries, veins, optic disc, and macula. The diabetic retinopathy lesions considered are hard exudates, soft exudates, haemorrhages, and microaneurysms. One of the major issues with Bayesian deep neural networks is validating the produced uncertainty estimates which is aimed to be solved in this work. Another unexplored area is the calibration of deep neural networks in the application to the fundus image segmentation.

Thus, the objectives are development of:

1. a Bayesian baseline for retinal artery-vein segmentation using different uncertainty quantification methods.
2. a Bayesian baseline for diabetic retinopathy retinal lesion segmentation and a validation procedure for predicted uncertainties.
3. Bayesian methods for hyperspectral retinal landmarks segmentation and comparative analysis with color fundus image segmentation.

To the best of the author's knowledge, this work is the first study of the topics specified in the list of the objectives above.

1.3 Outline

The rest of the thesis is structured as follows:

Chapter 2 introduces structure of the eye and fundus imaging. An overview of the used datasets is provided. A brief literature review of machine vision methods for disease screening is given.

Chapter 3 contains the theoretical background of Bayesian deep learning and semantic segmentation as well as the segmentation validation metrics. The discussion of related works as well as proposed methods is provided.

Chapter 4 concludes the thesis with the discussion of the results, major issues and limitations together with the possible directions of future research.

Chapter II

The eye, fundus imaging and computer-aided diagnosis

2.1 Structure and diseases of the eye

The human eye is an organ of sight which typically has a spherical shape and located in an orbital cavity. The human eye has a complicated structure. The main object of interest in this work is the eye fundus which is an interior surface of the eye opposite to the lens [29].

The fundus examination can give many insights on the patient's health. The examination is often performed using fundus photography. The fundus photography provides fundus images which contain different objects of interest such as anatomical landmarks or lesions. The normal landmarks of the fundus are as follows [29]:

1. Retinal vasculature consists of arteries and veins. The arteries transport an oxygenated blood from the heart all over the body and the veins transports it back.
2. Optic disc is a circular disc which is formed by the nerve fibre layer. Since there are no light-sensitive cells in the disc, it is also known as the blind spot. The optic nerve is a nerve that extends from the optic disc and transfers the visual information from the retina to the brain. The white circular area in the center of the optic disc is called the optic cup.
3. Macula is in the posterior part of the retina which is a pigmented area that consists of densely-packed photoreceptors (cones) enabling high visual acuity and color vision. The darker region in the center of the macula is called fovea.

Various diseases can affect the fundus in different ways [29] by either affecting the landmarks or causing different lesions.

Hypertensive retinopathy is a vascular disease caused by high blood pressure (hypertension). The risk factors for hypertension include obesity, alcohol abuse, tobacco use and stress. A patient might experience headaches, pain in the eyes or blurred vision. The

disease rarely causes visual loss but can be a sign of other vascular problems. Depending on the severity of the condition, different changes to the vasculature can occur, such as narrowing of the retinal arteries and changes in arteriovenous crossings. The arteriole-to-venule ratio (AVR) is also an important biomarker characterizing the retinal vasculature. The lower values of the ratio can indicate hypertension. The fundus photography can be used to identity these changes or to infer the biomarker [29].

Glaucoma is a chronic optic neuropathy causing damage to optic disc and loss of vision. Typically, glaucoma is caused by high intraocular pressure and the basis for the disease is mostly genetic. The symptoms include photophobia, worsening vision, nausea, ocular pain and eye redness. One way to detect glaucoma is to estimate the biomarker called cup-to-disc ratio (CDR) which is a ratio of the size of the optic cup to the size of the disc. The higher cup-to-disc ratio can provide evidence of the presence of the glaucoma. The biomarker can be inferred from fundus images [29].

Age-related macular degeneration is a condition which leads to the worsening of central vision and distorted and blurred vision. Apart from the advanced age, other risk factors are smoking, obesity and hypertension. The disease is caused by degeneration of arteries causing a lack of oxygen and other nutrients. Depending on the type of the disease, different types of lesions can appear near the macular region [29].

Diabetic retinopathy (DR) is a complication of diabetes damaging the retina and it is one of the leading causes of blindness. The disease affects the retinal vasculature by narrowing the arteries or fusiform venous dilatation. It can also be recognized by the appearance of different DR lesions depending on the grade of the disease. Depending on the proximity of exudative lesions to the macula region, diabetic maculopathy may be present. During the late stages of the disease retinal detachment appears leading to further increasing risks of loss of vision [29].

Apart from the landmarks, the objects of interest in this work include DR lesions [29]:

- Microaneurysms are one of the earliest signs of DR and resemble red small dots. Microaneurysms are caused by the damage to the retinal capillary walls.
- Haemorrhages are red lesions that appear after ruptured microaneurysms. Haemorrhages are bigger than microaneurysms and have unclear edges.
- Soft exudates which are also called cotton wool spots are exudates with blurred edges and contrast. They are the result of obstructed arterioles.
- Hard exudates are yellow lesions with high contrast and clear edges. They are accumulations of lipids under the retinal layer. These lipids leak from damaged blood vessels.

Figure 2.1 illustrates a fundus image with annotations for the landmarks and lesions.

2.2 Color and spectral imaging of the eye fundus

Eye fundus photography is a common imaging technique allowing noninvasive examinations of the fundus. The images of the fundus are acquired using fundus cameras which

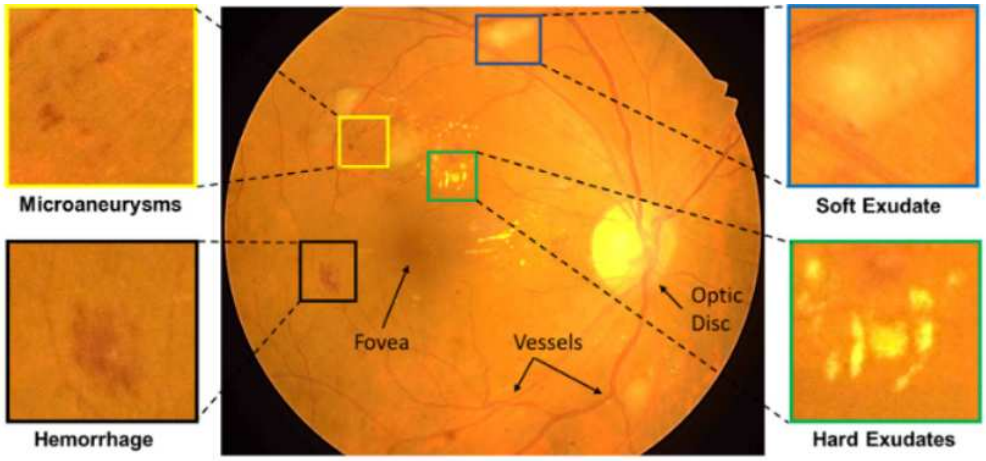


Figure 2.1: The structure of the fundus [46]. Reprinted from Medical Image Analysis, Porwal, P., Pachade, S., Kokare, M., et al., IDRiD: Diabetic retinopathy–segmentation and grading challenge, 101561, ©(2020) with permission from Elsevier.

are based on low power microscopes. In the process of fundus photography, the light from a light source is guided by the optical system to the eye of the patient. The reflected light is then registered by an imaging detector. A complementary metal–oxide–semiconductor or charge-coupled device (CCD) can be used as a detector. Based on these general principles, different devices producing different images can be constructed:

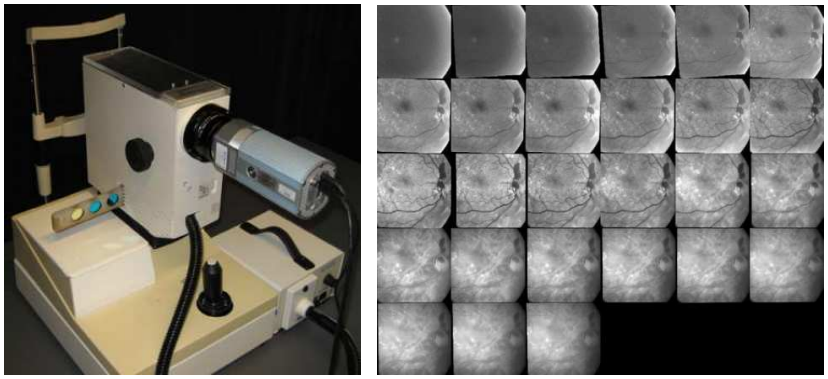
- Color photography provides images with red, green, and blue (RGB) channels.
- Spectral photography provides images where each channel corresponds to a certain wavelength or a limited band of the electromagnetic spectrum.

Color eye fundus photography is widely used for studying diabetic retinopathy, age-related macular degeneration and cardiovascular diseases [2]. The image in Figure 2.1 is an example of the RGB fundus image.

Whereas color fundus cameras provide RGB or grayscale images, spectral fundus imaging systems result in hyperspectral images. In these images each channel corresponds to particular spectral band, i.e., each pixel in the image contains information about the reflectance spectrum of the sample. Different chemical substances in a sample have different reflectance or absorbance spectra, thus, additional features are available for a more refined quantitative analysis [61].

A hyperspectral imaging setup can be a modified fundus camera with a light source with a broadband illumination, and a spectral device for selecting a spectral band. Fält et. al. [12] adapted a Canon CR5-45NM fundus camera to the spectral fundus camera by replacing the standard light source with a fibre optic illuminator consisting of a halogen lamp with illumination spectrum from 380 to 780 nm and 30 interference filters with 10

nm step are used for the wavelengths selection. As a detector, grayscale CCD camera with array size 2048×2048 pixels and 2×2 binning was used. The imaging setup is presented in Figure 2.2a.



(a) **(b)**

Figure 2.2: (a) The spectral fundus imaging setup; (b) An example of a spectral retinal image. The image was normalized for the visualization purpose [33].

2.3 Fundus image datasets

Fundus image datasets allow benchmarking and analysis of different machine vision methods for fundus image analysis. Typically, these datasets contain pairs of input images and corresponding ground truth data. Particular types of the input images and ground truth data depend on the type of problems the study aims to solve. The problems include the landmark segmentation, lesion segmentation, or disease grading.

The datasets used to benchmark the optic disc segmentation algorithms are:

1. Digital retinal images for optic nerve segmentation database (DRIONS-DB) [7] contains 110 color fundus images with spatial resolution 600×400 pixels. The ground truth is presented in the form of contours of the optic disc. The annotations were produced by two medical experts.

2. Open retinal image database for optic nerve evaluation (RIM-ONE) [11] is composed of 169 fundus images where 118 images are gathered from non-glaucomatous patients and the remaining patients have signs of glaucoma of different stages. The spatial resolution of the images is 2144×1424 pixels. The corresponding ground truth data is presented in a form of binary segmentation masks.

The datasets used to benchmark the retinal vasculature segmentation algorithms are:

1. High resolution fundus image database (HRF) [6] contains 45 fundus images with the corresponding binary segmentation masks for retinal blood vessels. The spatial resolution of the images is 3504×2336 pixels.

2. Structured analysis of the retina (STARE) dataset [20] contains 40 images with the corresponding binary segmentation masks for retinal blood vessels. The spatial resolution of the images is 700×605 pixels.

3. Retinal images vessel tree extraction (RITE) dataset [22] contains 40 images labelled for retinal arteries, veins and vessels segmentation. The spatial resolution of the images is 768×584 pixels. RITE dataset is an extension of Digital Retinal Images for Vessel Extraction (DRIVE) dataset.

The datasets used to study methods for detecting signs of diabetic retinopathy are:

1. Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR) dataset [8] contains 1200 images with different spatial resolutions. The images are annotated with DR grades and risks of macular edema. The grades are given based on the number and presence of different DR lesions.

2. DiaRetDB1 [31] is a dataset containing 89 images with the spatial resolution of 1500×1152 pixels. The ground truth segmentation masks are available for DR lesions.

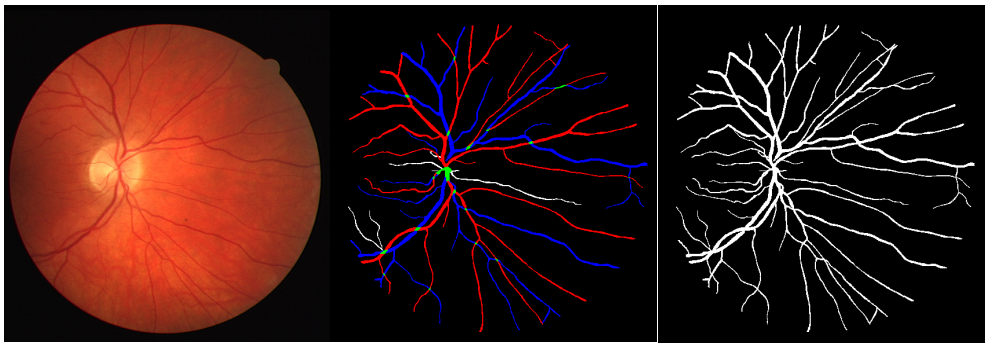
3. Indian diabetic retinopathy image dataset (IDRiD) [45] contains 81 image with the spatial resolution of 4288×2848 pixels. The ground truth information is presented by the segmentation masks for DR lesions, DR grade and binary masks for the optic disc segmentation.

The datasets used in this work were chosen based on the kind of ground truth data presented and availability for open access. RITE dataset is an open access dataset containing the ground truth data for both arteriovenous and vessels segmentation. IDRiD dataset contains the pixel-accurate segmentation masks for DR lesions. These datasets contain RGB fundus images. The exception is DiaRetDB2 dataset which contains 55 spectral images with the segmentation masks for retinal vasculature, optic disc, and macula. The access to DiaRetDB2 dataset was provided by University of Eastern Finland and University of Tampere.

2.3.1 DRIVE and RITE

Digital retinal images for vessel extraction (DRIVE) dataset [21] is a standard benchmark for the retinal vasculature segmentation. The dataset consists of 20 test and 20 train images with the corresponding ground truth segmentation masks for the blood vessels. The ground truth was collected by two experts. The spatial resolution is 768 pixels.

Retinal images vessel tree extraction (RITE) dataset [22] is based on the DRIVE dataset and augments it with the ground truth data for arteries and veins. Figure 2.3 illustrates an example RITE image with the corresponding ground truth. Ground truth labels for the arteries and veins contain labels for the arteries (red), veins (blue), branches (green), and uncertain (white).



(a) (b) (c)

Figure 2.3: The RITE data set: (a) An example test image; (b) corresponding artery-vein reference standard; (c) is the ground truth mask for the blood vessels [22].

2.3.2 IDRiD

Indian diabetic retinopathy image dataset (IDRiD) [45] is a database of fundus images developed for diabetic retinopathy screening research. The dataset contains ground truth data for the optic disc and fovea centers, diabetic retinopathy grade and pixel level segmentation masks for hard exudates, soft exudates, haemorrhages, and microaneurysms. There are 54 images for the train set and 27 images for the test set. The resolution of the input images is 4288×2848. An example image from the dataset is shown in Figure 2.4.

Due to different sizes of the lesions the dataset is very unbalanced. Figure 2.5 shows bar graphs with the number of positive pixels for each lesions and healthy tissue (background).

2.3.3 DiaRetDB2

DiaRetDB2 is a dataset of images with the spatial resolution of 1024×1024 and with the 30 channels where each channel corresponds to the specific wavelength. The dataset contains manual ground truth segmentation masks for the vasculature, optic disc and macula as well as field-of-view (FOV) masks which indicate informative image regions. The segmentation masks for the optic disc and macula were collected by medical experts. The blood vessel annotations were produced by the author during his Master studies. A montage of the spectral bands is shown in Figure 2.2. An example of the sample with the corresponding masks is shown in Figure 2.6.

2.4 Computer-aided diagnosis

The fundus imaging setups can be complemented with computer-aided diagnosis (CAD) systems similar to the one illustrated in Figure 1.1. Such systems utilize computer vision

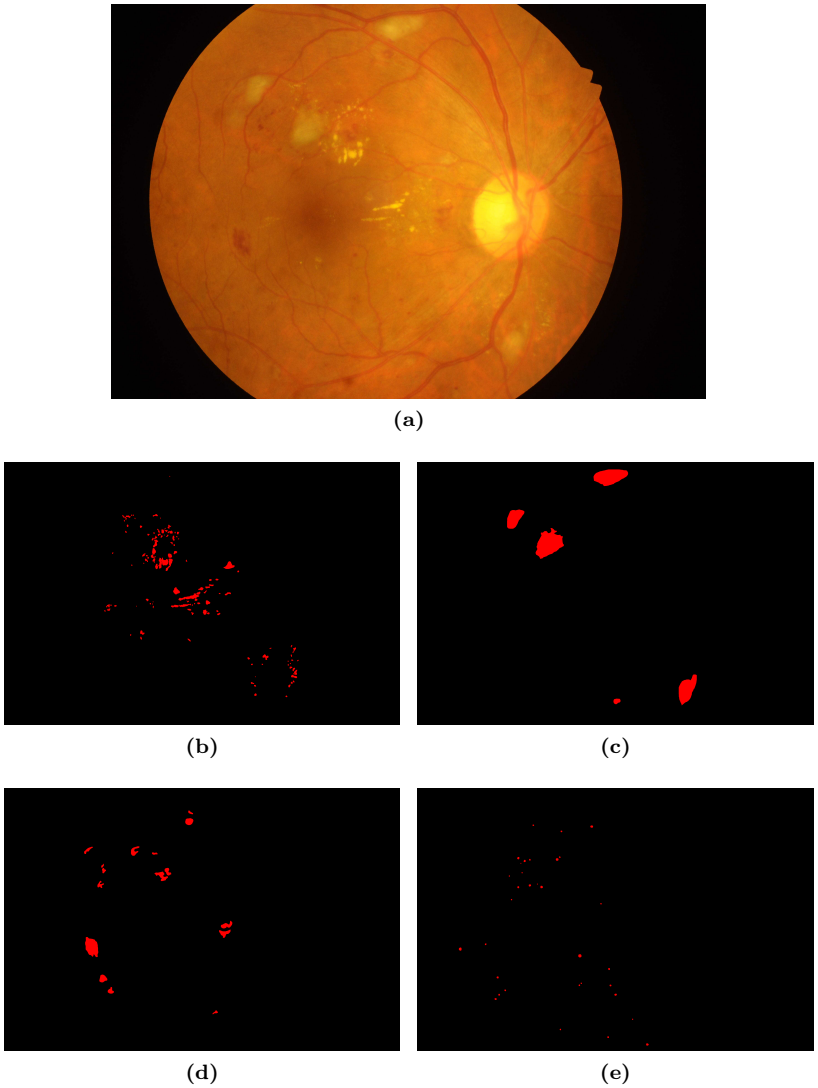


Figure 2.4: (a) An example of IDRiD image with ground truth masks for (b) hard exudates, (c) soft exudates, (d) haemorrhages, and (e) microaneurysms [46].

techniques to produce descriptions of images to help clinicians to make correct diagnoses. Depending on the requirements of the system, the possible implementations are [36]:

1. End-to-end methods that map retinal images to a disease grade [17,35,48,58]. These methods can be implemented as supervised machine learning algorithms trained on input images and the corresponding disease grades. It is also possible to visualize the image features that are relevant to the predicted grade [48, 51].
2. Biomarker-based methods that are algorithms that map retinal images to biomark-

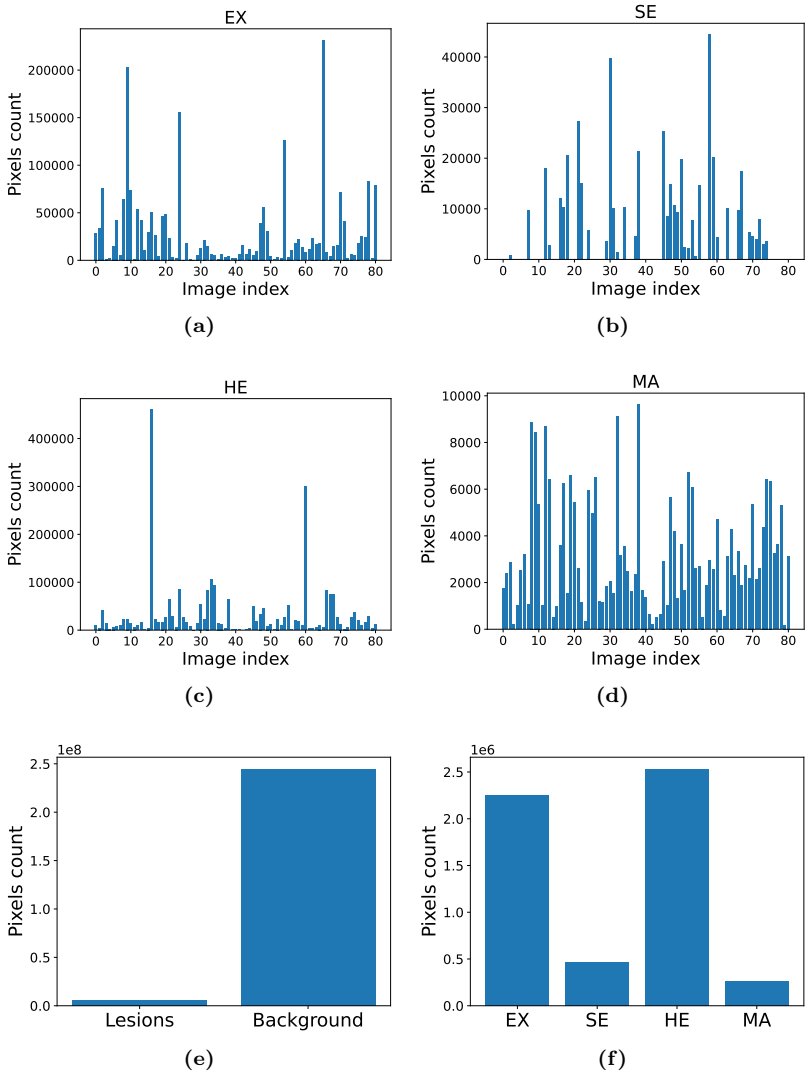


Figure 2.5: Lesion statistics in IDRiD dataset. The number of positive pixels per image for (a) hard exudates (EX), (b) soft exudates (SE), (c) haemorrhages (HE), and (d) microaneurysms (MA). (e) The number of pixels for the lesions and the background. (f) The number of positive pixels for each lesion for the whole dataset.

ers such as the AVR [3, 44] or CDR [24]. The diagnostic decision can be inferred from the predicted biomarkers.

3. Segmentation-based approaches that transform retinal images to segmentation maps where each pixel represents semantic information about the image content. The

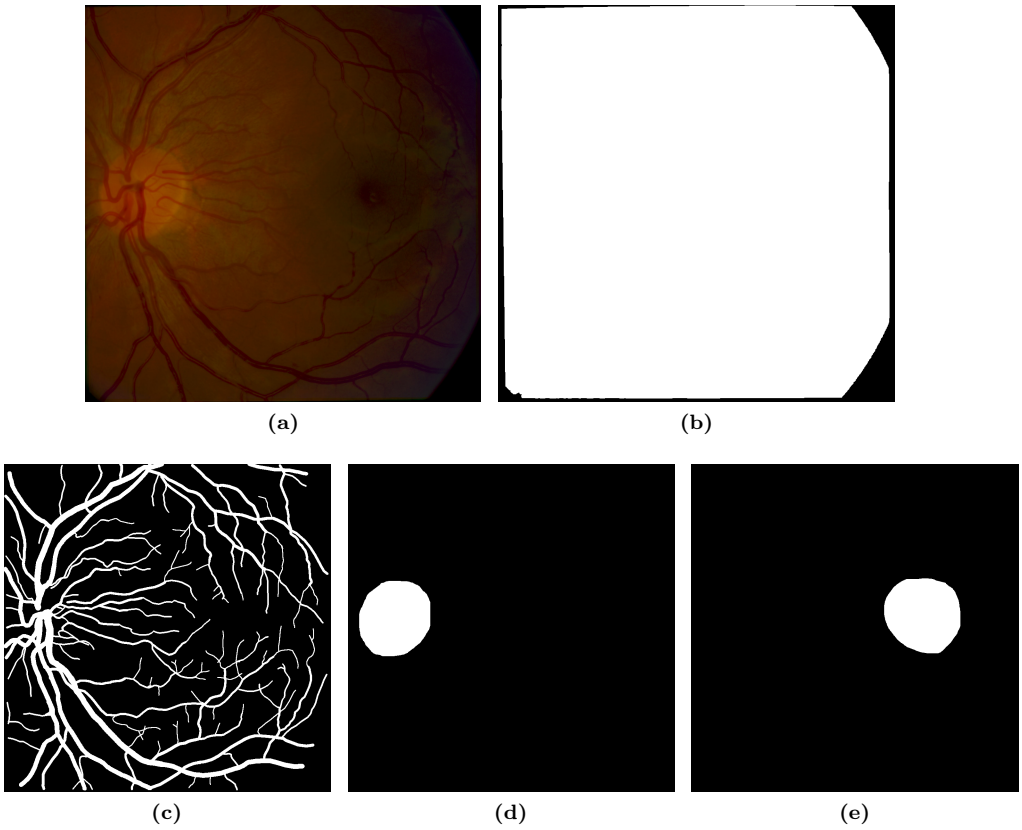


Figure 2.6: (a) An RGB image from the spectral retinal image dataset. (b) FOV mask and the corresponding segmentation masks for the (c) vessels, (d) optic disc and (e) macula.

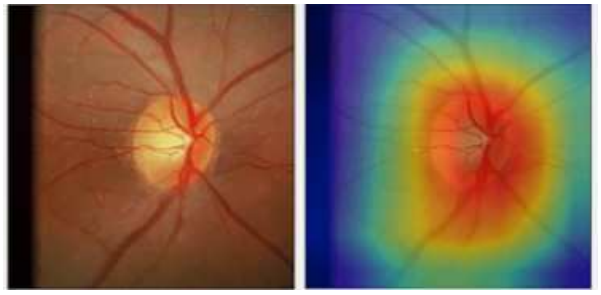
diagnostic decision can be inferred from the segmentation maps depending on the presence of specific lesions [62] or the state of the retinal landmarks [3].

