Lappeenranta University of Technology
School of Engineering Science
Master's Programme in Computational Engineering and Technical Physics
Intelligent Computing Major

Master's Thesis

Liubov Nedoshivina

ACTIVE LEARNING OF THE GROUND TRUTH FOR RETINAL IMAGE SEGMENTATION

Examiners:	Professor Lasse Lensu Ph.D. Maxim Peterson
Supervisors:	Professor Lasse Lensu

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Master's Programme in Computational Engineering and Technical Physics
Intelligent Computing Major

Liubov Nedoshivina

ACTIVE LEARNING OF THE GROUND TRUTH FOR RETINAL IMAGE SEGMENTATION

Master's Thesis

2018

53 pages, 23 figures, 5 tables.

Examiners: Professor Lasse Lensu
Ph.D. Maxim Peterson

Keywords: computer vision, active learning, retinal imaging, retinal image segmentation

Diabetic retinopathy and other eye-related diseases can be diagnosed from eye fundus images by medical experts who look for specific lesions in the images. Automated diagnosis methods can help medical doctors to increase the diagnosis accuracy and decrease the time needed. In order to have a proper dataset for training and evaluating the methods, a large set of images should be annotated by several experts to form the ground truth. To enable efficient utilization of expert's time, active learning is studied to accelerate the collection of the ground truth. Since one of the important steps in the retinal image diagnosis is the blood vessels segmentation, the corresponding approaches were studied. Two approaches were implemented and extended by proposed active learning methods for selecting the next image to be annotated. The performance of the methods in the case of standard implementation and active learning application was compared for several retinal images databases.

PREFACE

I would like to thank my supervisor Lasse Lensu for the guiding me during all the research. I would also like to express my special thanks to Pavel Vostatek who provided the essential information on the segmentation algorithms.

Lappeenranta, May 24, 2018

Liubov Nedoshivina

CONTENTS

LIST OF ABBREVIATIONS

Acc	Accuracy
BCNN	Bayesian Convolutional Neural Networks
CAD	Computer-Aided Diagnosis
CMIF	Collection of Multispectral Images of the Fundus
CNN	Convolutional Neural Networks
DC	Dice Coefficient
DR	Diabetic Retinopathy
DSC	Dice Similarity Coefficient
EBM	Evidence-Based Medicine
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
GT	Ground Truth
HRF	High-Resolution Fundus
kNN	k-Nearest Neighbours
LL	Logarithm of Likelihood, LogLikelihood
NPDR	NonProliferative Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
QBC	Query By Committee
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic (curve)
ROC	Retinopathy Online Challenge (database)

1 INTRODUCTION

1.1 Background

Evidence-based medicine (EBM) is the current practice in many subfields of medical science. In this approach, the medical diagnosis and planning of treatment is based on scientific knowledge and objective examination of each patient through biomedical measurements [1]. One example of the knowledge used in the process is images because of the versatile possibilities to examine the condition of the patient or her organs. From this viewpoint, the medical doctors base their decisions nowadays on a more complete and timely view to the condition.

Eye-related diseases like the diabetic retinopathy are diagnosed from eye fundus images by medical experts who look for specific lesions in the images. Not only the rethinopaty could be diagnosed from these kinds of images. For example, in the research conducted by R. Poplin et al. in 2018 [2], a deep neural network could predict a heart disease risk only by an eye fundus image.

The required attention of medical expert in a fundus examination restricts the possibility to perform broad screenings of eye diseases. For screening and monitoring a progressive disease, automatic image processing methods are a well-motivated possibility to help a single expert's work, or enable a wider screening program [3].

To develop and compare methods for automated image analysis, it is important to have reliable expert knowledge for the image content. In order to have a proper dataset for training and evaluating the methods, a large set of images should be annotated by several experts, and either the annotations should be fused to form the ground truth (GT) for the image content or the level of agreement and performance of the experts should be evaluated to define the gold standard [4]. This approach was successfully applied with different classification models. In the review [3], several examples of the implementations where high performance in the retinopathy diagnosis was achieved are described: for example, more than 90% of the accuracy in identification of the diabetic retinopathy stages [5], [6].

Collecting the annotations for a set of retinal images and the subsequent training of a deep neural network based on this knowledge were applied in [4]. According to the results of this research, high accuracy of retinopathy grading was achieved, but the amount of resources needed was quite large: the training set was consisted from more than 128 thousand images and 54 experts were recruited to label the data. The algorithm of grading was based on a neural network classifier. The performance of the automatic grader was close to the expert assessments.

Active learning is studied to accelerate the collection of the GT to allow reaching the expected diagnosis performance faster. Having a small annotated dataset, in the case of active learning, an algorithm by a special function queries an unlabeled set for a model training. The experiments in the field of automated retinopathy diagnosis with active leaning already exist [7], [8]. In [7], C.I. Sánchez et al. were able to reduce the size of training set by 80% while keeping a high success rate.

As it was determined in [2], the blood vessel condition is connected to the cardiovascular disease. One of the important steps in the retinal images based diagnosis is the blood vessel segmentation. In the segmentation tasks the result can be presented in the form of label map which is a binary image where 1 corresponds to vessel class and 0 to non-vessel class. Hence the segmentation task is a binary pixel-wise classification task. The active learning approaches can be applied to form the most informative and compact representation of the ground truth needed for the segmentation.

1.2 Objectives

The objectives of the research are as follows:

1. Study active learning algorithms which could be used for collecting of the ground truth in case of the automated medical diagnosis task.
2. Select the segmentation methods which would be suitable for the retinal blood vessel segmentation task and propose the active learning solutions for them.
3. Select the datasets of the retinal images where medical expert annotations are given and the GT is presented in form of a segmentation map.
4. Select the methods for evaluating the performance of the segmentation methods with the given GT collection method.
5. Assess the applicability of the proposed methods to the databases by comparing the performance.

1.3 Structure of the thesis

The key aspects of retinal imaging and automated diagnosis are presented in Chapter 2. Also basic principles of the active learning and a review of the learning algorithms are discussed in that Chapter in Section 2.4. Chapter 3 contains descriptions of the selected segmentation methods and the proposed active learning approaches. In Chapter 4, the evaluation methods and the selected datasets characteristics are described. In that Chapter in Section 4.3, the experiments and the achieved results of the conducted experiments are presented. Finally, in Chapter 5, the results are analyzed, the future work is proposed, and in Chapter 6 the general conclusions are given.

2 AUTOMATED ANALYSIS OF RETINAL IMAGES

In order to form the most informative ground truth, it is important to establish the features of the retinal based diagnosis process. The main purpose of this Chapter is to describe retinal imaging and the basic active learning approaches applied in particular to the automated analysis of the biomedical images.

2.1 Eye structure and retinal imaging

A human eye is a quite sensitive organ to study and has a complex structure. The main parts of the eye are the iris, pupil, lens, retina, optic nerve. The iris performs a function similar to a camera aperture controlling the amount of light coming to the eye through the pupil, whereas the flexible lens focuses the light similar to a camera lens. Eye fundus or retina is the light sensitive area. It performs the sensing like the camera sensor by conversion of light to the neural signals [9]. The whole eye is covered by the special protective tissue called sclera. The structure of the eye is presented in Fig. 1.