It is worth to note that the described approaches are not mutually exclusive. It is possible to build systems that utilize combinations of these approaches [3, 55, 62].

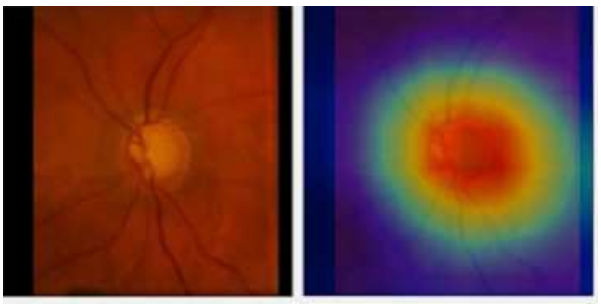
Manikis et al. [40] proposed an image processing framework for detecting early signs of hypertension. The framework includes retinal blood vessel segmentation, optic disc detection and AVR estimation. The authors achieved an accuracy of 0.937 for blood vessel segmentation on the DRIVE dataset. Agurto et al. [3] also relied on methods for retinal vasculature segmentation together with AVR estimation and additional texture feature extraction for the hypertension classification problem. The study was conducted using a private dataset and the authors achieved an accuracy of 0.8 for hypertension prediction. Triwijoyo et al. [58] trained a convolutional neural network in an end-to-end manner to classify images as hypertensive and non-hypertensive. The method achieved

an accuracy of 0.98 on the DRIVE dataset.

Medeiros et al. [41] proposed a deep neural network for retinal nerve fiber layer (RNFL) thickness prediction. The study was conducted using a private dataset that consisted of retinal color images, optical coherence tomography (OCT) scans and medical history for each patient. The RNFL thicknesses were inferred from the OCT scans. The neural network was trained end-to-end to solve the regression problem. Next the RNFL thickness was used to differentiate between glaucomatous and healthy eyes. The authors achieved an accuracy of 0.837 for the glaucoma classification problem. The authors also presented visualizations of the image areas relevant for the network to make the prediction. For this purpose, class activation maps (CAM) [51] were used and example visualizations are presented in Figure 2.7. From the visualizations it is clear that the network focuses more on the optic disc and cup, but can also capture certain areas outside of the optic disc.



(a)



(b)

Figure 2.7: (a) An RGB image and the corresponding CAM for a healthy eye; (b) An RGB image and the corresponding CAM for a glaucomatous eye [41]. The red color denotes more relevant parts, whereas blue represents less relevant parts. Reprinted from Ophthalmology, Vol. 126, Medeiros, F. A., Jammal, A. A., and Thompson, A. C., From machine to machine: An OCT-Trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs, 513-521, ©(2019) with permission from Elsevier.

Sreng et al. [55] proposed an algorithm for optic disc segmentation and glaucoma classification. A schematic illustration of the proposed system is given in Figure 2.8. The system uses a neural network for the optic disc segmentation, and another neural network for

the glaucoma classification. The segmentation network is trained on a database of retinal images with optic disc annotations. Further, the segmented images cropped around the optic disc are reused to train the classification network. The authors achieved an accuracy of 0.997 for the optic disc segmentation and 0.973 for the glaucoma classification. The presented results were achieved on RIM-ONE dataset.

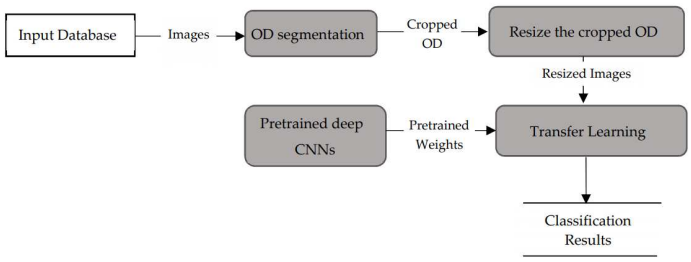


Figure 2.8: A schematic illustration of the method proposed by Sreng et al. [55].

Numerous end-to-end methods for DR grading have been proposed [17, 48]. Typically, they are trained to solve a classification task where the input is a color retinal image and the output is a vector of probabilities for each DR grade. Pratt et al. [48] studied feature visualization of convolutional neural networks for DR grading. Figure 2.9 presents CAMs for different DR grades. The CAMs can highlight areas with DR lesions. However, they are very coarse and it is possible that they highlight irrelevant parts of the images. The evaluation metric used for DR grading is quadratic weighted Kappa on the test data for the multi-class problem (larger values mean better performance). The authors achieved a Kappa value of 0.81.

Wei et al. [62] proposed a method aiming to solve both lesion segmentation and DR grading problem. The proposed network is a DR classification network with a side-stream for DR lesion segmentation and classification. Figure 2.10 presents a schematic representation of the proposed network. The authors achieved a state-of-the-art Kappa of 0.803 for DR grading and 0.801 for DR lesion classification. Figure 2.11 shows visualizations of the results of DR lesion segmentation and classification. From the figure it can be seen that the produced segmentation are more accurate than the CAM visualizations in Figure 2.9.

Based on the above examples, fundus image segmentation plays an important role in fundus image analysis. Fundus image segmentation can be used to help to assist in diagnosing hypertension, diabetic retinopathy, and glaucoma.

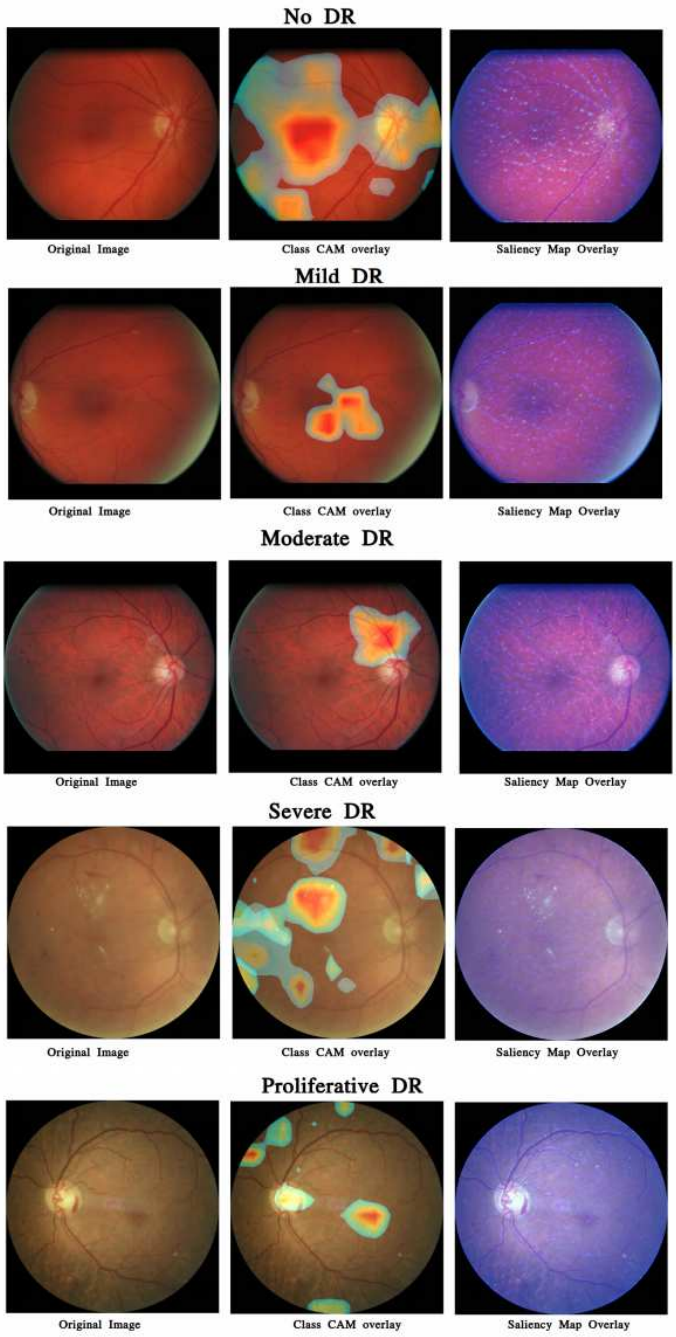


Figure 2.9: (Left) Fundus images from the Liverpool Diabetic Eye Screening Program (LDESP). (Middle) Class activation maps (CAMs) from the trained DenseNet multi-class DR model overlayed on the original image. (Right) Saliency map from the trained DenseNet multi-class diabetic retinopathy (DR) model overlayed on the original fundus image [48].

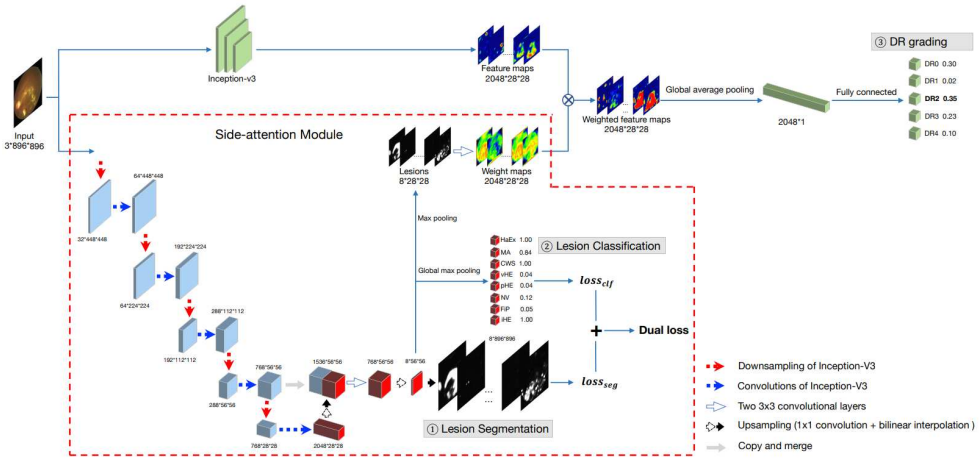


Figure 2.10: A schematic illustration of the method proposed by Wei et al. [62] (©2020 IEEE).

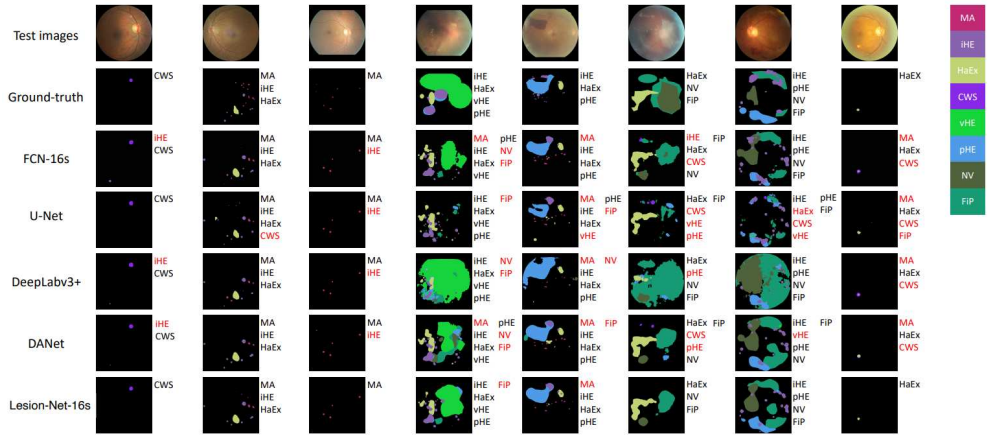


Figure 2.11: Qualitative results of lesion segmentation and classification. The red font indicates false alarms. The results are given for the method proposed by Wei et al. [62] (©2020 IEEE).

Chapter III

Fundus image segmentation

3.1 Methodology

3.1.1 Deep learning for semantic segmentation

Fundus image segmentation is a subproblem of semantic segmentation where pixels are grouped based on their semantic similarity. This problem can be efficiently solved using supervised deep learning methods [36]. Let $\mathcal{D} = \{ (\mathbf{x}, \mathbf{p})_i \}_{i=0}^{N-1}$ be a dataset of N input-output pairs where $\mathbf{x}$ is an input image and $\mathbf{p}$ is a corresponding ground truth segmentation map. Then the training can be formulated as an inference problem of parameters $\boldsymbol{\theta}$ of a model f that maps the input image to an estimate of the segmentation map $\hat{\mathbf{p}}$ [13]:

$$\hat{\mathbf{p}}_i = \text{sigmoid}\left(f\left(\mathbf{x}_i, \boldsymbol{\theta}\right)\right), \tag{3.1}$$

where $\text{sigmoid}(y) = (1 + \exp(-y))^{-1}$ is the sigmoid activation function mapping logits y to label probabilities.

The most common way of estimating the parameters is finding a maximum a posterior (MAP) estimate [13]

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p\left(\boldsymbol{\theta} \mid \mathcal{D}\right), \tag{3.2}$$

where $\hat{\boldsymbol{\theta}}$ is a MAP estimate of the parameters and $p\left(\boldsymbol{\theta} \mid \mathcal{D}\right)$ is the posterior probability distribution of the parameters defined as [13]

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \frac{p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(\mathcal{D}\right)}, \tag{3.3}$$

with the likelihood [13]

$$p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) = \prod_{i=0}^{N-1} p\left(\mathbf{p}_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right), \tag{3.4}$$

the prior over the parameters $p\left(\boldsymbol{\theta}\right)$, and the evidence $p\left(\mathcal{D}\right) = \int p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}$.

The problem (3.2) is typically reformulated as a minimization problem [13]:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} -\log p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) \\ &= \arg\min_{\boldsymbol{\theta}} -\left[\sum_{i=0}^{N-1} \log p\left(\mathbf{p}_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta}\right) - \log p\left(\mathcal{D}\right)\right] \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=0}^{N-1} \mathcal{L}\left(\mathbf{p}_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right) + \mathcal{R}\left(\boldsymbol{\theta}\right), \end{aligned} \tag{3.5}$$

where $\mathcal{L}$ is a negative log-likelihood which is responsible for the data fit also known as the loss function. $\mathcal{R}$ is the negative log-prior of the parameters which acts as a regularization term and the log-evidence $p(\mathcal{D})$ is cancelled being a constant not depending on the parameters.

In the case of image segmentation, it is natural to formulate the loss function as the binary cross-entropy (BCE)

$$\mathcal{L}\left(\mathbf{p}_i \mid \mathbf{x}_i, \boldsymbol{\theta}\right) = -\mathbf{p}_i \log \hat{\mathbf{p}}_i - \left(1 - \mathbf{p}_i\right) \log\left(1 - \hat{\mathbf{p}}_i\right). \tag{3.6}$$

In this work the prior distribution is modelled as a fully-factorized Gaussian distribution

$$p\left(\boldsymbol{\theta}\right) = \prod_{i=0}^{D-1} \mathcal{N}(0, \sigma_\theta^2), \tag{3.7}$$

where D is the number of parameters and σ_θ controls the regularization strength.

The optimization problem (3.5) is typically solved using gradient descent based methods. One of the basic examples of such techniques is the stochastic gradient descent algorithm (SGD). The method differs from the standard gradient descent method in estimating the gradient by using mini-batches of the data examples. This modification helps to save computational resources while solving the problem. Nowadays, different modifications that improve the convergence of SGD are used [50].

3.1.2 Bayesian deep learning

The approach described above produces only point estimates of the segmentation labels and the model's parameters. In order to better capture imperfect ground truth labelling and imaging conditions, it is possible to define the model's outputs and parameters as random variables and infer distributions over them. The first approach takes into account the aleatoric heteroscedastic uncertainty, while the latter models the epistemic uncertainty [32].

Aleatoric uncertainty

The aleatoric uncertainty is a data induced uncertainty that can be caused by the imperfect imaging conditions. It can be included into the model (3.1) by predicting standard deviations of the outputs together with the outputs themselves [32]:

$$\left[\hat{\mathbf{y}}_i, \boldsymbol{\sigma}_i\right] = f\left(\mathbf{x}_i, \boldsymbol{\theta}\right), \tag{3.8}$$

$$\hat{\mathbf{p}}_i = \text{sigmoid}\left(\hat{\mathbf{y}}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}\right), \tag{3.9}$$

where $\hat{\mathbf{y}}$ is a vector of logits, $\odot$ stands for the Hadamard product, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is a normally distributed noise with the identity covariance matrix $\mathbf{I}$.

Taking into account the modified model (3.8) it is possible to modify the loss function to work with multiple aleatoric samples $\hat{\mathbf{p}}_{ij}$

$$\mathcal{L}_A(\mathbf{p}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{j=0}^{N_A-1} -\mathbf{p}_i \log \hat{\mathbf{p}}_{ij} - (1-\mathbf{p}_i) \log(1-\hat{\mathbf{p}}_{ij}), \tag{3.10}$$

where $\hat{\mathbf{p}}_{ij}$ is a j-th sample inferred from the input image $\mathbf{x}_i$, and N_A is the number of aleatoric samples.

Epistemic uncertainty

The epistemic uncertainty captures the model's ignorance about the underlying problem. From (3.2) one can see that $\boldsymbol{\theta}$ is a random variable which can be marginalized during the inference [13]

$$p(\mathbf{p}_* \mid \mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{p}_* \mid \mathbf{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}. \tag{3.11}$$

Calculating the posterior predictive (3.11) is a difficult task, since it involves taking the intractable integral. Different approximating techniques are used instead [1].

Gal et al. [14] reinterpreted dropout [56] as a stochastic variational inference technique, where the complex posterior distribution (3.3) was replaced by a simpler variational approximant $q_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ with parameters $\boldsymbol{\omega}$. This approach is called Monte-Carlo dropout (MC-Dropout) The relationship between the true and approximate posteriors is given by [14, 56]

$$\boldsymbol{\omega} = \boldsymbol{\theta} \odot \mathbf{M}_D, \tag{3.12}$$

where $\mathbf{M}_D$ is a random binary dropout mask. In this case, the training algorithm aims to minimize the difference between the true posterior and approximant [14]:

$$\mathcal{L}_{\mathrm{VI}}(\boldsymbol{\omega}) = \int q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \log p(\mathcal{D} \mid \boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega} - D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega})), \tag{3.13}$$

where D_{KL} is the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega})) = \int q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \log \frac{q_{\boldsymbol{\theta}}(\boldsymbol{\omega})}{p(\boldsymbol{\omega})} d\boldsymbol{\omega}. \tag{3.14}$$

The formula (3.13) is similar to (3.5) in a sense that the second term penalizes the model to be close to the prior and the first term is responsible for the data fit and it is typically approximated using Monte-Carlo methods

$$\mathcal{L}_{\mathrm{VI}}(\boldsymbol{\omega}) \approx \sum_{i=0}^{N-1} \sum_{j=0}^{N_E-1} \frac{1}{N_E} \mathcal{L}(\mathbf{p}_i \mid \mathbf{x}_i, \boldsymbol{\omega}_j) + \mathcal{R}_{\mathrm{VI}}(\boldsymbol{\omega}), \tag{3.15}$$

where N_E is the number of epistemic samples, and the variational regularization term is $\mathcal{R}_{\mathrm{VI}}(\boldsymbol{\omega}) = D_{\mathrm{KL}}(q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \parallel p(\boldsymbol{\omega}))$.

Maddox et al. [39] proposed to model the posterior distribution of the parameters as a fully-factorized Gaussian distribution

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \mathcal{N}(\boldsymbol{\theta}_{\mathrm{SWA}}, \boldsymbol{\Sigma}_{\mathrm{SWAG}}), \tag{3.16}$$

the parameters of which are estimated during training. This technique is called stochastic weight averaging Gaussian (SWAG) and it is based on the stochastic weight averaging (SWA) proposed in [25].

In more traditional approaches, Monte-Carlo Markov chain (MCMC) methods are typically used. For deep models, however, it is difficult to scale them properly due to the high-dimensionality of the problem and costly likelihood evaluations. Ma et al. [38] formalized stochastic gradient extensions of the classical MCMC algorithms, which can work with subsets of the datasets to utilize stochastic gradient information to explore the distributions, and can be used to quantify epistemic uncertainty [26].

The theory above describes general principles of the Bayesian deep learning approach which can be applied to a variety of different architectures that formalize the model f in (3.1).

3.1.3 Neural network architectures

The most of modern architectures for deep semantic segmentation are encoder-decoder models. The encoder compresses the input images to a hidden representation. Then, this representation is reconstructed by decoders into a feature map which is further transformed to the segmentation map using a pixelwise classifier.

One of the basic examples of such architectures is SegNet [5]. The encoder is composed of blocks of convolutional layers, batch normalization (BN) and rectified linear units (ReLU) which are followed by max-pooling [16]. The decoder is a symmetric reflection of the encoder with the pooling layers replaced by the upsampling layers using pooling indices to recover feature maps. Figure 3.1 is a schematic illustration of the architecture.

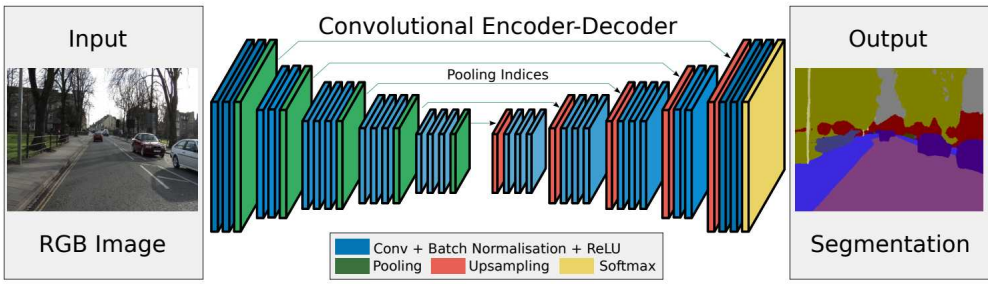


Figure 3.1: SegNet architecture [5] (©2020 IEEE).

Ronneberger et al. [49] proposed the U-Net architecture which follows similar principles but also allows the data leakage from the encoder to the decoder so that the high resolution feature maps are cropped, copied and concatenated with the decoded feature maps. This mechanism allows to preserve more information about border pixels and

fine details. The architecture was developed specifically for medical image segmentation problems and it is one of the most widely used architectures across a variety of different domains [67]. Figure 3.2 shows the U-Net architecture.

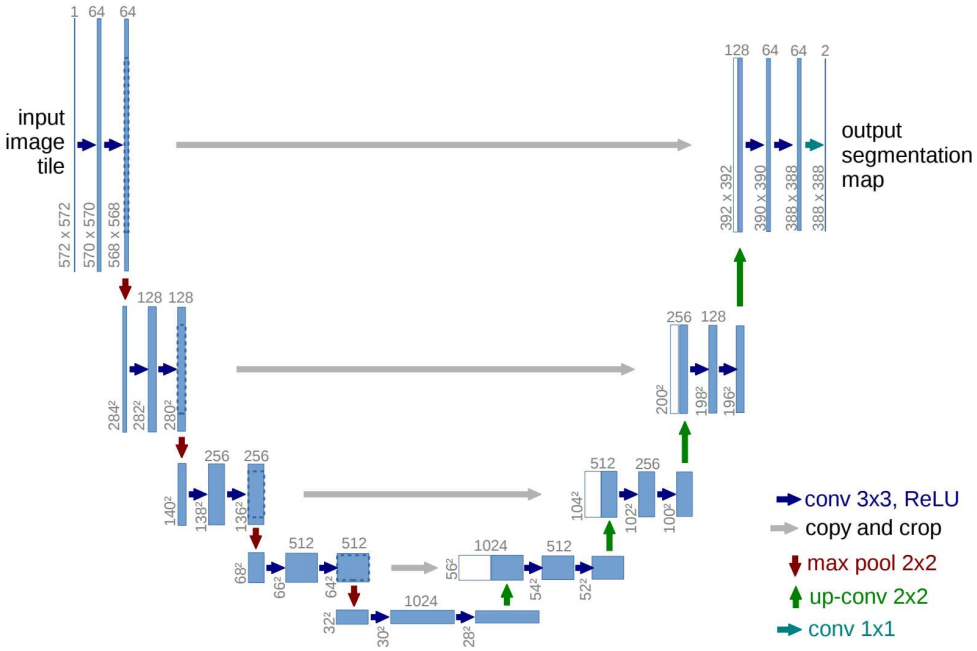


Figure 3.2: UNet architecture [49]. Reprinted by permission from Springer Nature Customer Service Centre GmbH, Springer Nature, Medical Image Computing and Computer- Assisted Intervention – MICCAI 2015. U-Net: Convolutional networks for biomedical image segmentation, Ronneberger, O., Fischer, P., and Brox, T., ©(2015).

In addition to reusing features from the encoder for the decoding purposes, it is also possible to adapt a similar approach to the internal components of the encoders and decoders. Gao et al. [23] proposed dense convolutional networks which connect outputs of previous convolutional layers to subsequent convolutional layers which form dense convolutional blocks (DCB). This approach helps to overcome the vanishing gradients problem and to make networks deeper without significantly increasing the number of parameters. Jégou et al. [27] adapted this approach to build dense fully-convolutional neural networks (Dense-FCN) for the semantic segmentation purposes. The architecture reuses features from different resolutions as well as features from different blocks of the encoder and decoder. Figure 3.3 illustrates an example of Dense-FCN architecture.