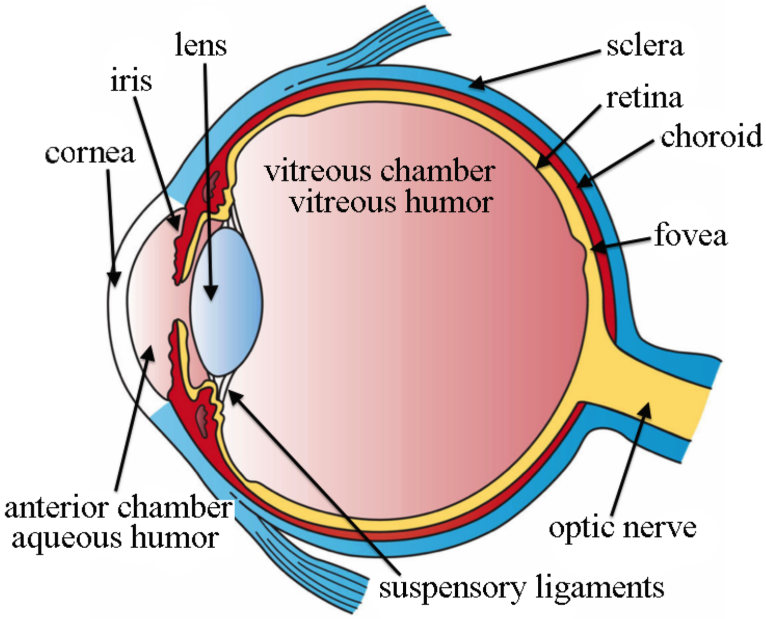


Figure 1. Structure of the eye [10].

Diseases like diabetic retinopathy could be diagnosed by studying the eye condition [4]. The invasive methods are not comfortable for the patient and could damage some parts of the eye, therefore, non-invasive techniques are needed. The main tool for the diagnosis of the eye-related diseases is the eye fundus camera (Fig. 2), which provides a non-invasive way to examine the current condition of the eye.

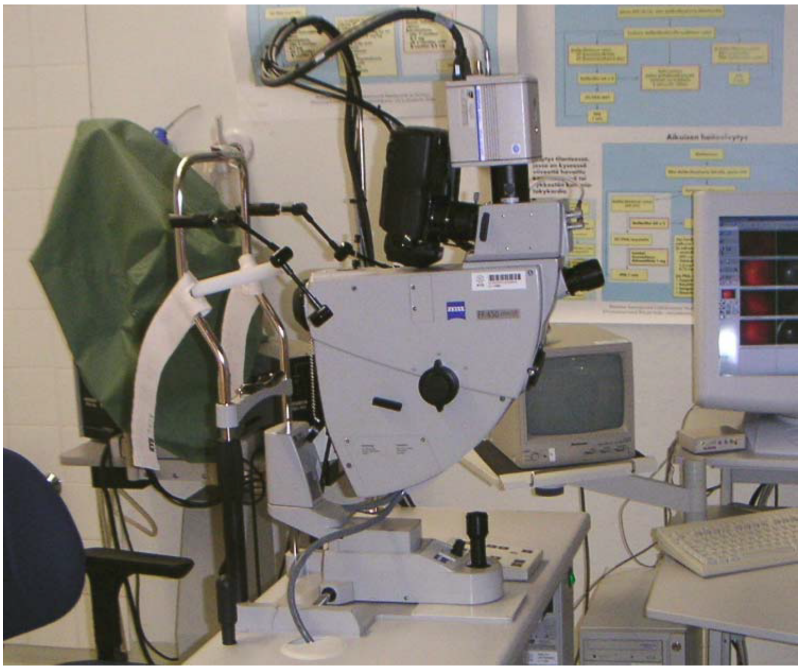


Figure 2. An eye fundus camera [9].

There are two kinds of cameras which are the mydriatic and non-mydriatic cameras. In order to use the mydriatic camera, the dilation of the pupil is needed, whereas while using the non-mydriatic cameras it is not necessary, but the obtained images will have worse quality and resolution. The camera could provide a color image of the retina and also highlight the important parts by applying special filters. Also images containing only green and blue channels or red-free images are often used in the retinopathy diagnosis. Another way to present the condition of the eye fundus is to use the fluorescein angiography. By using a fluorescent dye and a specialized camera, a special image (an angiogram) is obtained. The angiogram could give the important information on the vascular status [11]. An example of a color retinal image is shown in Fig. 3.

By using simple and fast way to obtain the retinal image and medical expert knowledge in the form of ground truth, it is possible to perform automatic diagnosis or detection of lesions related to the eye diseases. One of the signs of the diabetic retinopathy is the presence of the lesions of the different kind: red lesions and bright lesions. The latter category could be divided into called hard exudates, drusen and cotton wool spots (soft exudates). Also the types of retina damage such as microaneurysms, haemorrhages, neovascularization and macular edema are the signs which indicate the presence of the

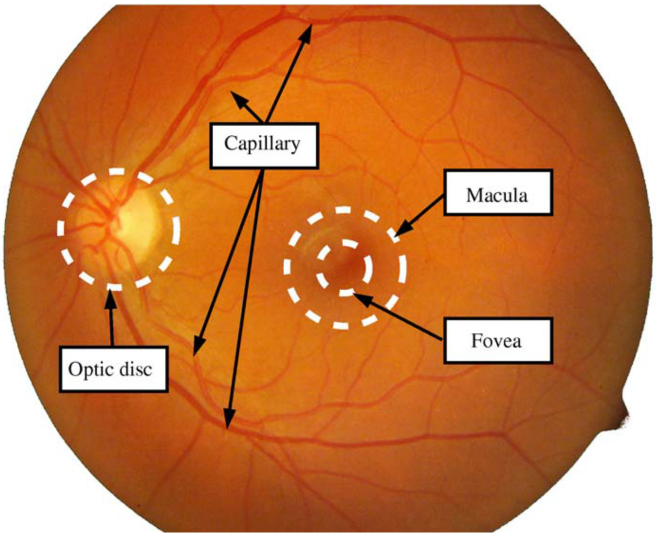


Figure 3. Example retinal image with the main parts labeled [9].

diabetes [9], [12]. If one can train an image analysis model to recognize these lesions, the automated diagnosis process could be performed. To establish the presence some of the lesions, the blood vessels segmentation methods are needed. An example of the image of the retina with abnormalities is shown in Fig. 4.

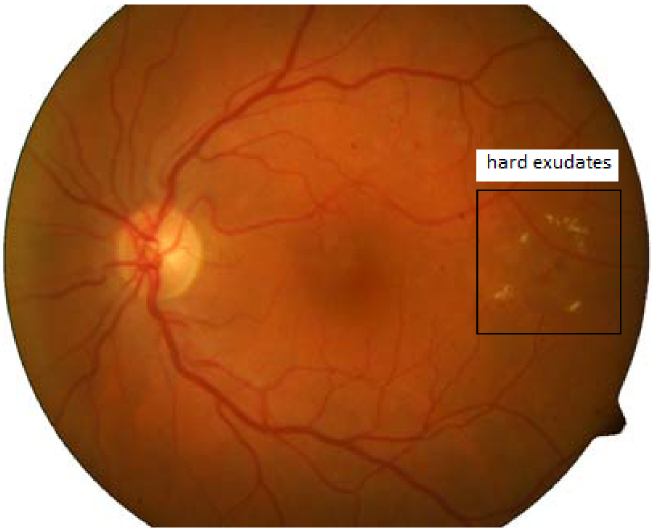


Figure 4. Example of the retina containing lesions (hard exudates) [9].