3.1.4 Segmentation and uncertainty validation

A trained model needs to be tested on a validation set with the goal of estimating its performance. Firstly, point estimates of the segmentation masks are obtained as average

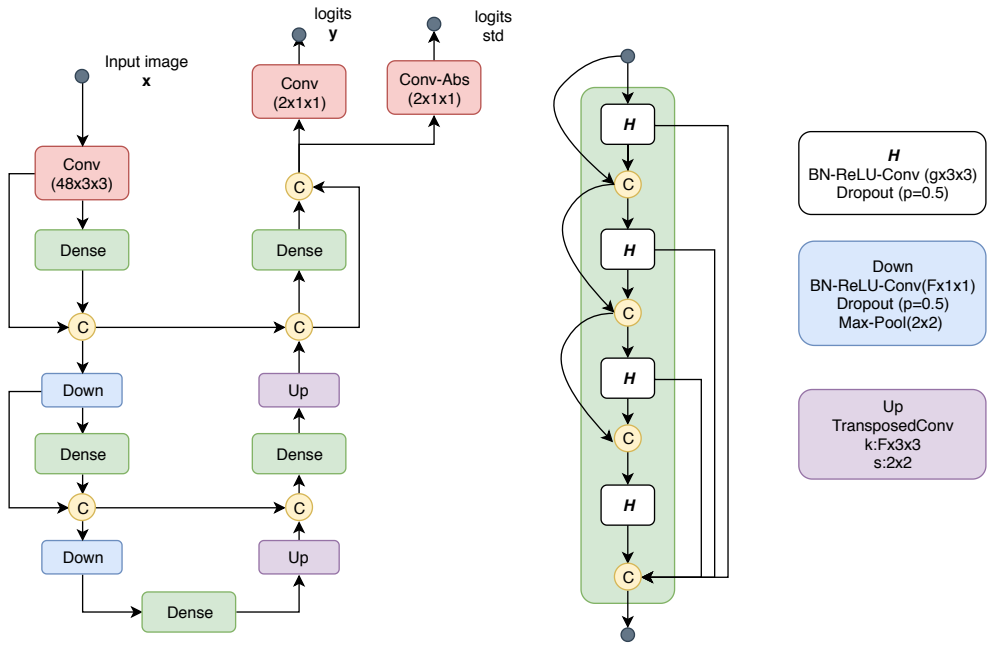


Figure 3.3: The Dense-FCN architecture: *Dense* stands for DCB; C is a tensor concatenation; H is a block consisting of the batch normalization, rectified linear unit and convolutional layer with growth rate g; *Down* is a transition-down block with F output feature maps; *Up* is a transition up with F output feature maps and 2×2 stride; *logits std* denotes standard deviations of logits.

probabilities using Monte-Carlo sampling:

$$\bar{\mathbf{p}}_i = \frac{1}{S} \sum_{s=0}^{S-1} f\left(\mathbf{x}_i, \boldsymbol{\omega}_s\right), \quad \boldsymbol{\omega}_s \sim q_{\boldsymbol{\theta}}, \tag{3.17}$$

where $S = N_E \times N_A$ is the total amount of samples produced by the model, and the variational parameters $\boldsymbol{\omega}_s$ are sampled from the approximate probability distribution $q_{\boldsymbol{\theta}}$.

The aleatoric U_A and epistemic U_E uncertainties can be estimated as

$$U_A = \mathbb{E}_q \left[\mathbb{V}_{p(\mathbf{p}|\mathbf{x},\boldsymbol{\omega})} \left[\mathbf{p}\right]\right], \tag{3.18}$$
$$U_E = \mathbb{V}_q \left[\mathbb{E}_{p(\mathbf{p}|\mathbf{x},\boldsymbol{\omega})} \left[\mathbf{p}\right]\right], \tag{3.19}$$
$$U_T = U_A + U_E, \tag{3.20}$$

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

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:

$$SE = \frac{TP}{TP + FN}, \tag{3.21}$$

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 takes into account false positives FP:

$$PPV = \frac{TP}{TP + FP}. \tag{3.22}$$

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

$$SP = \frac{TN}{TN + FP}, \tag{3.23}$$

where TN is the amount of true negative pixels.

- Intersection over union (IoU)

$$IoU = \frac{T \cap P}{T \cup P}, \tag{3.24}$$

where T is a set of target pixels and P is a set of predicted pixels.

- F1 score

$$F1 = \frac{TP}{TP + 0.5(FP + FN)}. \tag{3.25}$$

- The metrics above are calculated by thresholding the label probabilities (3.17). In this work the threshold value is 0.5. ROC-AUC is an integral metric regardless of the threshold 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 the precision-recall curve (PR-AUC) is another integral metric regardless of the threshold value. PR-AUC more realistically represents the segmentation performance in comparison to the area under receiver operating characteristic ROC-AUC [46].

- Expected calibration error (ECE) is used to assess a model's calibration [18]:

$$ECE = \mathbb{E}_{\hat{p}}\left[\left|\mathbb{P}\left(\hat{l} = l \mid \hat{p} = \pi\right) - \pi\right|\right], \quad \pi \in [0, 1], \tag{3.26}$$

where $\hat{p}$ is a confidence estimate of the predicted class $\hat{l}$, l is a true label and π is a true probability. Together with ECE, reliability diagrams can be 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.

Apart from evaluating the segmentation, it is also important to assess the estimated uncertainty. In this work the uncertainty evaluation procedure is based on the assumptions presented by Mobiny et al. [42] that the misclassified pixels must have higher uncertainties. Thus, in this work the uncertainty validation procedure is formulated as a binary classification problem where the estimated uncertainties are considered as predicted classification scores and the misclassifications are as ground truth labels. The uncertainty validation metrics used in this work are similar to those used to validate the segmentation results but with prefix U.

3.2 Retinal artery-vein segmentation

3.2.1 Background

The problem of the retinal artery-vein (AV) segmentation considered in this work is the simultaneous segmentation of the vasculature and its classification into arteries and veins. The problem can be solved by just applying regular frameworks for semantic segmentation but the major issue of segmenting thin vessels remains. In order to overcome the problem Girard et al. [15] proposed a post-processing technique which builds a vasculature tree and uses a likelihood propagation score to update the segmentation maps based on the connectivity patterns. Badawi et al. [4] aimed to solve the same problem by augmenting the BCE loss with an additional segment-level loss which is defined through the mismatch between the segments extracted from the vasculature tree. Zhang et al. [65] achieved better performance by training a refined U-Net that minimized a multi-scale loss. The utilized multi-scale loss was inspired by [37] and it sums loss values from different stages of decoding and the downscaled ground truth segmentation. Zhang et al. also proposed to use a cascade network which predicts the probabilities of the vessels labels and then sequentially passes the results to subnetworks for arteries and veins, Figure 3.4 illustrates this principle. More detailed review of artery-vein segmentation approaches is given in [43].

3.2.2 Research findings

This section presents the results from Publication II and Publication III. In this work the segmentation for the arteries and veins first produced $\mathbf{p} = [p_{\text{artery}}\ p_{\text{vein}}]$ and then the probabilities for the blood vessels are inferred:

$$p_{\text{vessel}} = p_{\text{artery}} + p_{\text{vein}} - p_{\text{artery}}p_{\text{vein}}. \tag{3.27}$$

The minimized loss function is a sum of three terms for each label:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\mathcal{L}_{\text{artery}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{vein}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{vessel}}(\boldsymbol{\theta})\right], \tag{3.28}$$

where $\mathcal{L}$ denotes the BCE loss for each corresponding label. The epistemic uncertainty was estimated using three different methods:

1. MC-Dropout is a baseline method;

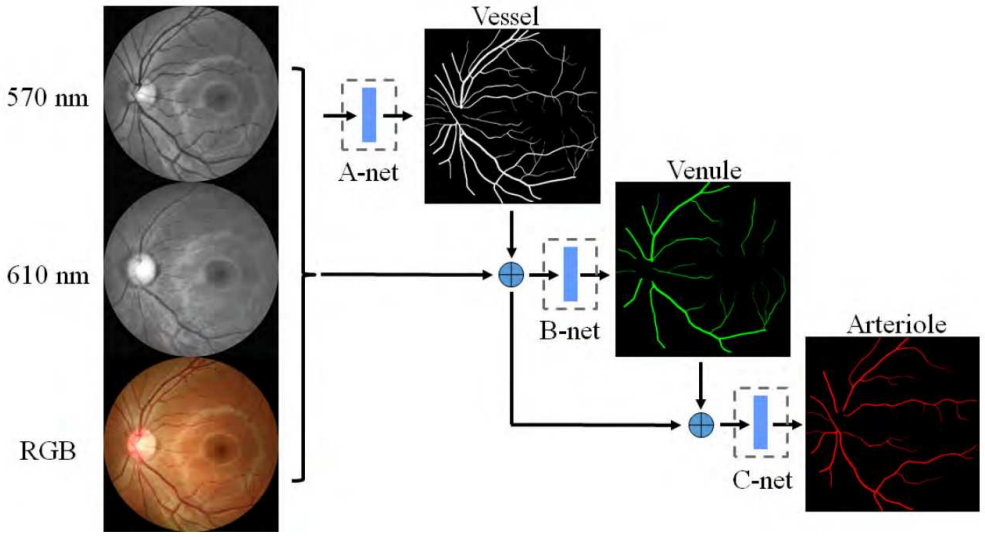


Figure 3.4: Cascade network for the artery-veins segmentation [65] (©2020 IEEE).

2. SWA-MC-Dropout is a method which used MC-Dropout but SWA was applied as a part of the training;

3. SWAG is a method which estimates an approximation of the normal distribution during SWA training stage.

The performance metrics for all three labels are given in Tables 3.1 – 3.3[1]. Figure 3.5 shows an example of the resulting AV segmentation.

Table 3.1: Network performance in artery classification (the best accuracy and calibration are presented in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.970	0.642	0.990	0.0988	0.974
SWA	**0.975**	0.690	0.992	0.0943	0.981
SWAG	0.973	0.706	0.989	**0.0871**	0.966

The examples of the estimated aleatoric and epistemic uncertainties are shown in Figure 3.6. From the images one can notice changes in aleatoric uncertainties when the weight averaging is applied. In the baseline case the aleatoric uncertainty is mostly higher near the optic disc and edges of the vessels. If the weight averaging is applied the pattern is similar but the uncertainties near optic disc are lower. It is also clear that just sampling around the found optimum using SWAG yields lower epistemic uncertainty than sampling

[1]Due to an error in the code calculating the average of the calibration errors the ECE values in Publication III are wrong. Here the corrected values are given.

Table 3.2: Network performance in vein classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.971	0.655	0.994	0.169	0.980
SWA	**0.974**	0.742	0.991	0.120	0.991
SWAG	0.971	0.804	0.983	**0.107**	0.980

Table 3.3: Network performance in vessel classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.957	0.723	0.989	0.221	0.980
SWA	**0.961**	0.782	0.986	**0.208**	0.983
SWAG	**0.961**	0.836	0.978	0.338	0.984

using MC-Dropout. Table 3.4 presents the total estimated uncertainties which gives the quantitative support to the claims above.

Table 3.4: Mean sums of estimated aleatoric and epistemic uncertainties per image.

Method	Aleatoric			Epistemic		
	Arteries	Veins	Vessels	Arteries	Veins	Vessels
Baseline	1276.2	1159.5	1807.5	4853.6	4066.4	5069.7
SWA	3.3	3.5	5.3	4038.6	3882.3	4659.7
SWAG	31.1	38.9	57.3	997.8	1104.3	1396.1

The proposed methods yield performance comparable to the state of the art methods without any additional preprocessing or multi-scale loss functions. Table 3.5 shows a comparison of the performance of recent works and proposed methods.

3.3 Diabetic retinopathy lesion segmentation

3.3.1 Background

The IDRiD challenge [45, 46] is the common benchmark for diabetic retinopathy lesion segmentation algorithms. The best performing algorithms in the challenge are presented by deep learning based techniques. The authors experimented with different architectures and custom loss functions such as combinations of BCE and dice loss or balanced BCE [46]. The dataset is highly imbalanced and the custom loss functions were employed to overcome these problems. The input images are very high-dimensional 4288×2848 and all the reported methods were trained on cropped patches.

Yan et al. [63] proposed an architecture which aims to solve the problem of high dimensionality of the inputs and a lack of the global context when training only using the

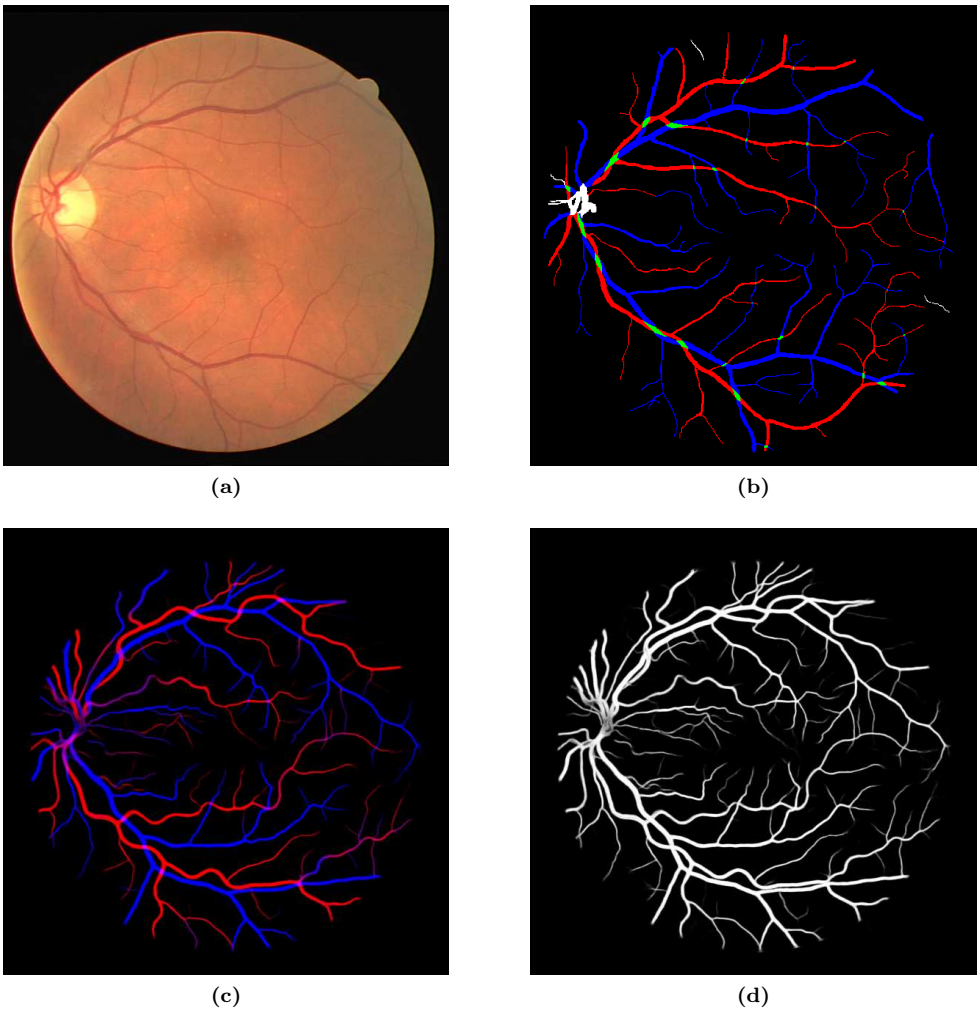


Figure 3.5: (a) The input image; (b) ground truth; (c) mean predicted AV probabilities; (d) mean predicted vessel probabilities. The results are obtained using stochastic weight averaging.

cropped images. The architecture consists of two U-Nets. The first network is the GlobalNet which processes a downscaled input image and produces a coarse segmentation map. The second network is the LocalNet which processes cropped patches and concatenates corresponding cropped features with the features from GlobalNet and produces the resulting segmentation map. The network is trained end to end using a combination of local and global supervision. Figure 3.7 illustrates the architecture proposed by Yan et al. The reported PR-AUC for hard exudates is 0.889, for soft exudates 0.697, for haemorrhages 0.703, and for microaneurysms 0.525.

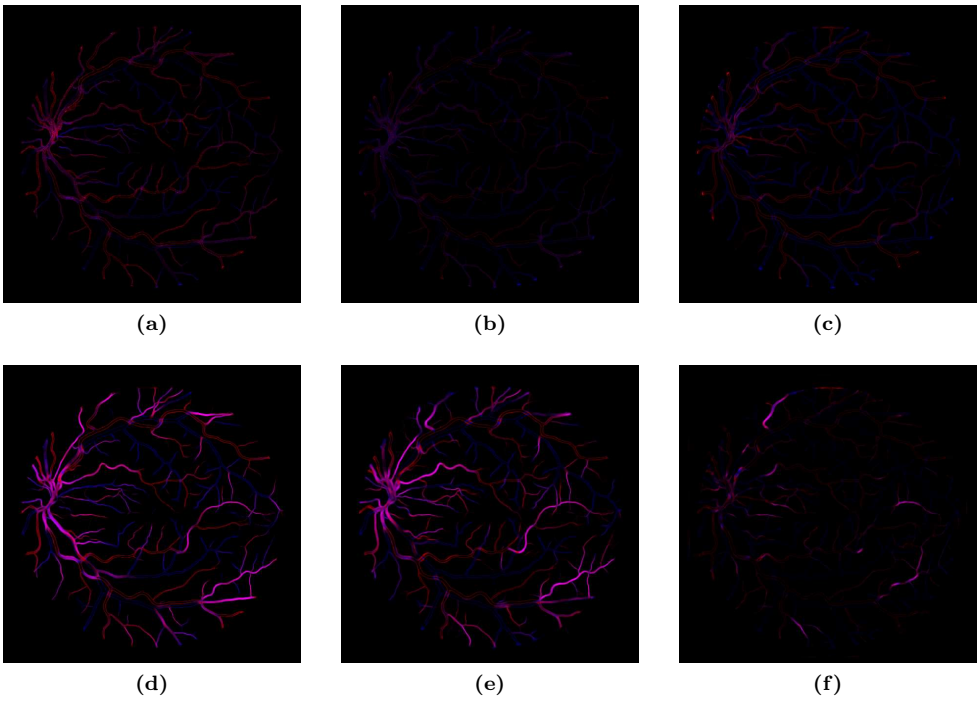


Figure 3.6: Aleatoric uncertainties calculated using (a) the baseline, (b) stochastic weight averaging, and (c) stochastic weight averaging Gaussian. Epistemic uncertainties calculated using (d) the baseline, (e) stochastic weight averaging, and (f) stochastic weight averaging Gaussian. The pseudo-colors represent different labels. The red channel shows the artery segmentation uncertainty, and the blue channel shows the vein segmentation uncertainty

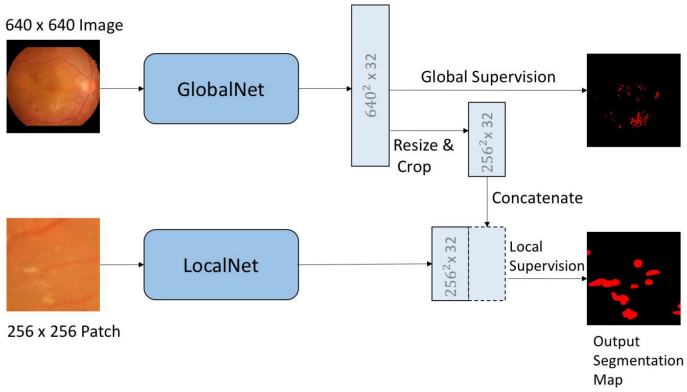


Figure 3.7: Local-Global U-Nets for DR lesion segmentation [63] (©2020 IEEE).

Table 3.5: Comparison of evaluation results (accuracies of each label). The datasets are specified with splitting methods used by the authors.

Method	Vessels	Arteries	Veins	Dataset
Girard et al. [15]	0.948	N/A	N/A	CT-DRIVE
Badawi et al. [4]	0.960	N/A	N/A	DRIVE (standard)
Hemelings et al. [19]	N/A	0.948	0.930	DRIVE (standard)
Zhang et al. [65]	N/A	0.977	0.975	DRIVE (5-fold CV)
Baseline	0.957	0.970	0.971	DRIVE (standard)
SWA	0.961	0.975	0.974	DRIVE (standard)
SWAG	0.961	0.973	0.971	DRIVE (standard)

3.3.2 Research findings

This section presents the results from Publication IV. In this work the basic Bayesian deep learning approach with Dense-FCN was used. The major challenge was to overcome the class imbalance problem. The most efficient approach in this work appeared to be oversampling. For the input batch positive and negative samples are selected with a probability of 0.5. The probability of selecting a certain image is a logarithm of the positive pixels in the image normalized to the total amount of positive pixels in the dataset. The probability of sampling a certain patch is a logarithm of the number of the positive pixels normalized to the total number of positive pixels in the image. The network is trained on 224 patches and processes downsampled images 2144×1440 as a whole. It was empirically found that a simple preprocessing based on gamma-correction and contrast limited adaptive histogram equalization [66] improves the segmentation performance.

Table 3.6 shows the performance metrics for the DR lesions segmentation using the proposed Bayesian method. From the table one can see that the trained segmentation models are very specific and the main issue is the sensitivity. The best achieved performance is for the hard exudate segmentation, since the hard exudates have clear edges and are relatively big. The soft exudates and haemorrhages typically have lower contrast and blurred edges in comparison with hard exudates. The microaneurysms are the most difficult to segment as they are the smallest lesions.

Table 3.6: Evaluation results of the baseline training scheme.

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

The examples of the resulted segmentations together with the visualizations of the misclassifications and uncertainties are given in Figures 3.8 – 3.11. From the images it is clear that there are certain similarities between the misclassifications and epistemic uncertainty visualizations. From Figure 3.9 it is clear that soft exudates can be easily confused with any yellow spots on the image, and from Figure 3.10 it can be noted

that haemorrhages can be confused with the vasculature, since the model trained by DR lesions segmentation did not learn anything about the retinal vasculature.

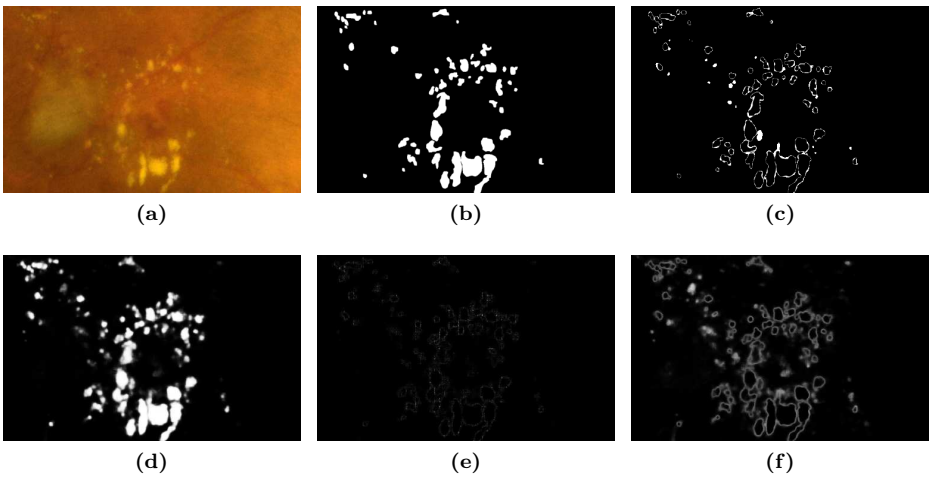


Figure 3.8: Inference results for hard exudates: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

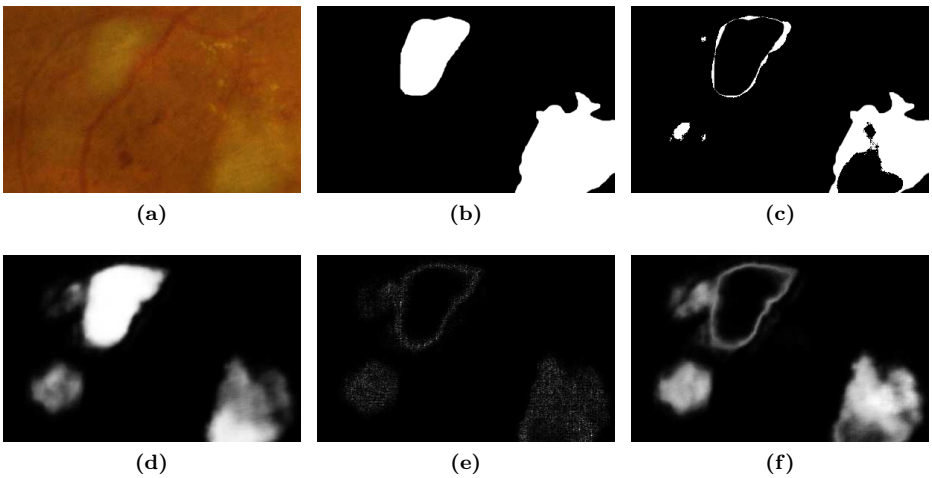


Figure 3.9: Inference results for soft exudates: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

Table 3.7 shows the evaluation results for the estimated uncertainties. The uncertain-

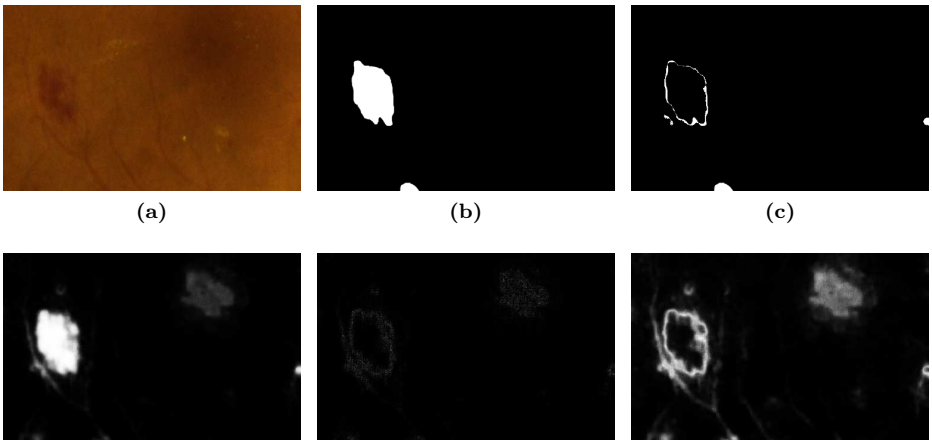


Figure 3.10: Inference results for haemorrhages: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

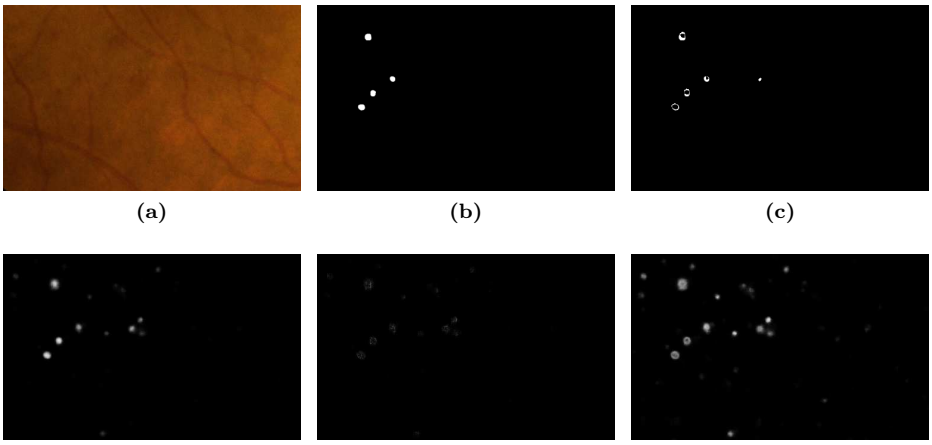


Figure 3.11: Inference results for microaneurysms: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

ties are very specific but the sensitivity is very low indicating that the uncertainty has been underestimated. Laves et al. [34] reported similar problems in the context of deep Bayesian regression. In this work different attempts to calibrate the uncertainty estimates

Table 3.7: Evaluation results for the estimated uncertainty maps.

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

were maid but significantly better results were not achieved.

3.4 Hyperspectral image segmentation

3.4.1 Background

The architectures described in Section 3.1.3 have can be adapted to process HSI by reducing dimensionality of the spectral features to appropriate dimensions suitable for the pretrained networks. The pretrained networks were trained using RGB images that have 3 channels, whereas HSI images contain 30 channels. The introduced dimensionality reduction approach enables a possibility to employ transfer learning techniques to train the segmentation models. The transfer learning can improve the models convergence by utilizing neural networks which already were pretrained on big datasets [16]. Jiao et al. [28] proposed an architecture fusing features from VGG16 encoder [54] and principal component analysis and Yu et al. [64] showed that dimensionality reduction (DR) blocks can be trained end-to-end in combination with the convolutional neural networks.

3.4.2 Research findings

This section presents the results from Publication I. In this work the approach similar to [64] was used. A series of 1×1 convolutional layers followed by ReLU was used to reduce the dimensionality of the spectral features. The architectures utilized in the experiments are based on SegNet and Dense-FCN. The SegNet based architectures (DR-SegNet) use the encoder pretrained on ImageNet [9], whereas the Dense-FCN based architectures (DR-DenseFCN) were not pretrained which allows the dimensionality of the output of the dimensionality reduction blocks to be tuned. The dimensionality reduction blocks in the case of Dense-FCN are also built using the principles of dense convolutional block. Additional experiments were performed without the dimensionality reduction blocks to assess the need for the dimensionality reduction layers before the base architecture. MC-Dropout was used to estimate epistemic uncertainty using $N_E = 100$ forward passes.

The architectures were trained in two stages: patchwise pretraining and full resolution fine-tuning. During the pretraining stage batches of three samples are sampled so that each sample is selected for the particular class. The mining of positive samples was also employed to handle the class imbalance related to the optic disc and macula segmentation tasks. The patches for blood vessels were selected uniformly, whereas the patches for optic disc and macula were selected so that they were centered around the mean of the true labels coordinates.

Table 3.8: Evaluation results. The best F1 scores are in bold.

Architecture	Vessels		Optic Disc		Macula	
	F1	IoU	F1	IoU	F1	IoU
DR-SegNet	0.8091	0.6802	**0.8947**	0.8356	**0.6566**	0.5291
RGB-SegNet	0.7925	0.6571	0.8802	0.8149	0.6033	0.4657
DR-2-Dense-FCN	**0.8243**	0.7019	0.7311	0.6257	0.3084	0.2414
HSI-Dense-FCN	0.7974	0.6647	0.7323	0.6202	0.2932	0.2282
RGB-Dense-FCN	0.8112	0.6840	0.7154	0.6153	0.1543	0.1070

Table 3.8 contains an evaluation of the performance of the proposed architectures. From the table it is clear that the blood vessels segmentation is the easiest task and segmenting the macula accurately is more difficult which is due to the lack of clear edges of the object. DR-SegNet architecture achieves the best performance for the optic disc and macula, whereas DR-DenseFCN with the spectral features reduced to two dimensional vectors achieves the best results for the vasculature segmentation task. Additionally, the evaluation results for the architecture (HSI-Dense-FCN) not pretrained on any data and without dimensionality reduction blocks are given. The results suggest that dimensionality reduction blocks help to more efficiently utilize spectral information leading to better segmentation performance.

The visualizations of the results are given in Figures 3.12 – 3.14. From the figures it is clear that the additional spectral information can help to localize the macula more accurately. It is also worth to note that the model trained using RGB images is more uncertain in the areas where labels overlap and the macula segmentation results contain artifacts where the macula and vasculature overlaps. The DR-SegNet shows less segmentation artifacts.

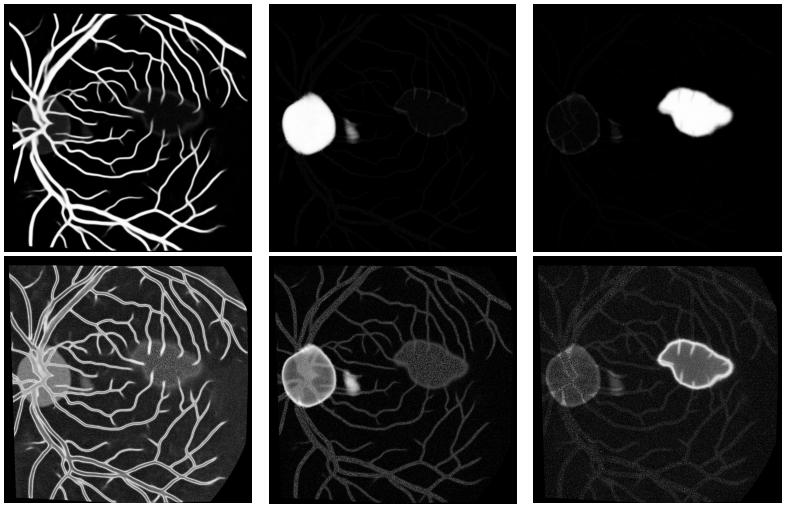


Figure 3.12: Top row: example segmentation results using DR-SegNet. Bottom row: standard deviations of the activations.

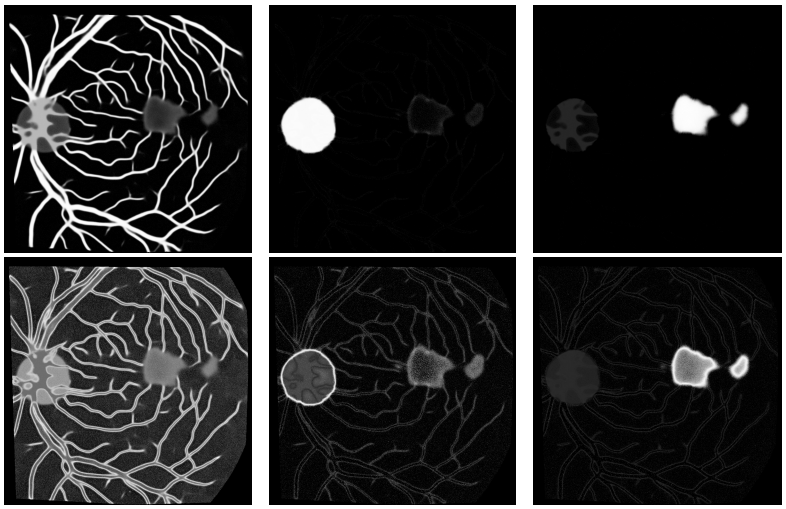


Figure 3.13: Top row: example segmentation results using RGB-SegNet. Bottom row: standard deviations of the activations.

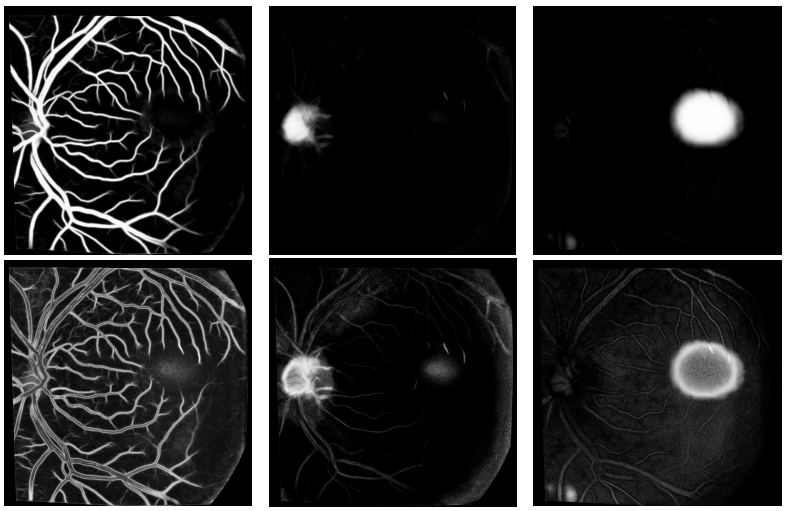


Figure 3.14: Top row: example segmentation results using DR-6-Dense-FCN. Bottom row: the epistemic uncertainties presented as standard deviations of the activations. Each column shows visualization for the tasks (from left to right): blood vessels, optic disc, macula.

Chapter IV

Discussion

4.1 Current results

This study focuses on the application of deep Bayesian approaches to the problem of fundus image segmentation. The fundus segmentation problem was studied from different perspectives including the landmark segmentation as well as DR lesion segmentation. In addition to different considered objects of interest, two modalities were considered: RGB and spectral images.

The proposed Bayesian baseline for artery-vein segmentation produced performance comparable to the previously published state-of-the-art results. The presented method uses the multi-label segmentation model and two different uncertainty quantification methods: MC-Dropout and SWAG. In this work, it is shown that the utilization of the weight averaging techniques can moderately improve the segmentation performance as well as the calibration of the model. It is shown that the highest uncertainties are concentrated near the optic disc, artery-vein crossings, and thin vessels. The high uncertainties near the optic disc and crossings can be explained by overlapping labels. The uncertain thin vessels are too far away from the optic disc and it is difficult to classify them into arteries and veins. It is shown that the weight averaging techniques can lead to vanishing aleatoric uncertainties. The utilization of aleatoric uncertainty inference can also be treated as a learned loss attenuation [32] and it seems that this attenuation plays lesser role when weight averaging is applied. The method proposed by Zhang et al. [65] still outperforms the methods proposed in this work. Nevertheless, the Bayesian methods proposed in this work keep the predictions more consistent, since the probabilities for blood vessels are inferred from the predicted probabilities for the arteries and veins. Thus, the objective of developing Bayesian baseline for retinal artery-vein segmentation using MC-Dropout and SWAG for uncertainty quantification is achieved.

This work extended the recent research on DR lesion segmentation on the uncertainty quantification of deep Bayesian models. The uncertainty validation procedure based on the idea of predicted misclassifications using the uncertainty estimates was introduced. The uncertainties estimated using Monte-Carlo dropout can clearly highlight problematic

image areas for the DR lesion segmentation. Nevertheless, straightforward utilization of the produced uncertainties to detect misclassifications has not provided satisfactory results. Apart from the previously mentioned problems, the proposed method suffers from the low sensitivity and significant miscalibration. The problem of low sensitivities is a common problem also reported in the recent studies. Thus, the objective of developing Bayesian baseline for diabetic retinopathy retinal lesion is achieved and the validation procedure is proposed.

Spectral imaging allows to capture additional information about the eye which potentially can improve the segmentation performance. In this work, it is shown that additional spectral information helps to moderately improve the retinal vasculature, optic disc and macula segmentation performance. The proposed method extended standard architectures with the dimensionality reduction layers and MC-Dropout for the epistemic uncertainty quantification. The utilization of the dimensionality reduction layers can help to adapt the standard architectures for the retinal HSI segmentation without significantly increasing the number of parameters of the models. The highest uncertainties highlighted unclear label edges and the areas of the overlapping objects. Thus, the objective of developing deep Bayesian models for hyperspectral fundus landmark segmentation is achieved. The comparison with RGB image segmentation is provided.

4.2 Future work

The low sensitivity is a common problem in many subproblems regarding retinal image segmentation including those sub-problems considered in this work. The issue is primarily due to the small size of the objects of interest such as microaneurysms or thin arteries and veins. Other objects such as the macula or soft exudates do not have clear edges. Another factor affecting the performance is related to unbalanced datasets where the objects of interest are underrepresented in comparison to the background classes. In medical image analysis and segmentation, the use of custom heuristic loss functions has become commonplace [30]. Nevertheless, they typically have hyperparameters that are difficult to tune. In this work, results outperforming the proposed baselines with cross-entropy were not achieved and, thus, are not published. Further research focused on evaluating these loss functions is needed before making any conclusions on their effectiveness. It is also interesting what kind of effects the loss functions have on the produced uncertainty estimates and model calibration.

In this work, the methods addressing the class imbalance problem are based on simple heuristics involving sampling the image patches near the objects of interest. An alternative approach is to use adaptive active learning methods which can utilize the uncertainty estimates to propose interesting patches. The future work can be focused on adapting those methods in order to address the class imbalance problem.

Another issue shown in this study is model miscalibration. There are a number of methods that address this problem. The traditional approaches are based on post-processing of training results to improve the calibration [18]. Thulasidasan et al. [57] showed that the calibration can be improved by applying the mix-up augmentation which blends the input samples. Seo et al. [52] modified the training procedure by adding an additional term to the loss function which penalizes the model for too overconfident or underconfident predictions. Seo et al. [52] also followed the Bayesian approach and used the

produced uncertainty estimates to calculate the penalizing miscalibration term. The future research can be focused on exploring the existing or new calibration methods to improve the fundus image segmentation methods.

Monte-Carlo based methods utilized in this work are computationally costly and require a number of forward passes to estimate the uncertainty of the outputs. Currently, methods for calculating the uncertainty using a single forward pass are being explored [60]. These methods could reduce the computation time for the inference. However, applying these methods to semantic segmentation in a straightforward manner can lead to a number of issues leading to poor uncertainty estimates [47, 59]. Thus, more research is needed to make the uncertainty estimation methods more practical for the fundus image segmentation problems.

Chapter V

Conclusion

In this work, the Bayesian multilabel baselines for the artery-vein segmentation are proposed. The impacts of weight averaging techniques on the segmentation performance, model calibration and produced uncertainties were assessed. The epistemic uncertainties were estimated using Monte-Carlo dropout and SWAG. It is shown that the weight averaging techniques can improve the overall performance but also lead to vanishing aleatoric uncertainties. SWAG also yields significantly lower epistemic uncertainty which can potentially lead to problems with the future risk analysis.

The deep Bayesian approach for diabetic retinopathy lesion segmentation is proposed. The analysis of the epistemic and aleatoric uncertainties is provided, and the calibration of the model is assessed. The extended lesion segmentation approach is presented which takes into account both the segmentation performance and the quality of uncertainty estimates. The proposed validation procedure suggests that the uncertainty estimates produced by Monte-Carlo dropout cannot be straightforwardly used to reliably estimate the misclassifications.

The spectral fundus image segmentation of retinal vasculature, optic disc and macula were studied. The problem is formulated as a multi-label segmentation problem. The formulation allows to form continuous segmentation maps for the retinal vasculature, optic disc and macula even in the case of overlapping labels. The neural network architectures for the segmentation of the landmarks are proposed. The proposed architectures are straightforward modifications of the existing semantic segmentation architectures. The modifications are formulated in a form of dimensionality reduction layers trained together with the rest of the architecture end-to-end. Different configurations of dimensionality reduction layers are analyzed and compared. It was shown that spectral information may give additional benefits in the landmark segmentation moderately improving the segmentation performance.

References

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* (2021).

[2] Abràmoff, M. D., Garvin, M. K., and Sonka, M. Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering 3* (2010), 169–208.

[3] Agurto, C., Joshi, V., Nemeth, S., Soliz, P., and Barriga, S. Detection of hypertensive retinopathy using vessel measurements and textural features. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2014), IEEE, pp. 5406–5409.

[4] Badawi, S., and Fraz, M. Multiloss function based deep convolutional neural network for segmentation of retinal vasculature into arterioles and venules. *BioMed Research International 2019* (04 2019), 1–17.

[5] Badrinarayanan, V., Kendall, A., and Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 12 (2017), 2481–2495.

[6] Budai, A., Odstrcilík, J., Kolár, R., Hornegger, J., Jan, J., Kubena, T., and Michelson, G. A public database for the evaluation of fundus image segmentation algorithms. *Investigative Ophthalmology & Visual Science 52* (2011), 1345–1345.

[7] Carmona, E. J., Rincón, M., García-Feijóo, J., and de-la Casa, J. M. Identification of the optic nerve head with genetic algorithms. *Artificial Intelligence in Medicine 43 3* (2008), 243–59.

[8] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordóñez-Varela, J.-R., Massin, P., Erginay, A., Charton, B., and Jc, K. Feedback on a publicly distributed image database: The MESSIDOR database. *Image Analysis & Stereology 33* (2014), 231–234.

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), IEEE, pp. 248–255.

[10] Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G., Kenton, Z., Smith, L., Alizadeh, M., de Kroon, A., and Gal, Y. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481* (2019).

[11] Fumero, F., Alayon, S., Sanchez, J. L., Sigut, J., and Gonzalez-Hernandez, M. RIM-ONE: An open retinal image database for optic nerve evaluation. In *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)* (2011), pp. 1–6.

[12] Fält, P., Hiltunen, J., Hauta-Kasari, M., Sorri, I., Kalesnykiene, V., Pietilä, J., and Uusitalo, H. Spectral imaging of the human retina and computationally determined optimal illuminants for diabetic retinopathy lesion detection. *Journal of Imaging Science and Technology 55*, 3 (2011), 253–263.

[13] Gal, Y. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.

[14] Gal, Y., and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning* (2016), pp. 1050–1059.

[15] Girard, F., Kavalec, C., and Cheriet, F. Joint segmentation and classification of retinal arteries/veins from fundus images. *Artificial Intelligence in Medicine 94* (2019), 96 – 109.

[16] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[17] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., and Webster, D. R. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA 316*, 22 (12 2016), 2402–2410.

[18] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning* (2017), PMLR, pp. 1321–1330.

[19] Hemelings, R., Elen, B., Stalmans, I., Van Keer, K., De Boever, P., and Blaschko, M. B. Artery-vein segmentation in fundus images using a fully convolutional network. *Computerized Medical Imaging and Graphics* (2019).

[20] Hoover, A., Kouznetsova, V., and Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging 19*, 3 (2000), 203–210.

[21] Hoover, A., Kouznetsova, V., and Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging 19*, 3 (2000), 203–210.

[22] Hu, Q., Abràmoff, M. D., and Garvin, M. K. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013* (Berlin, Heidelberg, 2013), K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds., Springer Berlin Heidelberg, pp. 436–443.

[23] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2261–2269.

[24] Issac, A., Sarathi, M. P., and Dutta, M. K. An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Computer Methods and Programs in Biomedicine 122*, 2 (2015), 229–244.

[25] Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. In *The Conference on Uncertainty in Artificial Intelligence* (2018).

[26] Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. What are Bayesian neural network posteriors really like? *arXiv preprint arXiv:2104.14421* (2021).

[27] Jegou, S., Drozdzal, M., Vázquez, D., Romero, A., and Bengio, Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. pp. 1175–1183.

[28] Jiao, L., Liang, M., Chen, H., Yang, S., Liu, H., and Cao, X. Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing 55*, 10 (Oct 2017), 5585–5599.

[29] Jogi. *Basic ophthalmology.* Jaypee Brothers Medical Publishers, New Delhi, India, 2008.

[30] Jun, M. Segmentation loss odyssey. *arXiv preprint arXiv:2005.13449* (2020).

[31] Kauppi, T., Kalesnykiene, V., Kämäräinen, J., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Uusitalo, H., Kälviäinen, H., and Pietilä, J. The DIARETDB1 diabetic retinopathy database and evaluation protocol. In *BMVC* (2007).

[32] Kendall, A., and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (2017), pp. 5574–5584.

[33] Laaksonen, L. *Spectral retinal image processing and analysis for ophthalmology.* PhD thesis, Lappeenranta University of Technology, 2016.

[34] Laves, M.-H., Ihler, S., Fast, J. F., Kahrs, L. A., and Ortmaier, T. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning* (2020), PMLR, pp. 393–412.

[35] Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports 7* (12 2017).

[36] Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., and Fu, H. Applications of deep learning in fundus images: A review. *Medical Image Analysis 69* (2021), 101971.

[37] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2980–2988.

[38] Ma, Y.-A., Chen, T., and Fox, E. B. A complete recipe for stochastic gradient mcmc. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA, USA, 2015), NIPS'15, MIT Press, p. 2917–2925.

[39] Maddox, W., Garipov, T., Izmailov, P., Vetrov, D. P., and Wilson, A. G. A simple baseline for Bayesian uncertainty in deep learning. In *NeurIPS* (2019).

[40] Manikis, G. C., Sakkalis, V., Zabulis, X., Karamaounas, P., Triantafyllou, A., Douma, S., Zamboulis, C., and Marias, K. An image analysis framework for the early assessment of hypertensive retinopathy signs. In *2011 E-Health and Bioengineering Conference (EHB)* (2011), pp. 1–6.

[41] Medeiros, F. A., Jammal, A. A., and Thompson, A. C. From machine to machine: An OCT-Trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology 126*, 4 (2019), 513–521.

[42] Mobiny, A., Nguyen, H., Moulik, S., Garg, N., and Wu, C. DropConnect is effective in modeling uncertainty of Bayesian deep networks. *ArXiv abs/1906.04569* (2019).

[43] Mookiah, M. R. K., Hogg, S., MacGillivray, T. J., Prathiba, V., Pradeepa, R., Mohan, V., Anjana, R. M., Doney, A. S., Palmer, C. N., and Trucco, E. A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis 68* (2021), 101905.

[44] Noronha, K., Navya, K., and Nayak, K. P. Support system for the automated detection of hypertensive retinopathy using fundus images. In *International Conference on Electronic Design and Signal Processing (ICEDSP)* (2012), pp. 7–11.

[45] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data 3*, 3 (2018).

[46] Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al. IDRiD: Diabetic retinopathy–segmentation and grading challenge. *Medical Image Analysis 59* (2020), 101561.

[47] Postels, J., Segu, M., Sun, T., Van Gool, L., Yu, F., and Tombari, F. On the practicality of deterministic epistemic uncertainty. *arXiv preprint arXiv:2107.00649* (2021).

[48] Pratt, H., Coenen, F., Harding, S. P., Broadbent, D. M., and Zheng, Y. Feature visualisation of classification of diabetic retinopathy using a convolutional neural network. In *CEUR Workshop Proceedings* (2019), vol. 2429, pp. 23–29.

[49] Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Cham, 2015), N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Springer International Publishing, pp. 234–241.

[50] Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).

[51] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 618–626.

[52] Seo, S., Seo, P. H., and Han, B. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 9022–9030.

[53] Sharma, A., Agrawal, M., Roy, S. D., and Gupta, V. *Automatic glaucoma diagnosis in digital fundus images using deep CNNs*. Springer Singapore, Singapore, 2020, pp. 37–52.

[54] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[55] Sreng, S., Maneerat, N., Hamamoto, K., and Win, K. Y. Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images. *Applied Sciences 10*, 14 (2020), 4916.

[56] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 56 (2014), 1929–1958.

[57] Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *arXiv preprint arXiv:1905.11001* (2019).

[58] Triwijoyo, B. K., Budiharto, W., and Abdurachman, E. The classification of hypertensive retinopathy using convolutional neural network. *Procedia Computer Science 116* (2017), 166–173. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

[59] van Amersfoort, J., Smith, L., Jesson, A., Key, O., and Gal, Y. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409* (2021).

[60] Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning* (2020), PMLR, pp. 9690–9700.

[61] Wang, L., and Zhao, C. *Hyperspectral Image Processing*. Springer, 2015.

[62] Wei, Q., Li, X., Yu, W., Zhang, X., Zhang, Y., Hu, B., Mo, B., Gong, D., Chen, N., Ding, D., and xin Chen, Y. Learn to segment retinal lesions and beyond. *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), 7403–7410.

[63] Yan, Z., Han, X., Wang, C., Qiu, Y., Xiong, Z., and Cui, S. 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.

[64] Yu, S., Jia, S., and Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing 219* (2017), 88–98.

[65] Zhang, S., Zheng, R., Luo, Y., Wang, X., Mao, J., Roberts, C. J., and Sun, M. Simultaneous arteriole and venule segmentation of dual-modal fundus images using a multi-task cascade network. *IEEE Access 7* (2019), 57561–57573.

[66] Zhou, M., Jin, K., Wang, S., Ye, J., and Qian, D. Color retinal image enhancement based on luminosity and contrast adjustment. *IEEE Transactions on Biomedical Engineering 65*, 3 (2018), 521–527.