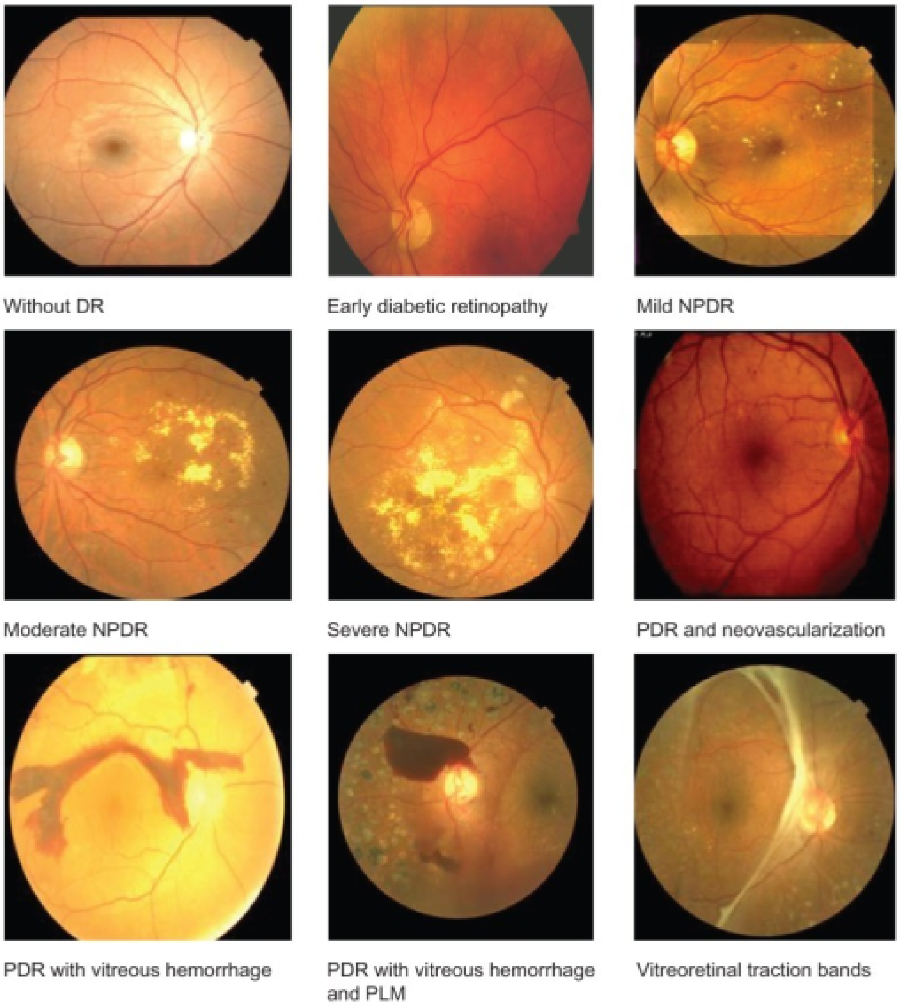
2.2 Computer-aided diagnosis

Since the possibility to perform the non-invasive examination of the body condition is available nowadays, the computer aided diagnosis systems (CAD) develop rapidly. It becomes crucial in the cases, where painful tests with subsequent complex chemical analysis are needed or where X-rays are used, which could not be applied too often because of its harmful effects. Also, such approaches may take a lot of time, whereas CAD systems could give at least an initial prediction on the presence of a disease very fast. By using only an image or the set of images (probably, of the special kind, like spectral images, an angiogram or a computer tomogram), a machine learning algorithm can detect a pathology. In the case of supervised automated methods, a set of medical annotations is needed to train the diagnosis model. A variety of image processing and classification methods has already been successfully applied to solve this task, which can be seen, for example, in [4], [7], [8], [13], [14], [15].

In the case of CAD of diabetic retinopathy, the task can be defined as a two-class to five-class classifications. In the two-class classification method, it is assumed that a model only tells whether the query image contains sings of the abnormalities or not. With the increasing number of classes, the stages of the retinopathy are added. These are non-proliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, proliferative diabetic retinopathy (PDR) and macular edema [12]. Stages of the rethinopathy as well as the health condition of the eye are presented in Fig. 5 [16]. Standard classification algorithms such as nowadays popular neural networks [17], support vector machine [18], k-Nearest Neighbours (kNN) [19], different statistical approaches has been successfully applied to solve this problem and in some cases 99% success rate has been achieved [12].

2.3 Databases of retinal images with the ground truth

As research in the field of the eye diseases diagnosis have been carried out for a long time, several large databases of labeled eye fundus images are available. One of the publicly available datasets of retinal images and the ground truth for such lesions as hard and soft exudates, microaneurysms and hemorrhages is DiaRetDB1 [20] which was created during the project [21]. The authors also proposed an annotation framework for the collection of the ground truth. Another database containing the ground truth information and way to mark the abnormalities was proposed by Michael D. Abramoff and is called ROC (Retinopathy Online Challenge) [22], [23].