[67] Zhou, T., Ruan, S., and Canu, S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array 3-4* (2019), 100004.

Publication I

Garifullin A., Kööbi P., Ylitepsa P., Ådjers K., Hauta-Kasari M., Uusitalo H., and Lensu L.
Hyperspectral Image Segmentation of Retinal Vasculature, Optic Disc and Macula

Reprinted with permission from
2018 Digital Image Computing: Techniques and Applications (DICTA)
pp. 1-5, 2018.

Hyperspectral Image Segmentation of Retinal Vasculature, Optic Disc and Macula

Azat Garifullin*, Peeter Kööbi†§, Pasi Ylitepsa†, Kati Ådjers†§, Markku Hauta-Kasari‡, Hannu Uusitalo†§ and Lasse Lensu*
* Machine Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta University of Technology, P.O. Box 20, 53851, Lappeenranta, Finland
† SILK, Department of Ophthalmology, ARVO F313, 33014, University of Tampere, Finland
‡ University of Eastern Finland, School of Computing, P.O. Box 111. 80101, Joensuu, Finland
§ Tays Eye Center, Tampere University Hospital, P.O. Box 2000, 33520, Tampere, Finland
Emails: {azat.garifullin, lasse.lensu}@lut.fi, {peeter.koobi, pasi.ylitepsa, hannu.uusitalo}@uta.fi, kati.adjers@pshp.fi, markku.hauta-kasari@uef.fi

***Abstract*—The most common approach for retinal imaging is the eye fundus photography which usually results in RGB images. Recent studies show that the additional spectral information provides useful features for automatic retinal image analysis. The current work extends recent research on the joint segmentation of retinal vasculature, optic disc and macula which often appears in different retinal image analysis tasks. Fully convolutional neural networks are utilized to solve the segmentation problem. It is shown that the network architectures can be effectively modified for the spectral data and the utilization of spectral information provides moderate improvements in retinal image segmentation.**

I. Introduction

Retinal diseases like diabetic retinopathy, age-related macular degeneration and glaucoma are the leading causes of blindness worldwide [1]. A diagnostic process to recognize the signs of these diseases is traditionally based on retinal RGB images. Recent developments in machine vision technologies provide various methods for automatic RGB images analysis. In [2], it has been shown that additional spectral features introduced to machine learning methods may improve the performance of lesions classification and enable new ways to analyze the retinal tissue layers. Thus, spectral retinal imaging can be treated as a useful alternative to traditional color fundus imaging. This work studies the joint segmentation of retinal vasculature, optic disc and macula for hyperspectral retinal images.

Deep convolutional neural networks is a common trend in both retinal and hyperspectral image analysis. Deep architectures similar to U-Net are extensively used for the vasculature [3], optic disc and cup segmentation [4] purposes. Tan et al. [5] studied the segmentation of optic disc, fovea and retinal vasculature using a single model trained on the DRIVE dataset [6]. All the mentioned approaches have been tested on RGB images. To the best of the authors' knowledge, this work is the first work which studies spectral retinal image segmentation using deep fully-convolutional neural networks.

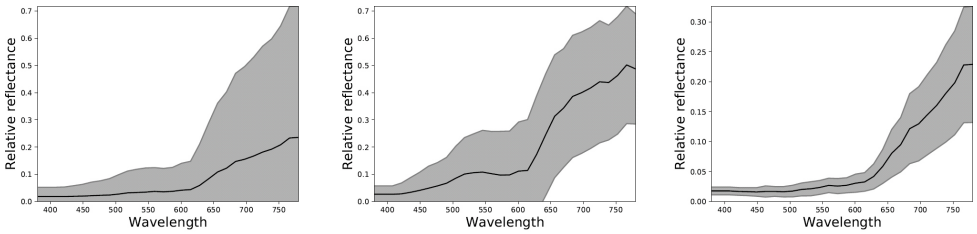
One way to build deep architectures for hyperspectral image (HSI) segmentation is to combine dimensionality reduction methods and convolutional neural networks. Jiao et al. [7] proposed to use the feature fusion from VGG16 encoder [8] and principal component analysis for HSI segmentation. Yu et al. [9] showed that dimensionality reduction blocks can be trained end-to-end altogether with the convolutional neural networks. Other approaches are based on 3D convolutional neural networks [10], [11] that can effectively extract both spatial and spectral features. However, it is more difficult to scale the 3D convolutions on high-resolution images. In this paper, we followed ideas similar to Yu et al. [9] and Jiao et al. [7], and adapted SegNet [12] and dense fully-convolutional neural networks (Dense-FCNs) [13] for the spectral retinal segmentation task. These architectures are trained and evaluated by using a spectral image dataset with manual ground truth for the vasculature, optic disc and macula.

II. Spectral retinal image dataset

Several spectral fundus imaging setups with different optical principles have been proposed. A typical hyperspectral imaging setup is an adapted fundus camera with a light source with a broadband illumination, and a spectral device for selecting a spectral band. Fält et al. [14] modified a Canon CR5-45NM fundus camera to spectral fundus camera by replacing the standard light source with a fibre optic illuminator including a halogen lamp with the illumination spectrum from 380 to 780 nm and 30 interference filters with 10 nm interval were used for the wavelength selection. As the detector, a grayscale charge-coupled device (CCD) camera with the sensor array size of 2048×2048 pixels and 2×2 binning was used.

The resulted dataset is a set of 1024×1024 images with the 30 channels where each channel corresponds to the certain wavelength. For each image in the dataset, the field-of-view (FOV) mask is provided. The FOV masks are binary images where the white areas correspond to regions of the fundus of the eye. Manual segmentation masks for the vasculature, optic disc and macula are available (Fig. 2). The dataset consists of 55 spectral retinal images acquired from patients with diabetic retinopathy: randomly selected 25 images are in the training set, and the rest are used as the testing set. Thus, the amount of data in the dataset is comparable to the amount

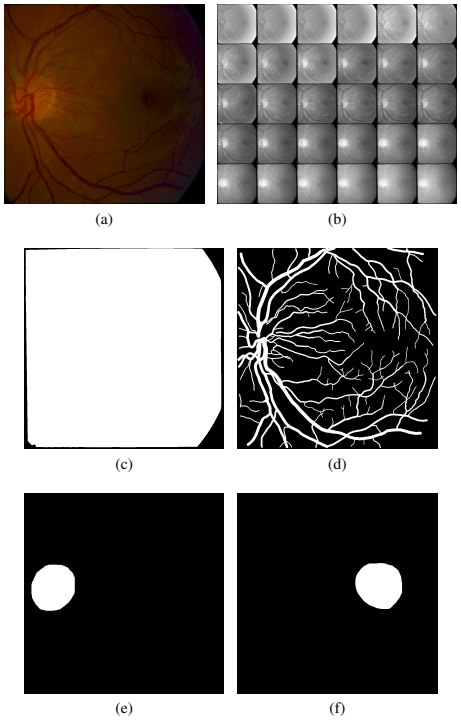
978-1-5386-6602-9/18/$31.00 ©2018 IEEE



(a) (b) (c)

Fig. 1. Visualization of the mean spectrum and 3σ range of (a) blood vessels, (b) optic disc and (c) macula

of labelled data in DRIVE [6] and STARE [15] datasets which are typically used for blood vessel segmentation algorithms benchmarking. Examples of the mean and variation of the spectra are presented in Fig. 1.



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

Fig. 2. (a) An RGB image from the spectral retinal image dataset. (b) Montage of the channel images composing an example hyperspectral retinal image. The image was normalized for the visualization purpose. (c) FOV mask and the corresponding segmentation masks for the (d) vessels, (e) optic disc and (f) macula.

III. Semantic segmentation architectures

A. SegNet

SegNet is an encoder-decoder architecture for semantic segmentation. In this work, we used basic-SegNet with small modifications for the HSI segmentation. The modified architecture consists of a dimensionality reduction block, encoder and decoder. The dimensionality reduction block is a sequence of blocks consisting of 1×1 convolutional layer and rectified linear unit (ReLU) activation. As the encoder, VGG16 pre-trained on ImageNet [16] was used. The decoder is a sequence of transposed convolutions and convolutional layers followed by batch normalization (BN), ReLU activation and dropout. The scheme of the architecture is shown in Fig. 3.

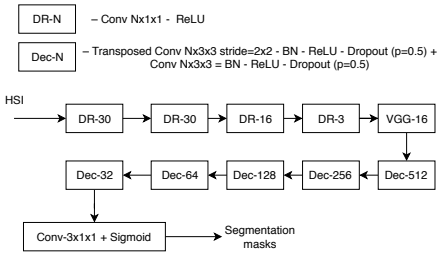


Fig. 3. DR-SegNet architecture.

B. Dense-FCN

It has been shown that Dense-FCNs have less parameters and may outperform the SegNet architecture in a variety of different segmentation tasks [13]. Here we adapted the Dense-FCN architecture for the retinal HSI segmentation task.

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

The main concept of Dense-FCN is similar to SegNet in a sense that the input is first compressed to a hidden

representation by the downsampling part, and then the segmentation masks are recovered by an upsampling part. The downsampling part consists of DCBs and downsampling transitions (DT) with skip connections to the upsampling part. The upsampling part consists of DCBs and upsampling transitions (UT). The scheme of the utilized architecture is given in Fig. 4.

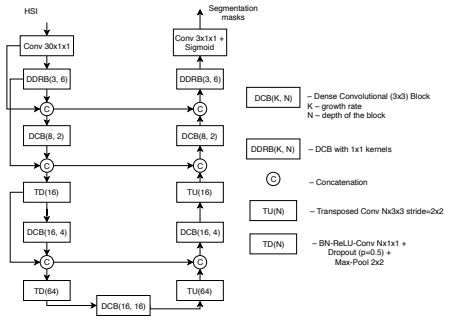
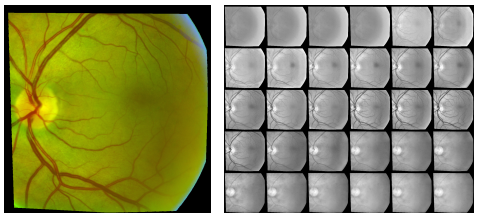


Fig. 4. DR-6-Dense-FCN architecture.

C. Image preprocessing

First, each channel of a spectral image is normalized to values between 0 and 255. After the normalization step, contrast limited adaptive histogram equalization [17] with the clip limit of 2 and the grid size of 8×8 is applied to each channel of the input. The described preprocessing scheme was applied to both RGB and spectral images. The preprocessing scheme was used to reduce the effects of uneven illumination fields of the channel images and indirectly affect the inter-person variation in the limited dataset. The scheme was found to improve the convergence and performance of the trained models. Examples of the preprocessed RGB and spectral images are given in Fig. 5.



(a) (b)

Fig. 5. Preprocessed (a) RGB and (b) spectral images.

D. Training details

Dense-FCNs are pretrained for 50 epochs with 500 steps per epoch on random patches 512×512 with the batch size equal to 3. Each batch consists of examples for blood vessels, optic disc, and macula. The vasculature examples are sampled uniformly, whereas the patch centers of the examples of the optic disc and macula are sampled from the normal distribution the parameters of which are estimated as the sample mean and covariance of true label coordinates.

The common training step for both architectures is fine-tuning on full size images. The number of epochs used was 50 for Dense-FCN and 100 for SegNet.

In both cases, the weights were initialized using HeNormal [18]. Binary cross-entropy was used as the loss function. In addition to dropout, $l2$ regularization with the weight decay factor 10^{-4} was used. As the optimizer, Adadelta [19] with the learning rate $l = 1$ and the decay rate $\rho = 0.95$ was used for the both pretraining and fine-tuning. The learning rate was dropped by a factor of 10 if the training loss was not decreased by 0.005 for 5 epochs. Data augmentation through flipping, reflecting and rescaling (with scale rates 0.8 and 1.2) was applied in both cases. The parameter values were determined empirically based on initial experiments.

IV. Experiments and results

The trained networks were evaluated on the full size images from the testing set using Monte Carlo dropout [20] in the test phase with 100 forward passes. The standard F1 measure, intersection over union (IoU) for each class and mean IoU over the classes were used as the evaluation metrics and they are presented in Table I. The evaluation metrics were calculated only inside the FOV. In order to distinguish between the RGB and spectral architectures, the DR prefix was added for the architectures for spectral images and the RGB prefix for the architectures for RGB images. The Dense-FCN architecture was also tested for the hyperspectral images without the dimensionality reduction layers (HSI prefix) and with a different number of output channels of the DR layers (DR-6 means 6 output channels).

From the table, it is clear that the vessel segmentation is the easiest task for all the architectures, since F1 score is comparable in all the cases, whereas the macular region is the most difficult to segment. The latter can be explained by the fact that there are normally no clearly defined structural characteristic in the macula. Furthermore, there are numerous images where the macula is only partly visible, and these images happen to be present only in the test set, because of what all the considered models were unable to generalize to a partly visible macula. Another interesting fact is that the region may have different shapes in different spectral channels.

The VGG encoder pretrained on ImageNet allows to significantly improve the performance on the optic disc and macula segmentation tasks. In the case of Dense-FCNs, it is difficult to achieve satisfactory performance for both the optic disc and macula segmentation, and the performance depends on the number of output channels of the dimensionality reduction layers. The segmentation results for the image presented in Fig. 5 are presented in Fig. 6 – 8.

Comparing the segmentation results for hyperspectral and RGB images in Fig. 6 and Fig. 7 shows that in some cases the

TABLE I
EVALUATION RESULTS. THE BEST F1 SCORES ARE IN BOLD.

Architecture	Vessels		Optic Disc		Macula		Mean IoU	# Parameters
	F1	IoU	F1	IoU	F1	IoU		
DR-SegNet	0.8091	0.6802	**0.8947**	0.8356	**0.6566**	0.5291	0.6816	21795786
RGB-SegNet	0.7925	0.6571	0.8802	0.8149	0.6033	0.4657	0.6458	21793379
DR-1-Dense-FCN	0.8125	0.6853	0.7128	0.6122	0.3962	0.3223	0.5399	730028
DR-2-Dense-FCN	**0.8243**	0.7019	0.7311	0.6257	0.3084	0.2414	0.5230	730383
DR-3-Dense-FCN	0.8069	0.6776	0.6986	0.5947	0.2822	0.2120	0.4948	730738
DR-4-Dense-FCN	0.8200	0.6958	0.6880	0.5867	0.3389	0.2665	0.5163	731093
DR-5-Dense-FCN	0.8006	0.6772	0.6977	0.5949	0.3843	0.2954	0.5225	731448
DR-6-Dense-FCN	0.8021	0.6714	0.7427	0.6394	0.4244	0.3402	0.5503	731803
HSI-Dense-FCN	0.7974	0.6647	0.7323	0.6202	0.2932	0.2282	0.5043	745837
RGB-Dense-FCN	0.8112	0.6840	0.7154	0.6153	0.1543	0.1070	0.4688	729043

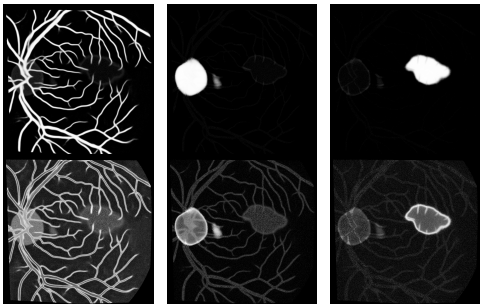


Fig. 6. Top row: example segmentation results with DR-SegNet. Bottom row: standard deviations of the activations.

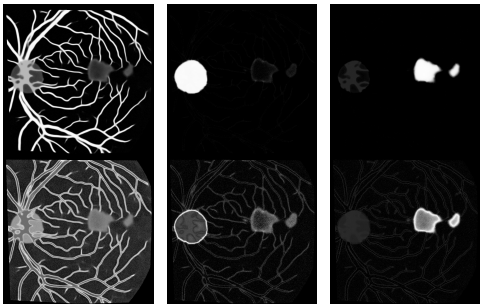


Fig. 7. Top row: example segmentation results with RGB-SegNet. Bottom row: standard deviations of the activations.

models trained on RGB fail to localize the macula properly, whereas the utilization of spectral information may help to avoid such problems. It is worth to mention that in both cases the segmentation artefacts are present in areas where the labels overlap. This is mainly caused by the VGG encoder pretrained on ImageNet, since there were no signs of such artefacts when it was not used or used without pretraining. It is also clear that the results obtained with DR-SegNet show less artefacts compared to RGB-SegNet.

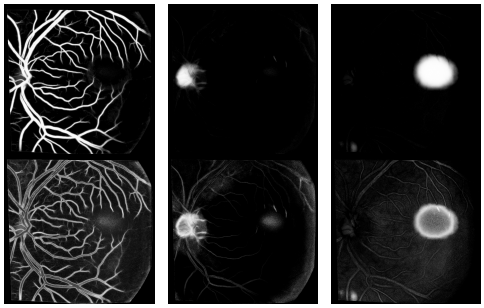


Fig. 8. Top row: example segmentation results with DR-6-Dense-FCN. Bottom row: standard deviations of the activations.

The dimensionality reduction layers allowed to adapt standard convolutional architecures for the hyperspectral image segmentation without significantly increasing the amount of parameters of the model. In addition, the utilization of the dimensionality reduction layers may slightly improve the performance of vessels and macula segmentation. In Fig. 9, the outputs of the dimensionality reduction layers for the both spectral architectures are shown.

In Fig. 9 one can see that the models try to emphasize areas where certain labels are most visible. For example, in the first image of the bottom row the macula is clearly seen, whereas in the second and third images the optic disc is visible more clear compared to the first image. Nevertheless, training the dimensionality reduction layers is a challenging task, and it is not always possible to train them to extract useful features. In the case of DR-SegNet, the pretrained VGG encoder makes the training easier. However, if the network is trained from the scratch, it is difficult to get results comparable to the results obtained with a pretrained model. We also tried to add skip connections to DR-SegNet in a manner similar to DR-Dense-FCNs, but in the case of DR-SegNet, it just confuses the model

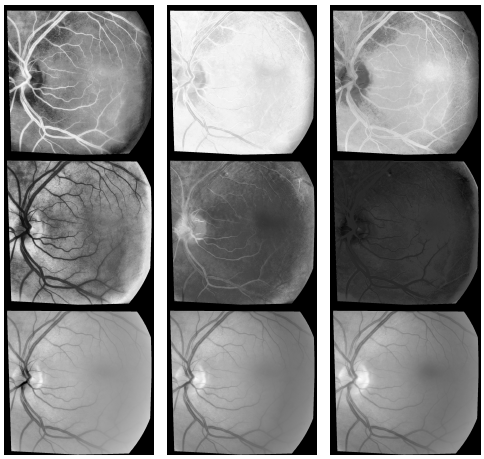


Fig. 9. Visualization of dimensionality reduction layers. The top two rows illustrate the results for DR-6-Dense-FCN and the bottom row for DR-SegNet.

more and the performance decreases, whereas in the case of DR-Dense-FCNs, it boosts performance significantly.

Previously, Laaksonen [2] has shown that diabetic retinopathy lesions classification algorithms trained on spectral data outperform algorithms trained on RGB data. Furthermore, spectral information can also be utilized for the histological analysis of fundus images [2]. From the presented results, it also clear that the utilization of additional spectral information may also improve the segmentation results compared to RGB images.

V. Conclusions

In this work, multilabel segmentation of retinal vasculature, optic disc and macula for spectral retinal images was studied. It was shown that spectral information may give additional advantages in optic disc and macula segmentation moderately improving the segmentation performance.

The results also show that it is necessary to study more the dimensionality reduction layers to find a way to train them effectively on small datasets. The future work will be concentrated on further improvements of training the Dense-FCN architecture in order to achieve comparable performance with the architectures pretrained on ImageNet. Another direction of the future work is the development of a gold standard based on label data from multiple experts.

Acknowledgements

The authors wish to thank CSC for the computational resources for some of the experiments.

References

[1] "World Health Organization: causes of blindness and visual impairment," http://www.who.int/blindness/causes/en/, accessed: 2018-06-09.

[2] L. Laaksonen, "Spectral retinal image processing and analysis for ophthalmology," Ph.D. dissertation, Lappeenranta University of Technology, 2016.

[3] L. Giancardo, K. Roberts, and Z. Zhao, "Representation learning for retinal vasculature embeddings," in *Fetal, Infant and Ophthalmic Medical Image Analysis*, M. J. Cardoso, T. Arbel, A. Melbourne, H. Bogunovic, P. Moeskops, X. Chen, E. Schwartz, M. Garvin, E. Robinson, E. Trucco, M. Ebner, Y. Xu, A. Makropoulos, A. Desjardin, and T. Vercauteren, Eds. Cham: Springer International Publishing, 2017, pp. 243–250.

[4] 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*, M. J. Cardoso, T. Arbel, A. Melbourne, H. Bogunovic, P. Moeskops, X. Chen, E. Schwartz, M. Garvin, E. Robinson, E. Trucco, M. Ebner, Y. Xu, A. Makropoulos, A. Desjardin, and T. Vercauteren, Eds. Cham: Springer International Publishing, 2017, pp. 168–176.

[5] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, "Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network," *Journal of Computational Science*, vol. 20, pp. 70–79, 2017.

[6] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[7] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5585–5599, Oct 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, 2017.

[10] M. He, B. Li, and H. Chen, "Multi-scale 3d deep convolutional neural network for hyperspectral image classification," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 3904–3908.

[11] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2018.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[13] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and 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.

[14] P. Fält, J. Hiltunen, M. Hauta-Kasari, I. Sorri, V. Kalesnykiene, J. Pietilä, and H. Uusitalo, "Spectral Imaging of the Human Retina and Computationally Determined Optimal Illuminants for Diabetic Retinopathy Lesion Detection," *Journal of Imaging Science and Technology*, vol. 55, no. 3, pp. 253–263, 2011.

[15] M. Michael Goldbaum, "Structured analysis of the retina," http://www.cecas.clemson.edu/ ahoover/stare, 2003, online.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[17] K. Zuiderveld, "Graphics gems iv," P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press Professional, Inc., 1994, ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485. [Online]. Available: http://dl.acm.org/citation.cfm?id=180895.180940

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[19] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[20] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

Publication II

Garifullin, A., Lensu, L., and Uusitalo, H.
On the Uncertainty of Retinal Artery-Vein Classification with Dense Fully-Convolutional Neural Networks

Reprinted with permission from
Advanced Concepts for Intelligent Vision Systems (ACIVS)
pp. 87–98, 2020.

On the Uncertainty of Retinal Artery-Vein Classification with Dense Fully-Convolutional Neural Networks

Azat Garifullin[1(✉)], Lasse Lensu[1], and Hannu Uusitalo[2,3]

[1] LUT University, P.O. Box 20, 53851 Lappeenranta, Finland
{azat.garifullin,lasse.lensu}@lut.fi
[2] SILK, Department of Ophthalmology, Tampere University, ARVO F313, 33014 Tampere, Finland
hannu.uusitalo@tuni.fi
[3] Tays Eye Center, Tampere University Hospital, P.O. Box 2000, 33520 Tampere, Finland

Abstract. Retinal imaging is a valuable tool in diagnosing many eye diseases but offers opportunities to have a direct view to central nervous system and its blood vessels. The accurate measurement of the characteristics of retinal vessels allows not only analysis of retinal diseases but also many systemic diseases like diabetes and other cardiovascular or cerebrovascular diseases. This analysis benefits from precise blood vessel characterization. Automatic machine learning methods are typically trained in the supervised manner where a training set with ground truth data is available. Due to difficulties in precise pixelwise labeling, the question of the reliability of a trained model arises. This paper addresses this question using Bayesian deep learning and extends recent research on the uncertainty quantification of retinal vasculature and artery-vein classification. It is shown that state-of-the-art results can be achieved by using the trained model. An analysis of the predictions for cases where the class labels are unavailable is given.

Keywords: Bayesian deep learning · Blood vessels segmentation · Artery-vein classification

1 Introduction

A number of eye and systemic diseases influence the vasculature of the retina in different ways. The blood vessel characteristics in retinal images may provide visible evidence about numerous diseases such as hypertensive retinopathy, diabetic retinopathy, as well as other cardio- and cerebrovascular diseases [12]. The related characteristics include the shape and size of retinal vessels, arteriovenous ratio and arteriovenous crossing [14]. These characteristics may be obtained by using blood vessel segmentation masks produced by automatic machine learning techniques [5].

J. Blanc-Talon et al. (Eds.): ACIVS 2020, LNCS 12002, pp. 87–98, 2020.
https://doi.org/10.1007/978-3-030-40605-9_8

The topic of blood vessels segmentation is well studied by the community [1]. However, the artery-vein (AV) classification task remains challenging not only for machines, but also for humans. Despite the fact that discriminative features based on color and geometry are described, it is still difficult to distinguish arteries from veins [14] due to imperfect imaging conditions and limited visibility of the retinal blood vessels.

Recently, deep convolutional neural networks have become a common trend for retinal vasculature segmentation and AV classification because of the ability to automatically learn meaningful features. Welikala et al. [16] proposed a method based on a convolutional neural network (CNN) classifying arteries and veins in a patch-wise manner. The authors considered the problem as a multi-class classification task placing a softmax layer at the end of the network. The UK Biobank database was used from which 100 images were labeled and classification accuracy of 82.26% for arteries and veins was reported. Girard et al. [5] proposed to use a modified U-Net [15] with likelihood score propagation in the minimum spanning tree effectively utilizing information about the global vessel topology. The approach was tested on the DRIVE data set [8] and it achieved 94.93% accuracy for the AV classification. Badawi et al. [2] proposed to train a CNN with multiloss function consisting of pixelwise cross entropy loss and segment-level loss to overcome training issues appearing because of inconsistent thickness of blood vessels. The authors also created a new data set consisting of labeled subsets of EPIC and MESSIDOR [3] data sets and classification accuracy of 96.5% was reported. Hemelings et al. [7] applied the U-Net architecture for the task of AV classification stating the problem as a multi-class classification problem predicting labels for four classes (background, vein, artery, and unknown) with classification accuracy of 94.42% and 94.11% for arteries and veins, respectively. Zhang et al. [18] proposed cascade refined U-net which modifies the original model with multi-scale loss training and includes sub-networks for simultaneous AV and blood vessel segmentation. The authors achieved 97.27% arteriovenous classification accuracy evaluated on the automatically detected vessels.

In this work, a multi-label classification approach is considered with the uncertainty quantification experiments presented. Our approach is most similar to the method proposed by Zhang et al. [18] in a way how three-component loss is used. The main difference is that in this work, classification of arteries and veins are not conditioned on blood vessel predictions, but vessel labels are conditioned on arteries and veins. Using the multi-label classification approach, there is no need to separately model the AV crossings and background. To the best of authors' knowledge, this work is the first presenting uncertainty quantification experiments for the of AV classification. For the experiments, the RITE data set is utilized.

2 Data and Methods

2.1 DRIVE and RITE Data Sets

The DRIVE database is a common benchmark for the retinal blood vessel segmentation task [8]. It contains 20 train and 20 test images with two sets of manual blood vessel segmentations. The RITE data set [9] extends DRIVE with an AV reference standard containing four types of labels: arteries (red), veins (blue), overlapping (green), and uncertain vessels (white). An example test image is shown in Fig. 1.

2.2 AV Classification

Let f be a model with parameters $\boldsymbol{\theta}$ that maps an input image $\mathbf{x}$ to a map of logits with the same spatial dimensionality as the original image:

$$\hat{\mathbf{y}} = f(\mathbf{x}, \boldsymbol{\theta}). \tag{1}$$

Given predicted logits $\hat{\mathbf{y}} = [\hat{y}_{\text{artery}}\ \hat{y}_{\text{vein}}]$, probabilities of assigning labels to arteries and veins can be calculated as follows:

$$p_{\text{artery}} = \text{sigmoid}(\hat{y}_{\text{artery}}), \tag{2}$$

$$p_{\text{vein}} = \text{sigmoid}(\hat{y}_{\text{vein}}). \tag{3}$$

In the multi-label setup, the same pixel can be classified with both artery and vein labels, which is meaningful in the case of AV crossings. A vessel probability label can then be naturally inferred by a simple formula:

$$p_{\text{vessel}} = p_{\text{artery}} + p_{\text{vein}} - p_{\text{artery}} p_{\text{vein}}. \tag{4}$$

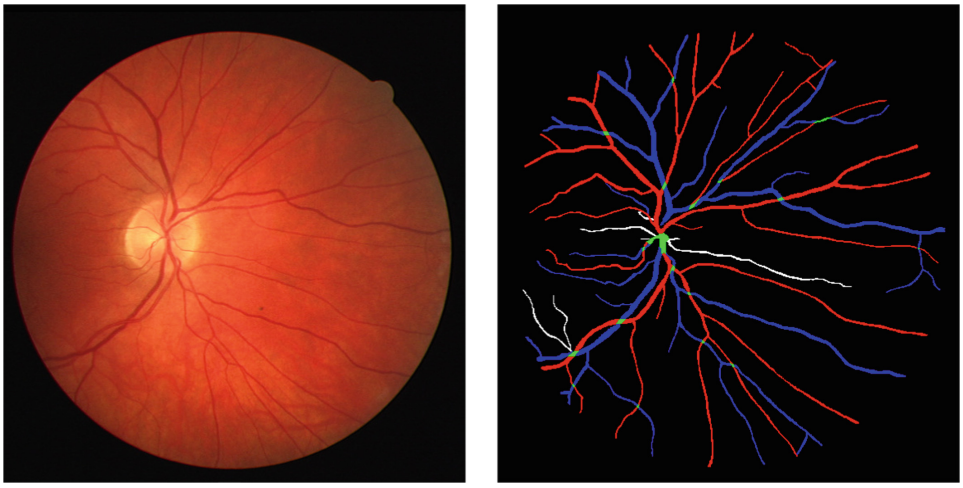
Since the data set contains the masks for both the AV classification and blood vessel segmentation, it is possible to state the following optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\mathcal{L}_{\text{artery}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{vein}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{vessel}}(\boldsymbol{\theta}) \right], \tag{5}$$

where $\mathcal{L}$ denotes the binary cross entropy loss for the corresponding labels. This way even if the labels for arteries and veins are not given for uncertain vessel labels, it is possible to enforce a model to predict correct labels for the blood vessels.

2.3 Aleatoric and Epistemic Uncertainties

The approach described in the previous section gives only point estimates for the label probabilities and the model parameters are considered to be deterministic. In order to better capture imperfect data labeling and image noise, one can consider the model outputs and the parameters to be random variables. The first approach captures the heteroscedastic aleatoric uncertainty that depends



(a) (b)

Fig. 1. The RITE data set: (a) An example test image and (b) corresponding artery-vein reference standard. (Color figure online)

on the input data, whereas the second represents the epistemic uncertainty that models a distribution of the learned parameters. More detailed explanations for the uncertainties can be found in [13] and [4]. In this work, a brief explanation for the AV classification task is given below.

Aleatoric uncertainty can be captured by modifying the original model to predict the mean and standard deviations of logits:

$$[\hat{\mathbf{y}}, \boldsymbol{\sigma}] = f(\mathbf{x}, \boldsymbol{\theta}). \tag{6}$$

In order to predict standard deviations, a second layer similar and parallel to the one used for logits is added to the output of the network. In order to ensure that the predicted standard deviations are positive, an additional absolute value activation is added to the output of the layer. The probabilities of the labels can then be calculated as follows:

$$\hat{\mathbf{p}} = \text{sigmoid}(\hat{\mathbf{y}} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

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

The main inference scheme for AV remains the same with the exception that instead of a point estimate, the model now yields N_A samples that are then used to calculate the loss (5). The final minimized loss is just an average over the predicted losses for each sample.

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

$$p(\mathbf{y} \mid \mathbf{x}, \mathcal{D}) = \int p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}, \tag{8}$$

where $\mathcal{D}$ denotes a data set of input-output pairs. Typically, the parameter's posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ for complex models such as deep neural networks is intractable and variational approximations are used. The posterior in (8) can be replaced by a simpler distribution $q(\boldsymbol{\theta})$ and the training procedure can then be formulated as the minimization of the Kullback-Leibler divergence between the true posterior and the approximation.

In this work, the model f is parameterized as a dense fully-convolutional network (Dense-FCN) and Monte-Carlo dropout [4] is used for the variational approximation. The description of the utilized architecture is given below.

2.4 Architecture

The architecture utilized in this work is a Dense-FCN. 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 [11]. Here we adapt the Dense-FCN architecture for the AV classification tasks.

The main building block of Dense-FCN 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), ReLU, convolution and dropout $p = 0.5$ layers resulting in g feature maps (growth rate).

The main concept of Dense-FCN is similar to other encoder-decoder architectures in the sense that the input is first compressed to a hidden representation by the downsampling part, and then 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 given in Fig. 2. The architectural parameters used are given below:

- Growth rate for all DCBs: $g = 16$.
- Downsampling path consists of five DCBs with depths $D_{\text{down}} = [4, 5, 7, 10, 12, 15]$.
- 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. 2.

2.5 Image Preprocessing

It was noticed in the experimental part of the work that simple preprocessing involving contrast enhancement and channel normalization improves the convergence and performance of the trained models. First, contrast-limited adaptive histogram equalization [19] with the clip limit of 2 and the grid size of 8×8 is applied and then each image channel is normalized to values between 0 and 255. The preprocessing scheme was used to reduce the effects of uneven illumination fields of the channel images (Fig. 3).

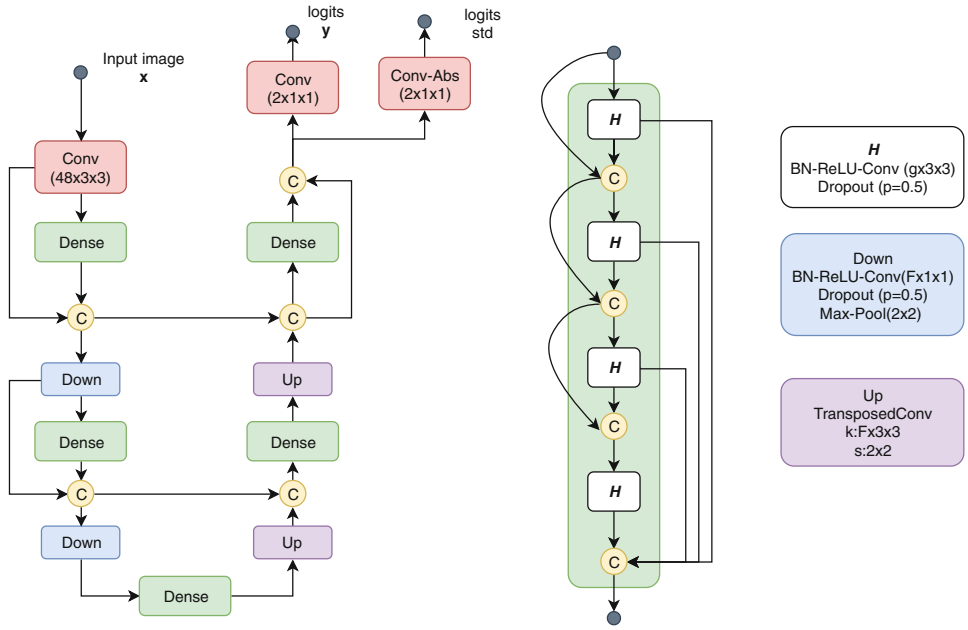


Fig. 2. Dense-FCN architecture: *Dense* stands for a DCB; C is a tensor concatenation; H is a block consisting of BN, ReLU and a convolutional layer with growth rate g; *Down* is a transition down block with F output feature maps; *Up* is a transition up with F output feature maps and 2×2 stride. *logits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

2.6 Training Details

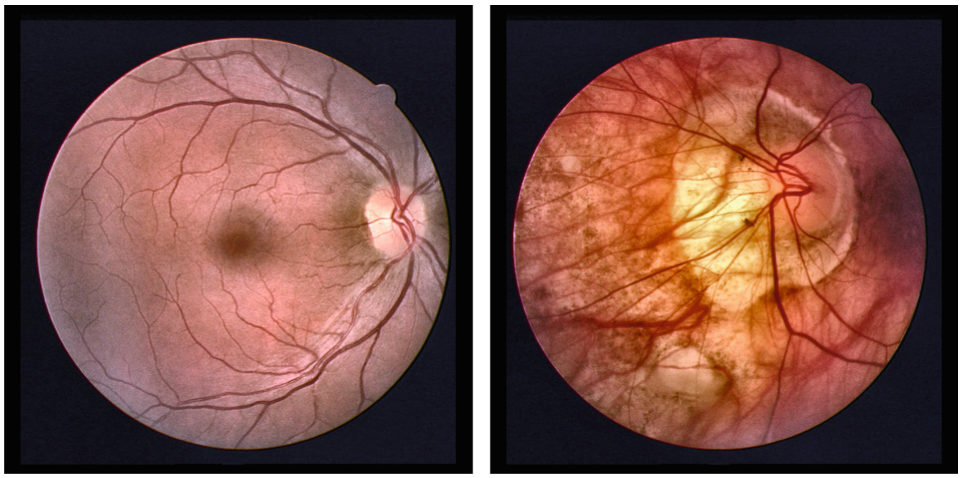
The Dense-FCN was pretrained for 200 epochs with 1000 steps per epoch on random patches 224×224 with the batch size equal to 5. Then it was fine-tuned for 50 epochs with 500 steps per epoch on full size images with the batch size equal to 1.

The weights were initialized using HeNormal [6]. In addition to dropout, $l2$ regularization with the weight decay factor 10^{-4} was used. As the optimizer, Adadelta [17] with the learning rate $l = 1$ and the decay rate $\rho = 0.95$ was used for both the pretraining and fine-tuning. The learning rate was dropped by a factor of 10 if the training loss was not decreased by 0.005 for 10 epochs. Data augmentation by using flipping, reflecting and rescaling (with scale rates 0.8 and 1.2) was applied in both cases. During the fine-tuning stage, the images were randomly padded to size 608×608 so that the size is divisible by 32 and could be properly compressed by the downsampling path. The parameter values were determined empirically based on initial experiments with the RITE database.

3 Experiments and Results

3.1 Training and Evaluation Strategies

Considering the given reference standard, the question arises of how to use the uncertain class labels and its effect on the final training results. Possible ways for utilizing this information are to consider these pixels to be arteries and veins simultaneously including uncertain (IU), or to exclude them from the training completely excluding uncertain (EU). In this work, a comparison of both training strategies is provided. The crossing labels are considered to be veins and arteries simultaneously. Both strategies are evaluated against the reference standard with excluded uncertain labels, and the vessels classification metrics are given by evaluating against the reference standard provided by the second expert.



(a) (b)

Fig. 3. Two examples of preprocessed RITE images.

Since the AV classification problem stated being multilabel, binary classification metrics were calculated for each class separately: area under receiver operating characteristic curve (ROC-AUC), accuracy, sensitivity and specificity.

During the inference stage, the model parameters are sampled 100 times and the number of inferred samples is $N_A = 50$. The final posterior predictive mean is calculated over all predicted samples, and the outputs aleatoric uncertainty U_A and epistemic uncertainty U_E are calculated as in [10]:

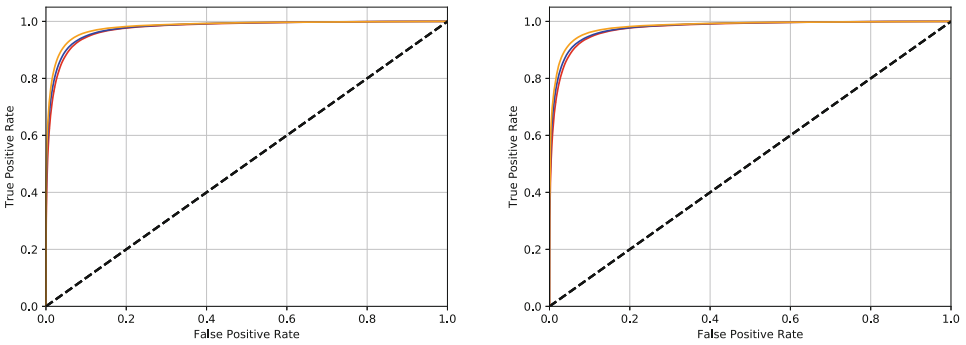
$$U_A = \mathbb{E}_q \left[\mathbb{V}_{p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})} [\mathbf{y}] \right], \tag{9}$$

$$U_E = \mathbb{V}_q \left[\mathbb{E}_{p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})} [\mathbf{y}] \right], \tag{10}$$

where $\mathbb{E}$ and $\mathbb{V}$ denote expectation and variance, respectively.

3.2 Experimental Results

The receiver operating characteristic (ROC) curves calculated after training with both strategies are shown in Fig. 4. The corresponding performance metrics are given in Tables 1 and 2. From the tables, it is clear that the AV classification performances are high, not far from the vessel pixel classification performance. Including uncertain labels into the training set leads to reduced classification accuracy for arteries and veins, but it slightly improves the performance of vessel classification. It is also clear that the Including uncertain strategy increases classification sensitivity, since the training procedure now takes all labeled vessels into account during the AV inference stage.



(a) (b)

Fig. 4. ROC curves for arteries (red), veins (blue) and vessels (orange): (a) Excluding uncertain and (b) including uncertain strategies. (Color figure online)

Table 1. Evaluation results for the excluding uncertain strategy.

Label	ROC-AUC	Accuracy	Sensitivity	Specificity
Arteries	0.973	0.970	0.607	0.992
Veins	0.976	0.970	0.669	0.992
Vessels	0.980	0.960	0.749	0.989

The segmentation results for two example images from the test set are illustrated in Fig. 5. Comparing the results for the training strategies shows that the network trained with the EU strategy tends to be more discriminative for arteries and veins in the areas closer to the optic disc. The common issue for both strategies is the learned bias about the thin vessels being arteries and incapacity to capture connectivity patterns of the predicted segmentation masks inferring vein branches to be arteries.

The aforementioned problems can also be visualized as predicted epistemic and aleatoric uncertainties which are presented in Fig. 6 for the same images

shown in Fig. 5. From the figure, it is clear that the epistemic uncertainty is larger near the optic disc where blood vessels cross. Further away from the optic disc it is concentrated mostly on the vessels' edges with a pattern similar to the one of the aleatoric uncertainty. Similar observations can be made from Fig. 7 where the uncertainties are compared for the two training strategies. The regions of highest uncertainty include vessel crossings and thin vessels even in the case correct classification.

3.3 Comparison with the State of the Art

The Table 3 shows a comparison of the proposed method with recently proposed methods. It is troublesome to directly compare the methods, since the evaluation methods and metrics used by different authors vary. The method proposed by Zhang et al. [18] is clearly superior compared to all the other methods, including the method studied in this work, but the authors use 5-fold cross-validation split, meaning that they have at least 32 images in the training set, whereas in this work the experiments were carried out using standard split with 20 images in the training set. Nevertheless, the performance obtained in this work is comparable with those recently published by Girard et al. [5] and Hemelings et al. [7].

Table 2. Evaluation results for the including uncertain strategy.

Label	ROC-AUC	Accuracy	Sensitivity	Specificity
Arteries	0.973	0.968	0.636	0.988
Veins	0.976	0.966	0.752	0.982
Vessels	0.981	0.961	0.797	0.984

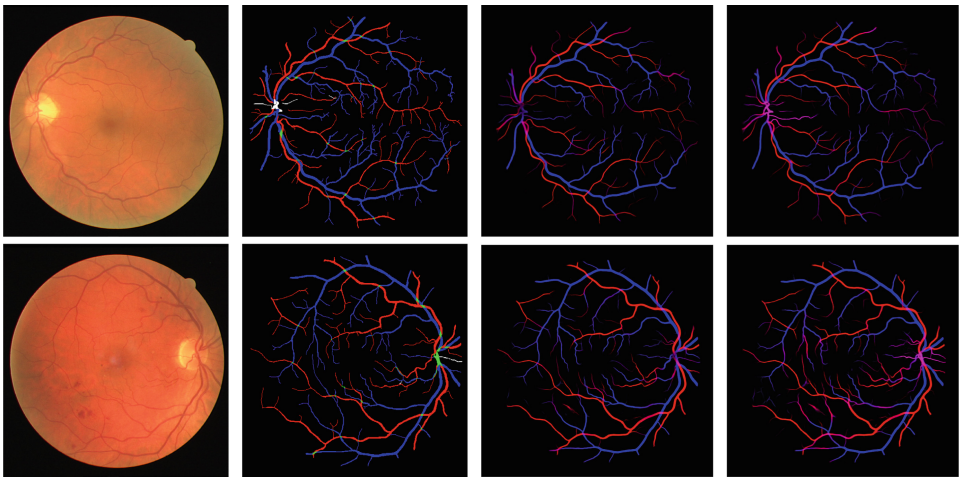


Fig. 5. Visualization of inference result: from left to right, the original image, reference standard, posterior predictive mean obtained with the excluding uncertain strategy with the including uncertain strategy.

Table 3. Comparison of evaluation results. The datasets are specified with splitting methods used by authors.

Method	Vessels accuracy	Arteries accuracy	Veins accuracy	Dataset
Girard et al. [5]	0.948	N/A	N/A	CT-DRIVE
Badawi et al. [2]	0.960	N/A	N/A	DRIVE (standard)
Hemelings et al. [7]	N/A	0.948	0.930	DRIVE (standard)
Zhang et al. [18]	N/A	0.977	0.975	DRIVE (5-fold CV)
This work	0.960	0.970	0.970	DRIVE (standard)

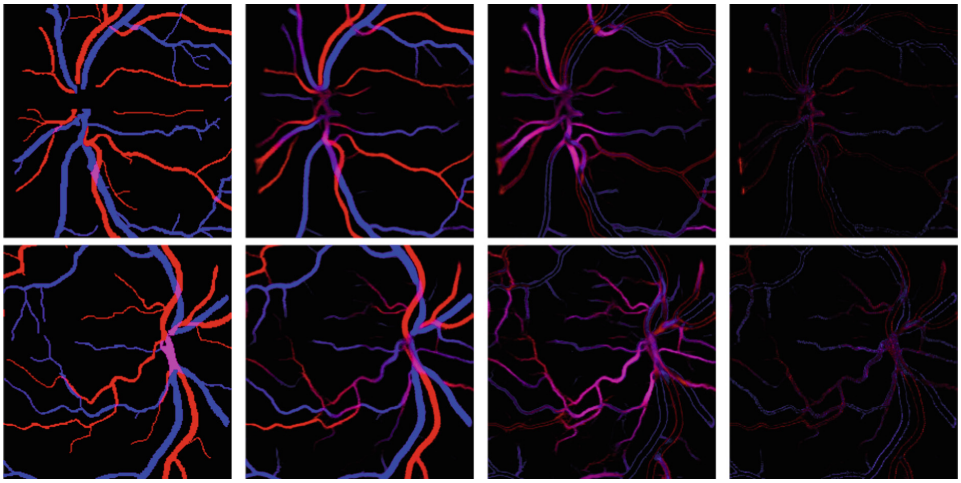


Fig. 6. Visualization of estimated uncertainty: from left to right, targets with removed uncertain labels and crossings, posterior predictive mean, epistemic uncertainty and aleatoric uncertainty. The results are obtained using the excluding uncertain strategy.

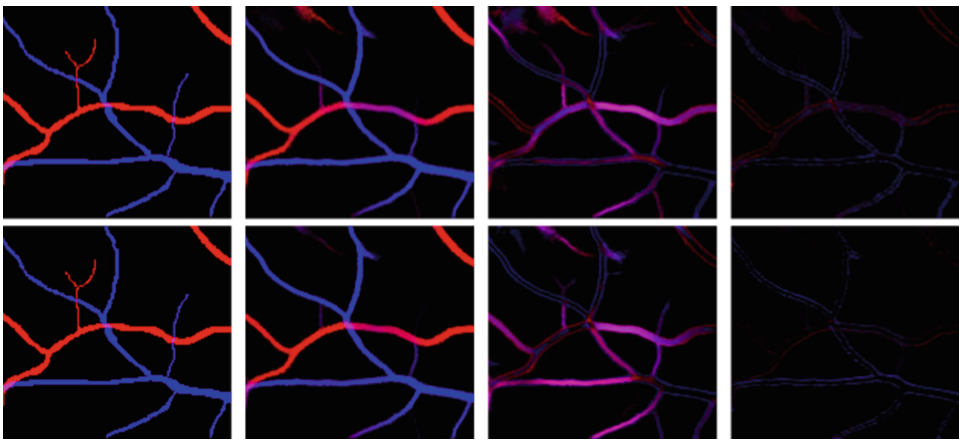


Fig. 7. Visualization of estimated uncertainty: from left to right, targets with removed uncertain labels and crossings, posterior predictive mean, epistemic uncertainty, and aleatoric uncertainty. The results are obtained using the excluding uncertain (top row) and including uncertain (bottom row) strategy.

4 Conclusion

In this work, multilabel classification of arteries and veins using a Bayesian fully-convolutional network was studied. It was shown that the misclassified areas on the images can be visualized using uncertainty estimates. The proposed approach is comparable with recent state-of-the-art approaches for blood vessel segmentation and AV classification methods.

The main topics for the future research are how to reduce the epistemic uncertainty and more careful study on the classification of uncertain labels in the RITE database. Retinal vasculature segmentation and AV classification methods typically include preprocessing procedures that affect the data. One of the opened questions, how different preprocessing techniques change the aleatoric uncertainty estimates. Other possible directions include differentiable end-to-end methods for modeling the connectivity and regularizations similar to [5] and [2].

References

1. Almotiri, J., Elleithy, K., Elleithy, A.: Retinal vessels segmentation techniques and algorithms: a survey. Appl. Sci. **8**(2), 155 (2018)
2. Badawi, S., Fraz, M.: Multiloss function based deep convolutional neural network for segmentation of retinal vasculature into arterioles and venules. BioMed Res. Int. **2019**, 1–17 (2019). https://doi.org/10.1155/2019/4747230
3. Decencière, E., et al.: Feedback on a publicly distributed image database: the messidor database. Image Anal. Stereol. **33**(3), 231–234 (2014)
4. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
5. Girard, F., Kavalec, C., Cheriet, F.: Joint segmentation and classification of retinal arteries/veins from fundus images. Artif. Intell. Med. **94**, 96–109 (2019). https://doi.org/10.1016/j.artmed.2019.02.004
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
7. Hemelings, R., Elen, B., Stalmans, I., Van Keer, K., De Boever, P., Blaschko, M.B.: Artery-vein segmentation in fundus images using a fully convolutional network. Comput. Med. Imaging Graph. **76**, 101636 (2019)
8. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging **19**(3), 203–210 (2000)
9. Hu, Q., Abràmoff, M.D., Garvin, M.K.: Automated separation of binary overlapping trees in low-contrast color retinal images. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 436–443. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_54
10. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. arXiv preprint arXiv:1907.01949 (2019)

11. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1175–1183. IEEE (2017)
12. Jogi, R.: Basic Ophthalmology. Jaypee Brothers Medical Publishers, New Delhi (2008)
13. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, pp. 5574–5584 (2017)
14. Malek, J., Tourki, R.: Blood vessels extraction and classification into arteries and veins in retinal images. In: 10th International Multi-conferences on Systems, Signals Devices 2013 (SSD 2013), pp. 1–6, March 2013
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Welikala, R., et al.: Automated arteriole and venule classification using deep learning for retinal images from the UK Biobank cohort. Comput. Biol. Med. **90**, 23–32 (2017). https://doi.org/10.1016/j.compbiomed.2017.09.005
17. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. Technical report, December 2012. arXiv: 1212.5701 http://arxiv.org/abs/1212.5701
18. Zhang, S., et al.: Simultaneous arteriole and venule segmentation of dual-modal fundus images using a multi-task cascade network. IEEE Access **7**, 57561–57573 (2019). https://doi.org/10.1109/ACCESS.2019.2914319
19. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Heckbert, P.S. (ed.) Graphics Gems IV, pp. 474–485. Academic Press Professional Inc., San Diego (1994). http://dl.acm.org/citation.cfm?id=180895.180940

Publication III

Lindén, M., Garifullin A., and Lensu, L.
Weight Averaging Impact on the Uncertainty of Retinal Artery-Venous Segmentation

Reprinted with permission from
Uncertainty for Safe Utilization of Machine Learning in Medical Imaging,
pp. 52–60, 2020.

Weight Averaging Impact on the Uncertainty of Retinal Artery-Venous Segmentation

Markus Lindén, Azat Garifullin(✉), and Lasse Lensu

LUT University, P.O. Box 20, 53851 Lappeenranta, Finland
{markus.linden,azat.garifullin,lasse.lensu}@lut.fi

Abstract. By examining the vessel structure of the eye through retinal imaging, a variety of abnormalities can be identified. Owing to this, retinal images have an important role in the diagnosis of ocular diseases. The possibility of performing computer aided artery-vein segmentation has been the focus of several studies during the recent years and deep neural networks have become the most popular tool used in artery-vein segmentation. In this work, a Bayesian deep neural network is used for artery-vein segmentation. Two algorithms, that is, stochastic weight averaging and stochastic weight averaging Gaussian are studied to improve the performance of the neural network. The experiments, conducted on the RITE and DRIVE data sets, and results are provided along side uncertainty quantification analysis. Based on the experiments, weight averaging techniques improve the performance of the network.

Keywords: Uncertainty quantification · Bayesian deep learning · Artery-vein segmentation · Blood vessel segmentation · Weight averaging

1 Introduction

Eye diseases have become a rapidly increasing health threat worldwide. Retinal images are a great tool for detecting some of the many ocular disease and diseases such as diabetic retinopathy and glaucoma can be detected from retinal images [12]. Ocular diseases are typically detected from retinal images by analyzing the vessel structure. The use of retinal images enables the diagnosis of ocular diseases in their early stages. The task of analyzing the vessel structure has been traditionally left to medical experts. The attention required by the medical experts in this tasks is, however, great and the task is very consuming and expensive. Studying the possibilities in making this process faster is for that reason important, as it would enable wider screenings for ocular diseases from retinal images. Automated image processing methods are a well-motivated possibility in solving this problem [3].

The possibility to use computers in performing artery-vein segmentation has been the focus of a number of studies during the recent years.

C. H. Sudre et al. (Eds.): UNSURE 2020/GRAIL 2020, LNCS 12443, pp. 52–60, 2020.
https://doi.org/10.1007/978-3-030-60365-6_6

However, artery-vein segmentation still remains a challenging tasks for both humans and machines alike. Some of the difficulties in artery-vein segmentation are related to the imaging conditions in which the retinal images are taken. The images tend to suffer from low contrast and changing lighting conditions, both of which make the segmentation process harder.

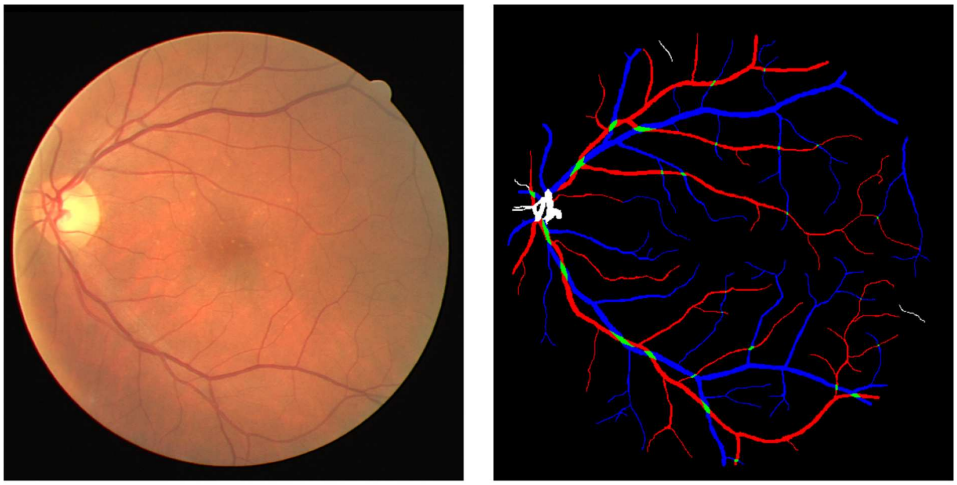
The deep convolutional neural network (DCNN) has recently become the most common tool used in artery-vein segmentation of retinal images, due to the DCNNs ability to automatically learn meaningful features from images. In a paper by Welikala et al., a convolutional neural network (CNN) was used in artery-vein segmentation. The CNN managed to achieve a 82.26% classification rate using UK Biobanks' retinal image database [13]. Hemelings et al. proposed the usage of U-Net architecture for artery-vein classification [5]. In the paper, Hemeling et al. considered the task as a multi-class classification problem with the goal of labeling pixels into four classes: background, vein, artery and unknown. The problem was solved using the retinal images found in DRIVE data set [6] and it achieved classification rates of 94,42% and 94.11% for arteries and veins. Girard et al. [3] modified the U-Net for artery-vein segmentation and found out that using likelihood score in the minimum spanning tree it was possible to improve the performance of the network in the case of smaller vessels. The method was tested using DRIVE data set, achieving an accuracy of 94.93%. Zhang et al. proposed cascade refined U-net to be used in artery-vein classification [14]. The cascade refined U-net consisted of three sub-networks. The task of the first sub-net (A-net in their paper) was to detect all the vessels from the input image, B-net segmented veins from the predicted vessels from the A-net, and finally the C-net segmented the arterioles from the outputs of the previous nets. In the paper, a classification rate of 97.27% was achieved using the automatically detected vessels from the RITE data set. In a paper by Garifullin et al., a dense fully convolutional neural network (Desne-FCN) was used in the task of artery-vein classification [2]. Using the Dense-FCN architecture and the RITE data the authors were able to achieve classification rates of 96%, 97% and 97% for vessels, arteries and veins respectively. In addition to that the authors performed uncertainty quantification on the results obtained using Monte-Carlo dropout [1] for variational approximation. In the aforementioned article, however, the authors did not illustrate the model calibration and the experiments were conducted with one training setup for different labelling strategies. Thus, the question of reliability of the shown uncertainty estimates arises.

This work illustrates how stochastic weight averaging affects the estimated uncertainties. In addition, differences between two epistemic uncertainty estimation techniques are illustrated. Both more traditional binary classification metrics as well as uncertainty quantification metrics are used to evaluate the algorithms.

2 Data

The retinal image data set chosen to be used in this work was the DRIVE data set [6]. The DRIVE data set contains 20 RGB images for testing and 20 for training. The images are of size 584×565.

The AV references standard used in this work is the RITE data set [7]. The RITE data set extends the DRIVE data set with references for arteries, veins, overlapping vessels and uncertain vessels. Red labels in the DRIVE data set stand for arteries, blue labels for veins, green for overlapping vessels and white ones for uncertain vessels. An example of a retinal image from the DRIVE data set as well as the corresponding data labels from the RITE data set can be seen in Fig. 1. During the training the labels for crossings were replaced by labels for both arteries and veins simultaneously and the uncertain labels were omitted for arteries and veins and left for the vessels.



(a) (b)

Fig. 1. (a) Retinal image from the DRIVE data set. (b) Retinal image labels from RITE dataset. (Color figure online)

3 Bayesian AV Classification

3.1 Baseline

Garifullin et al. followed a standard approach for deep Bayesian classification. First, a neural network f is used to estimate the distribution of logits parametrized through the estimate of the mean $\hat{\mathbf{y}}$ and variance $\boldsymbol{\sigma}$ of logits for arteries and veins:

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

The probability vector $\mathbf{p} = [p_{\text{artery}}\ p_{\text{vein}}]$ of the labels can then be calculated as follows:

$$\hat{\mathbf{p}} = \text{sigmoid}\left(\hat{\mathbf{y}} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}\right), \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right). \tag{2}$$

Given the probability vector for arteries and veins the probability for the vessels can be inferred based on the addition law of probability:

$$p_{\text{vessel}} = p_{\text{artery}} + p_{\text{vein}} - p_{\text{artery}} p_{\text{vein}}. \tag{3}$$

The resulting optimisation objective is a sum of binary cross-entropy functions for all three labels over all produced aleatoric samples.

The formulae (1)–(3) take into account heteroscedastic aleatoric uncertainty which is a type of uncertainty dependent on the data capturing imperfect imaging conditions, labeling and image noise. The second kind of uncertainty is epistemic uncertainty representing the model's ignorance. By considering the parameters of the model as a random variable with the posterior $p\left(\boldsymbol{\theta} \mid \mathcal{D}\right)$ the posterior predictive distribution over logits can be calculated as follows:

$$p\left(\mathbf{y} \mid \mathbf{x}, \mathcal{D}\right) = \int p\left(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) \mathrm{d}\boldsymbol{\theta}. \tag{4}$$

Typically, the integral (4) is intractable and stochastic approximations are used in order to estimate the posterior predictive. One of the most common techniques is to use stochastic variational approximation called MC-Dropout [1] which employs dropout as a Monte Carlo sampling technique in order to obtain samples from the model's posterior. Another widely used method is stochastic weight averaging Gaussian [11] where the model's posterior is approximated by a normal distribution the moments of which are estimated during the training procedure.

3.2 Stochastic Weight Averaging

Izmailov et al. found out that the values traversed by SGD would be around the flat regions of the loss surface, without actually reaching the center of this area [9]. By equally averaging these points traversed by SGD, Izmailov et al. found out that points that are inside this more desirable part of the loss surface would be achieved. They named this method stochastic weight averaging (SWA) and it was shown to improve the results and generalization of networks on a variety of architectures and in multiple applications. Given initial pre-trained weights SWA can be implemented as a running average of the weights calculated while continuing training with an additional computation of batch normalization statistics after (see [9] for more details).

3.3 Stochastic Weight Averaging Gaussian

SWAG was first introduced by Maddox et al. [11] for model averaging and uncertainty estimation. The main idea behind is to use SWA to calculate the mean of

the model's parameters and at the same time to estimate a diagonal approximation of the covariance matrix. Thus, the approximated posterior of the model's parameters is a normal distribution:

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \mathcal{N}\left(\boldsymbol{\theta}_{\mathrm{SWA}}, \boldsymbol{\Sigma}_{\mathrm{SWAG}}\right), \tag{5}$$

where $\boldsymbol{\theta}_{\mathrm{SWA}}$ is a parameter vector estimated with SWA and $\boldsymbol{\Sigma}_{\mathrm{SWAG}}$ is a corresponding diagonal covariance matrix.

4 Experiments and Results

4.1 Description of Experiments

The parameters and methodologies presented here were selected so that the baseline model used in this work would be as similar as possible to [2]. The utilized architecture is Dense-FCN-103 [10]. The baseline model was, however, re-implemented and the experiments reproduced to some degree in this work.

In all the experiments, the network was first pre-trained on RITE dataset with random patches of the input images of size 224×224. The batch size used in the pre-training was 5 and the network was pre-trained with 100 epochs and 1000 steps per epoch.

After the pre-training, the networks were fine-tuned with full-size images that were padded to size of 608×608 so that they could be properly compressed by the downsampling part of the network. The main optimizer used in all of the experiments was Adadelta with learning rate of 1 and decay rate of 0.95. The use of either SWA or SWAG would start on a later epochs of full resolution training.

To increase the diversity of the data set data augmentation techniques were used. The augmentation was performed by applying rotation, flipping, and scaling to the input data. The rotation angles used were 90, 180 and 270 degrees and the scaling rates were 0.8, 0.9, 1.0, 1.1 and 1.2.

The aleatoric and epistemic uncertainties were estimated using formulae from [8]. The uncertainties are estimated as an average sum standard deviations per image $S_p = \sum_i \sum_j \sigma_j / N_{\mathrm{test}}$, where i is an index of the image, j is an index of the pixel, and N_{test} is the total number of test images (Table 4).

Baseline. The fine-tuning of the network used as baseline was done using 50 epochs with 500 steps per epoch to match the hyperparameters used in [2]. The batch size used in the fine-tuning of the baseline was selected to be 1. MC-Dropout was used to quantify epistemic uncertainty.

SWA. The SWA implementation also had 50 epochs with 500 steps in each epoch in the full resolution training. Like in the baseline the batch size used was 1. The starting epoch for SWA was selected to be 10 and it was only used in the fine-tuning of the network. The starting epoch was selected through empirical experimentation. MC-Dropout was used to quantify epistemic uncertainty.

SWAG. The hyperparameters used in the SWAG implementation were 500 epochs with 50 steps per epoch. This was done so that the Gaussian posteriori approximation formed by SWAG would be generated from a higher number of epochs. Like in the baseline the batch size used was 1. The SWAG starting epoch was selected to be 100. The epistemic uncertainty was quantified by sampling the model's parameters from Gaussian distribution (5). Whereas the sampling is performed from the posterior estimated with SWAG, dropout is still used during the training phase.

4.2 Performance of the Networks

Due to the fact that artery-vein classification was considered a multilabel problem, the performance metrics used in were calculated for arteries, veins and vessels separately. The selected classification metrics were accuracy, sensitivity, specificity, Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and Estimated Calibration Error (ECE) [4].

By examining the performance metrics presented in Tables 1, 2 and 3, it can be seen that SWA improved the network performance overall compared to the baseline and SWAG models including the model calibration.

Table 1. Network performance in artery classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.970	0.642	0.990	0.00988	0.974
SWA	**0.975**	0.690	0.992	0.00943	0.981
SWAG	0.973	0.706	0.989	**0.00871**	0.966

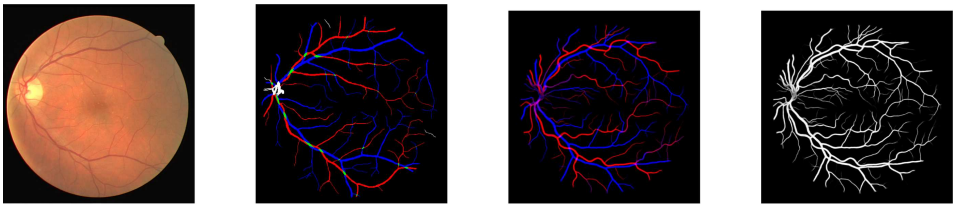
Table 2. Network performance in vein classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.971	0.655	0.994	0.0169	0.980
SWA	**0.974**	0.742	0.991	0.0120	0.991
SWAG	0.971	0.804	0.983	**0.0107**	0.980

Table 3. Network performance in vessel classification (the best accuracy and calibration are in bold)

Method	Accuracy	Sensitivity	Specificity	ECE	ROC-AUC
Baseline	0.957	0.723	0.989	0.0221	0.980
SWA	**0.961**	0.782	0.986	**0.0208**	0.983
SWAG	**0.961**	0.836	0.978	0.0338	0.984

The example of the segmentation results for SWAG is given in Fig. 2. The segmentation examples for the baseline and SWA look similar. The uncertainties of the results were visualized and example figures can be seen in Fig. 3. In the figure, the intensities of the colors describe the uncertainty in that region as standard deviations of the predicted probabilities: the higher intensity the higher the uncertainty.



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

Fig. 2. (a) The input image; (b) ground truth; (c) mean predicted AV probabilities; (d) mean predicted vessels probabilities. The results are obtained using SWAG.

From the tables and figures, it can be concluded that the aleatoric uncertainty of the baseline is much higher than those of SWA and SWAG. It can also be concluded that sampling the network weights from the Gaussian posterior generated by SWAG to create the variational approximation, rather than using Monte-Carlo dropout, has a reducing effect on the levels of epistemic uncertainty present in the predictions. This could probably be explained by the fact that the variance is estimated only around a local optimum during the late stages of the training, whereas MC-Dropout is enabled during the whole training process. From the estimated performance metrics, however, it is difficult to conclude whether it is a positive or negative effect. One noticeable pattern is the high epistemic uncertainty near the optic disc when estimated with MC-Dropout. On the other hand, sampling from Gaussian distribution leads to the high uncertainties mostly near the end points of the blood vessels and the areas after the crossings which is also present in the case of MC-Dropout.

At the same time one can see that aleatoric uncertainties change when SWA or SWAG are utilized. Kendall et al. [1] describe the aleatoric uncertainty as a loss attenuation mechanism allowing the model to adapt the loss dependent on the data and labelling. While the aleatoric uncertainty is meant to be data dependent, the changes to the training procedure affecting the model's convergence and the parameters of the layers predicting variances also affect the predicted aleatoric uncertainties. For the baseline and SWAG, we can see a similar pattern of the higher aleatoric uncertainty levels near the optic disc and borders of the vasculature, whereas the aleatoric uncertainties almost vanish when estimated using MC-Dropout trained with SWA.

Table 4. Mean sums of estimated aleatoric and epistemic uncertainties per image.

Method	Aleatoric			Epistemic		
	Arteries	Veins	Vessels	Arteries	Veins	Vessels
Baseline	1276.2	1159.5	1807.5	4853.6	4066.4	5069.7
SWA	3.3	3.5	5.3	4038.6	3882.3	4659.7
SWAG	31.1	38.9	57.3	997.8	1104.3	1396.1

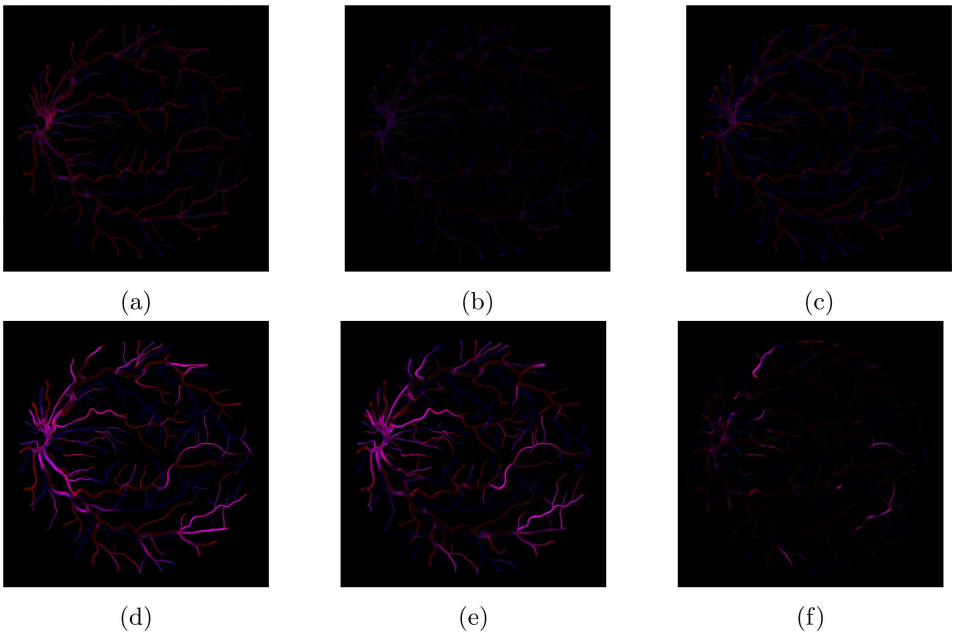


Fig. 3. Aleatoric uncertainties calculated using (a) the baseline, (b) SWA, and (c) SWAG. Epistemic uncertainties calculated using (d) the baseline, (e) SWA, and (f) SWAG.

4.3 Conclusions

In this work, the focus was on blood vessel segmentation from retinal images and on artery-vein classification by using a deep neural network. More specifically, two algorithms were studied to improve the classification performance and help in the model calibration. SWA and SWAG algorithms were implemented on top of the baseline and experimented with the DRIVE and RITE data sets.

The use of SWA improved the performance of the deep neural network on most of the binary classifications as well as the calibration metrics. SWAG showed slight improvements in the vessels and artery classification tasks. The weight averaging as a process significantly affecting the model's convergence seems to lead to diminishing aleatoric uncertainties and sampling from the normal distribution captures less epistemic uncertainty.

References

1. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)
2. Garifullin, A., Lensu, L., Uusitalo, H.: On the uncertainty of retinal artery-vein classification with dense fully-convolutional neural networks. In: Blanc-Talon, J., Delmas, P., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2020. LNCS, vol. 12002, pp. 87–98. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-40605-9_8
3. Girard, F., Kavalec, C., Cheriet, F.: Joint segmentation and classification of retinal arteries/veins from fundus images. Artif. Intell. Med. **94**, 96–109 (2019). https://doi.org/10.1016/j.artmed.2019.02.004
4. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017)
5. Hemelings, R., Elen, B., Stalmans, I., Van Keer, K., De Boever, P., Blaschko, M.B.: Artery-vein segmentation in fundus images using a fully convolutional network. Comput. Med. Imaging Graph. **76**, 101636 (2019)
6. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Trans. Med. Imaging **19**(3), 203–210 (2000)
7. Hu, Q., Abràmoff, M.D., Garvin, M.K.: Automated separation of binary overlapping trees in low-contrast color retinal images. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 436–443. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_54
8. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 137–145. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_16
9. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D.P., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. In: The Conference on Uncertainty in Artificial Intelligence (2018)
10. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1175–1183. IEEE (2017)
11. Maddox, W., Garipov, T., Izmailov, P., Vetrov, D.P., Wilson, A.G.: A simple baseline for Bayesian uncertainty in deep learning. In: NeurIPS (2019)
12. Miri, M., Amini, Z., Rabbani, H., Kafieh, R.: A comprehensive study of retinal vessel classification methods in fundus images. J. Med. Signals Sens. **7**(2), 59 (2017)
13. Welikala, R., et al.: Automated arteriole and venule classification using deep learning for retinal images from the UK biobank cohort. Comput. Biol. Med. **90**, 23–32 (2017). https://doi.org/10.1016/j.compbiomed.2017.09.005
14. Zhang, S., et al.: Simultaneous arteriole and venule segmentation of dual-modal fundus images using a multi-task cascade network. IEEE Access **7**, 57561–57573 (2019). https://doi.org/10.1109/ACCESS.2019.2914319

Publication IV

Garifullin, A., Lensu, L., and Uusitalo, H.

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

Reprinted with permission from
Computers in Biology and Medicine
Vol. 136, pp. 104725, 2021.

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.

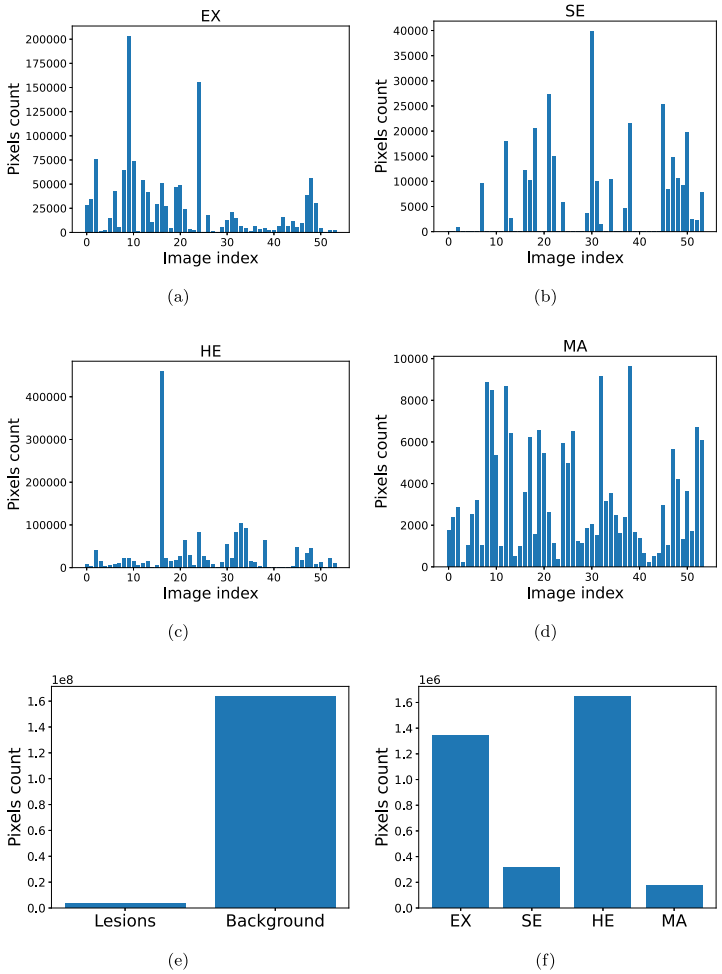
Guo et al. [11] proposed a multi-scale feature fusion method to handle issues with small lesion detection. Binary cross-entropy (BCE) loss with balancing coefficients was used to improve the sensitivity. The model was trained with full images resized to 1440×960 pixels without any further preprocessing. 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.



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

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.

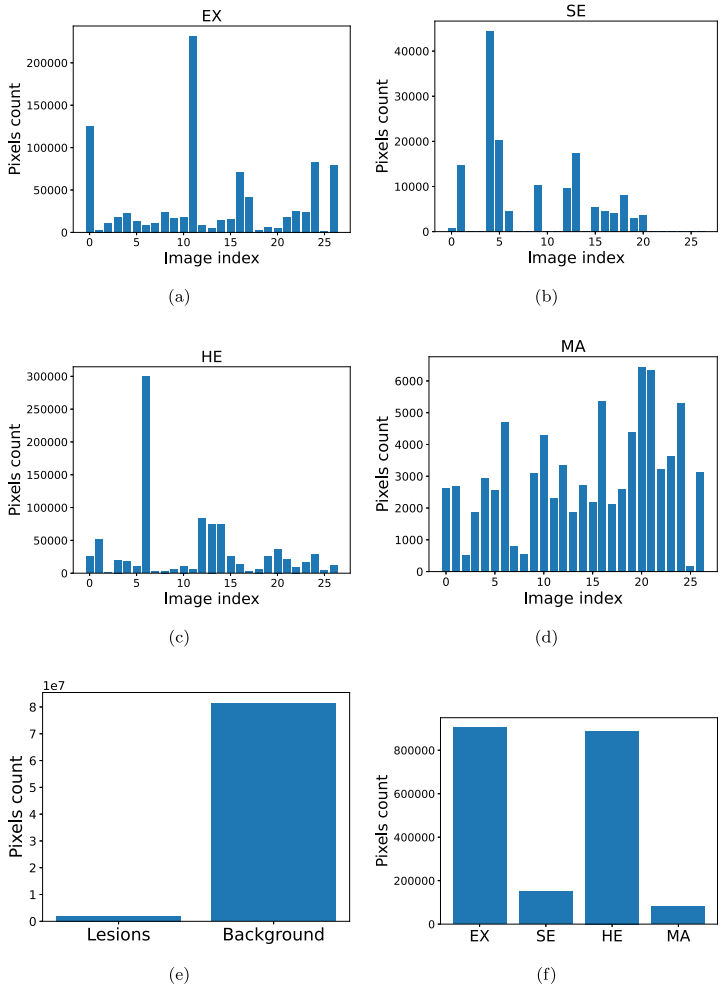


Fig. 3. Statistics of lesions in the test set. The number of positive pixels per image for (a) hard exudates (EX), (b) soft exudates (SE), (c) haemorrhages (HE), and (d) microaneurysms (MA). (e) The number of pixels for the lesions and the background. (f) The number of positive pixels for each lesion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

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

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

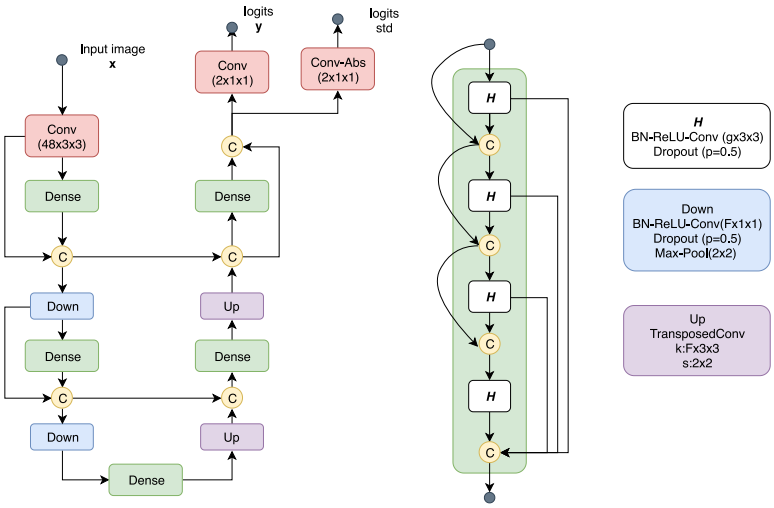


Fig. 4. The Dense-FCN architecture: *Dense* stands for a dense convolutional block; *C* is a tensor concatenation; *H* is a block consisting of batch normalization (BN), rectified linear unit (ReLU) and a convolutional layer with growth rate *g*; *Down* is a transition-down block with *F* output feature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

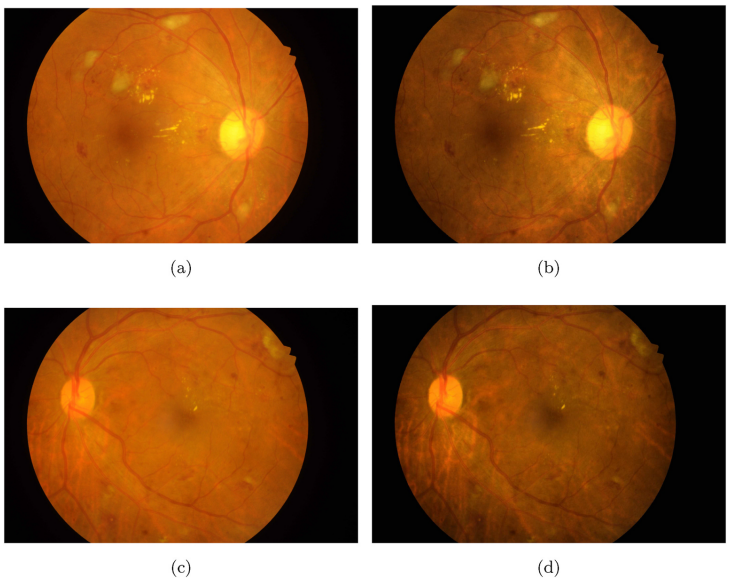


Fig. 5. Two examples of the original (left column) and enhanced (right column) images.

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.

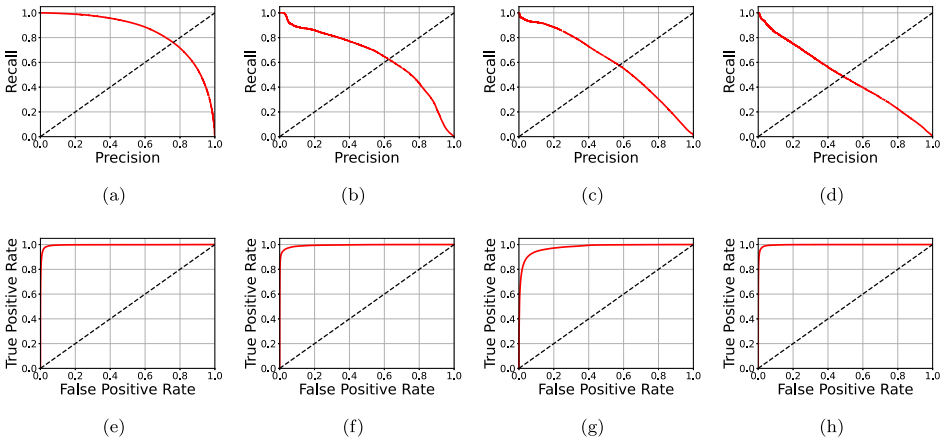


Fig. 6. Precision-recall and receiver operating characteristic curves for (a), (e) hard exudates; (b), (f) soft exudates; (c), (g) haemorrhages; (d), (h) microaneurysms.

- 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.

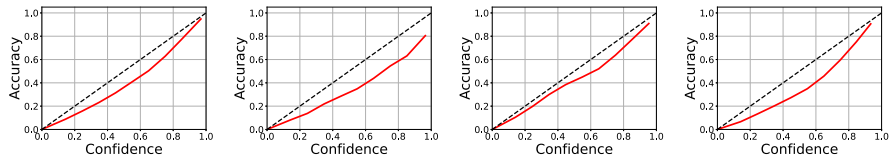
To summarize, the extended validation approach consists of the analysis of the produced segmentation masks as well as comparison of the produced uncertainties and the misclassification maps.

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

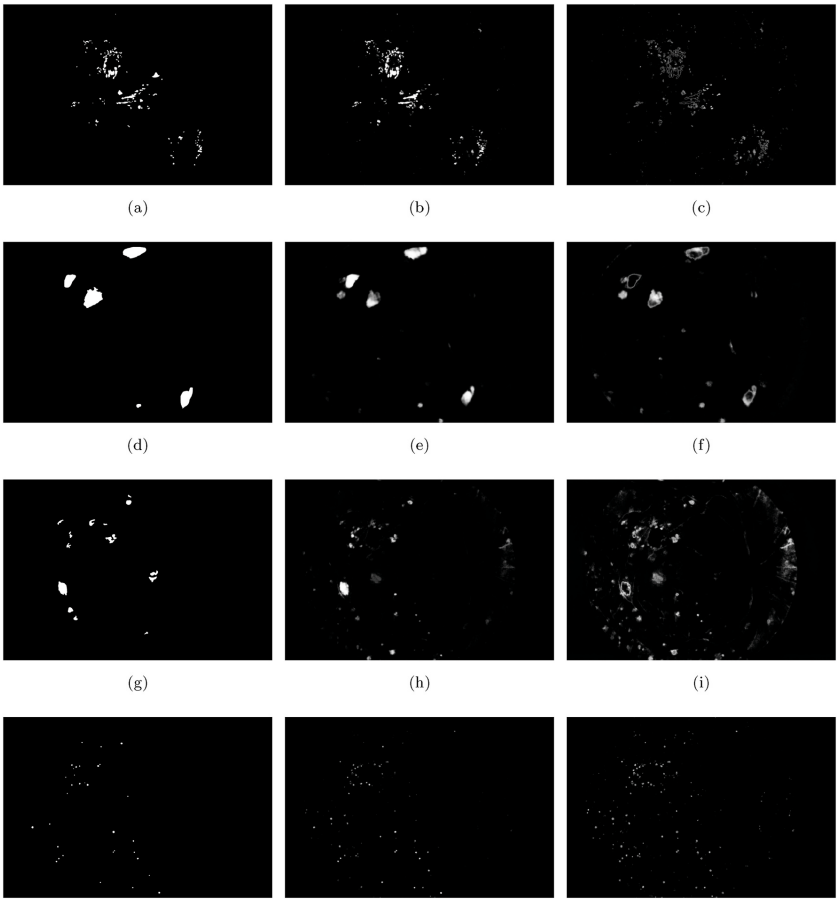
Table 1
Evaluation results of the baseline training scheme. The abbreviations of the evaluation metrics are explained in the text.

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



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

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



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

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).

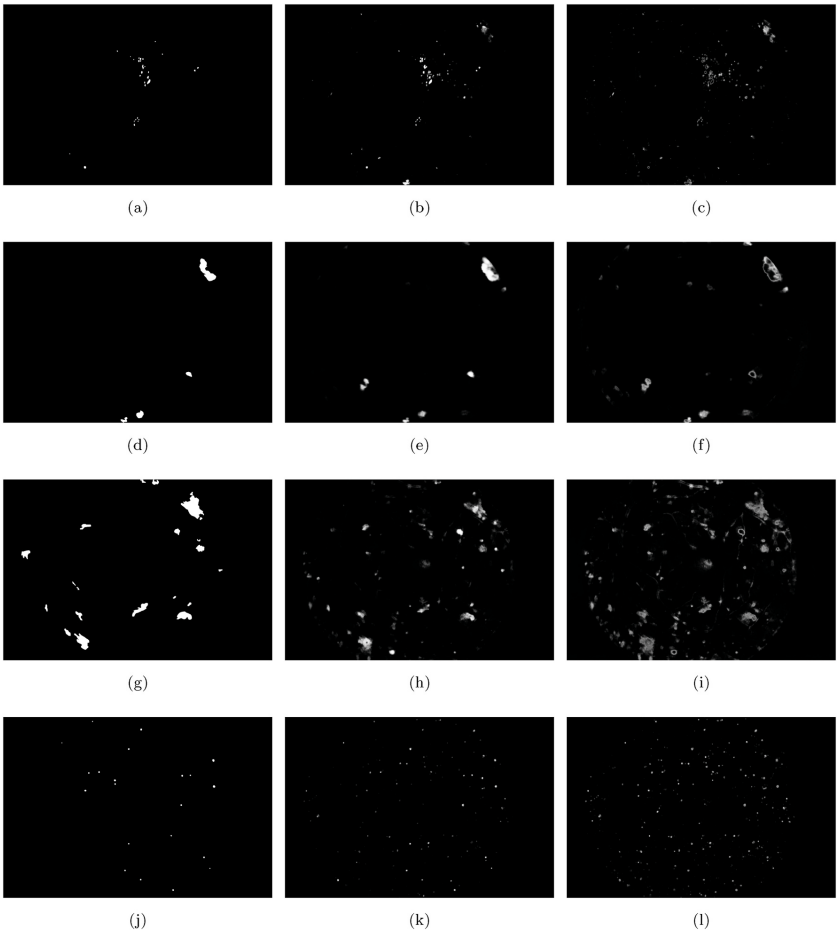


Fig. 9. Visualizations of inference results for input image 5d for lesions: (a), (b), (c) hard exudates; (d), (e), (f) soft exudates; (g), (h), (i) haemorrhages; (j), (j), (l) microaneurysms. The first column shows the ground truth masks, the second shows the mean inferred probabilities and the third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

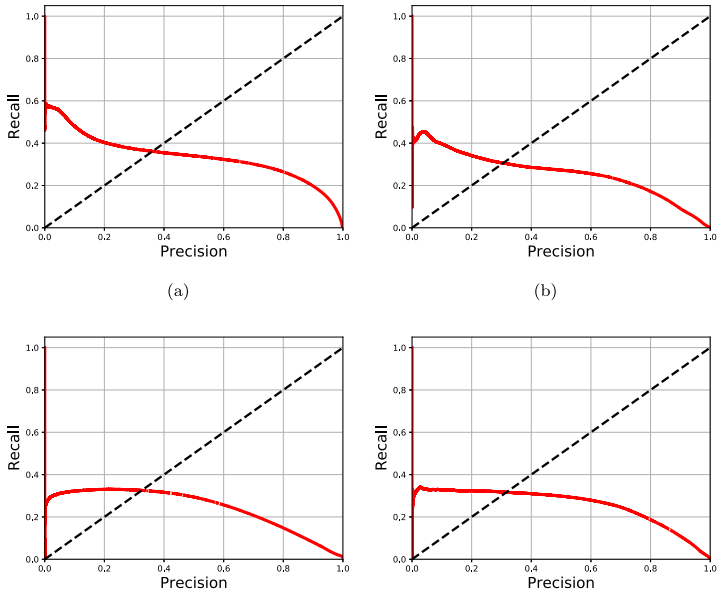


Fig. 10. Uncertainty precision-recall curves for (a) hard exudates; (b) soft exudates; (c) haemorrhages; (d) microaneurysms.

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.

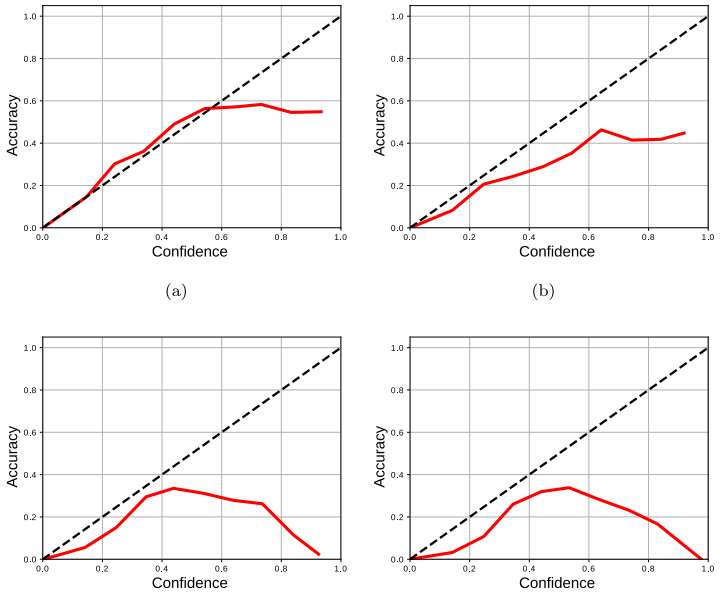


Fig. 11. Uncertainty reliability diagrams for (a) hard exudates; (b) soft exudates; (c) haemorrhages; (d) microaneurysms.

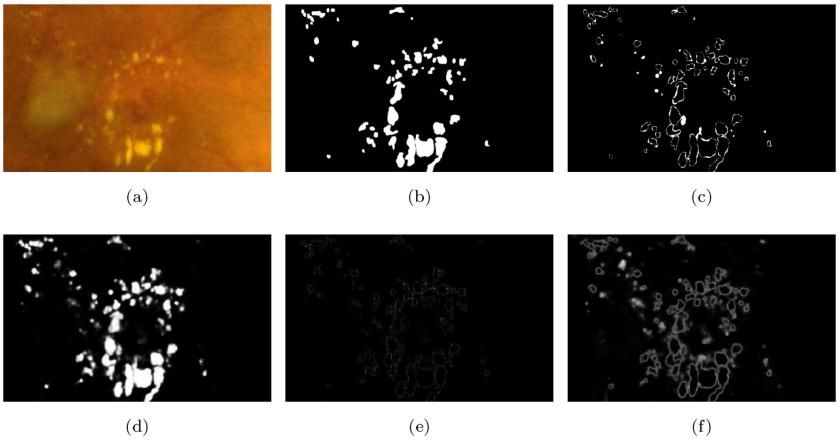


Fig. 12. Inference results for hard exudates with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

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

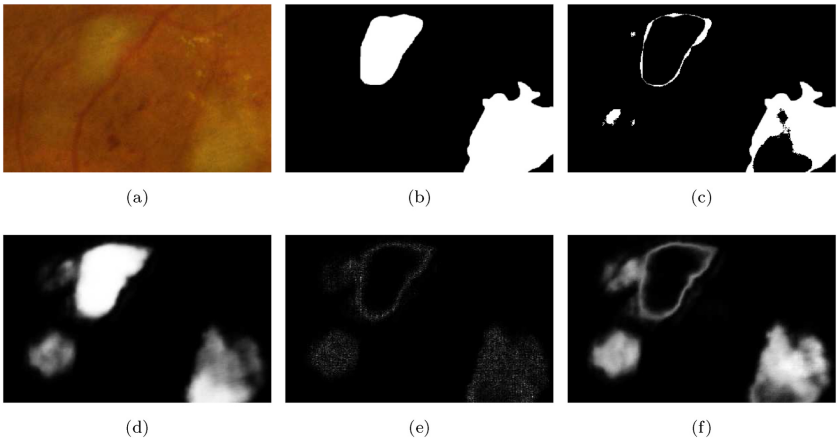


Fig. 13. Inference results for soft exudates with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

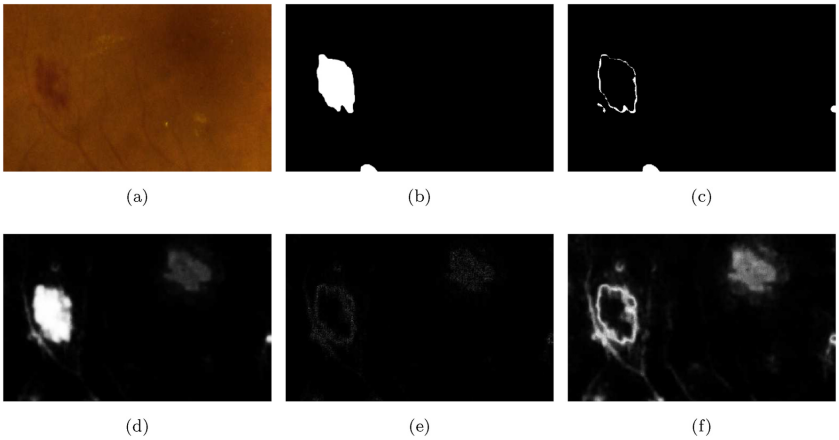


Fig. 14. Inference results for haemorrhages with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

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

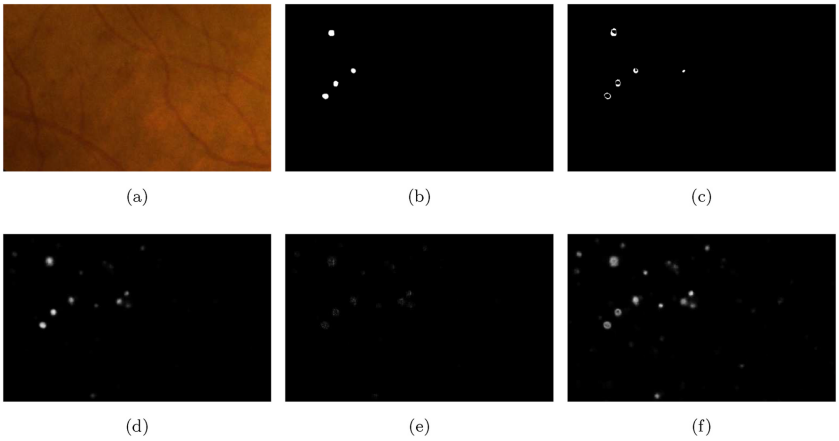


Fig. 15. Inference results for microaneurysms with magnified input image 5b: (a) input image; (b) ground truth mask; (c) misclassifications; (d) mean inferred probabilities; (e) aleatoric uncertainties (standard deviations of probabilities); (f) epistemic uncertainties (standard deviations of probabilities).

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.

ACTA UNIVERSITATIS LAPPEENRANTAENSIS

965. ZHANG, TAO. Intelligent algorithms of a redundant robot system in a future fusion reactor. 2021. Diss.

966. YANCHUKOVICH, ALEXEI. Screening the critical locations of a fatigue-loaded welded structure using the energy-based approach. 2021. Diss.

967. PETROW, HENRI. Simulation and characterization of a front-end ASIC for gaseous muon detectors. 2021. Diss.

968. DONOGHUE, ILKKA. The role of Smart Connected Product-Service Systems in creating sustainable business ecosystems. 2021. Diss.

969. PIKKARAINEN, ARI. Development of learning methodology of additive manufacturing for mechanical engineering students in higher education. 2021. Diss.

970. HOFFER GARCÉS, ALVARO ERNESTO. Submersible permanent-magnet synchronous machine with a stainless core and unequal teeth widths. 2021. Diss.

971. PENTTILÄ, SAKARI. Utilizing an artificial neural network to feedback-control gas metal arc welding process parameters. 2021. Diss.

972. KESSE, MARTIN APPIAH. Artificial intelligence : a modern approach to increasing productivity and improving weld quality in TIG welding. 2021. Diss.

973. MUSONA, JACKSON. Sustainable entrepreneurial processes in bottom-of-the-pyramid settings. 2021. Diss.

974. NYAMEKYE, PATRICIA. Life cycle cost-driven design for additive manufacturing: the frontier to sustainable manufacturing in laser-based powder bed fusion. 2021. Diss.

975. SALWIN, MARIUSZ. Design of Product-Service Systems in printing industry. 2021. Diss.

976. YU, XINXIN. Contact modelling in multibody applications. 2021. Diss.

977. EL WALI, MOHAMMAD. Sustainability of phosphorus supply chain – circular economy approach. 2021. Diss.

978. PEÑALBA-AGUIRREZABALAGA, CARMELA. Marketing-specific intellectual capital: Conceptualisation, measurement and performance. 2021. Diss.

979. TOTH, ILONA. Thriving in modern knowledge work: Personal resources and challenging job demands as drivers for engagement at work. 2021. Diss.

980. UZHEGOVA, MARIA. Responsible business practices in internationalized SMEs. 2021. Diss.

981. JAISWAL, SURAJ. Coupling multibody dynamics and hydraulic actuators for indirect Kalman filtering and real-time simulation. 2021. Diss.

982. CLAUDELIN, ANNA. Climate change mitigation potential of Finnish households through consumption changes. 2021. Diss.

983. BOZORGMEHRI, BABAK. Finite element formulations for nonlinear beam problems based on the absolute nodal coordinate formulation. 2021. Diss.

984. BOGDANOV, DMITRII. Transition towards optimal renewable energy systems for sustainable development. 2021. Diss.

985. SALTAN, ANDREY. Revealing the state of software-as-a-service pricing. 2021. Diss.

986. FÖHR, JARNO. Raw material supply and its influence on profitability and life-cycle assessment of torrefied pellet production in Finland – Experiences from pilot-scale production. 2021. Diss.

987. MORTAZAVI, SINA. Mechanisms for fostering inclusive innovation at the base of the pyramid for community empowerment - Empirical evidence from the public and private sector. 2021. Diss.

988. CAMPOSANO, JOSÉ CARLOS. Integrating information systems across organizations in the construction industry. 2021. Diss.

989. LAUKALA, TEIJA. Controlling particle morphology in the in-situ formation of precipitated calcium carbonate-fiber composites. 2021. Diss.

990. SILLMAN, JANI. Decoupling protein production from agricultural land use. 2021. Diss.

991. KHADIM, QASIM. Multibody system dynamics driven product processes. 2021. Diss.

992. ABDULKAREEM, MARIAM. Environmental sustainability of geopolymer composites. 2021. Diss.

993. FAROQUE, ANISUR. Prior experience, entrepreneurial outcomes and decision making in internationalization. 2021. Diss.

994. URBANI, MICHELE. Maintenance policies optimization in the Industry 4.0 paradigm. 2021. Diss.

995. LAITINEN, VILLE. Laser powder bed fusion for the manufacture of Ni-Mn-Ga magnetic shape memory alloy actuators. 2021. Diss.

996. PITKÄOJA, ANTTI. Analysis of sorption-enhanced gasification for production of synthetic biofuels from solid biomass. 2021. Diss.

997. MASHLAKOV, ALEKSEI. Flexibility aggregation of local energy systems—interconnecting, forecasting, and scheduling. 2021. Diss.

998. NIKITIN, ALEKSEI. Microwave processes in thin-film multiferroic heterostructures and magnonic crystals. 2021. Diss.

999. VIITALA, MIRKA. The heterogeneous nature of microplastics and the subsequent impacts on reported microplastic concentrations. 2021. Diss.

1000. ASEMOKHA, AGNES. Understanding business model change in international entrepreneurial firms. 2021. Diss.

1001. MUSTO, JIRI. Improving the quality of user-generated content. 2021. Diss.

1002. INKERI, EERO. Modelling of component dynamics and system integration in power-to-gas process. 2021. Diss.

LUT University

ISBN 978-952-335-761-7
ISBN 978-952-335-762-4 (PDF)
ISSN-L 1456-4491
ISSN 1456-4491
Lappeenranta 2021