Without DR

Early diabetic retinopathy

Mild NPDR

Moderate NPDR

Severe NPDR

PDR and neovascularization

PDR with vitreous hemorrhage

PDR with vitreous hemorrhage and PLM

Vitreoretinal traction bands

Figure 5. Stages of the diabetic retinopathy. The abbreviations are as follows: diabetic retinopathy (DR), proliferative diabetic retinopathy (PDR), previous laser marks (PLM), non-proliferative diabetic retinopathy (NPDR) [16].

A database of 400 annotated images without the GT collection framework was created during the STARE project (STructured Analysis of the Retina), which was started in 1975 [24]. It could be applied in the segmentation tasks such as blood vessel segmentation. Also for the segmentation purposes mainly, DRIVE [25] database was created. 40 segmented images are available in the database. CHASEDB1 [26] consists of the right and left eyes color images and contains 28 images with the ground truth in the form of the segmented images. One of the commonly used datasets which contain relatively large amount of manually segmented retinal images (143) is ARIADB [27].

In the large-scale research on diabetic retinopathy grading described in [4], the authors created their own annotated dataset consisted from 128 175 images and DR grade from 54 experts. They also used as validation sets such databases as EYEPACS-1 (9963 images collected by the authors in United States and Indian hospitals) [28] and the freely available

MESSIDOR-2 (1728 images) [29].

Spectral imaging of the eye fundus is also possible. CMIF database (Collection of multispectral images of the fundus) [30], [31] contains several images for the visible range of electromagnetic spectrum. Key characteristics of the databases can be found in Table 1.

Table 1. Key characteristics of the retinal images databases. N_{GT} is the amount of GT sets per each image.

Database	**Number of images**	**Manually segmented images**	N_{GT}
EYEPACS-1 [28]	9963	-	-
MESSIDOR-2 [29]	1728	-	-
STARE [24]	400	+	2
CMIF [30]	281	-	-
ARIADB [27]	143	+	2
ROC [22]	100	-	-
DiaRetDB1 [20]	89	-	-
DRIVE [25]	40	+	2
CHASEDB1 [26]	28	+	1

There is another point of view concerning the necessity of such databases. A large amount of already properly diagnosed images could be used for educational and training purposes of medical students. It would be a great source of information for practicing doctors which could reduce uncertainty in the difficult cases. Finally, with suitable framework, patients could get information on their condition and would notice the necessity to visit a doctor.

2.4 Segmentation of biomedical images

Since the blood vessel segmentation is an essential step of the automated retinal image diagnosis, multiple solutions have been proposed [32]. All the methods can be divided into two groups. The first group includes the conventional feature-based description of the preprocessed images. The Soares et al. described a two-step supervised vessel segmentation method in [33]. Gabor wavelets are used to form a feature description of a retinal image in the first step. In the second step, Bayesian classification is performed. A supervised algorithm by Sofka et al. [34] represents a likelihood ratio test consisting of multi-scale matched filtering and measures of vessel edges and their confidence.

In the segmentation method proposed by Nguyen et al. [35], line detectors at different scales are applied to an image. Having a set of rotated lines, the algorithm can segment vessels at multiple angles. The method training is performed in an unsupervised manner. Bankhead et al. proposed another unsupervised blood vessel segmentation method [36] in a similar way as it was in the Soares et al. work [33]. The first step involves the wavelet transform to form a feature description. Spline fitting is applied to determine the orientation of vessels. Based on perpendiculars to the vessel, zero-crossings the second derivative are determined. Azzopardi et al. [37] proposed a filter selectively responding to blood vessels based on a pool of Difference of Gaussians filters. This method is also unsupervised.

The second group of the segmentation methods involves convolutional neural networks usage. O. Ronneberger et al. proposed the CNN of a specific architecture constructed mainly for the segmentation purposes [38] and called it the U-Net. Since this is the network, a labeled dataset of images is required for the training.

2.5 Taxonomy of active learning approaches

The primary problem in the image-based diagnosis is the necessity to have a large database of annotated images which should be labeled by medical expert. Such work is very time demanding, therefore, some methods for reducing the amount of the expert's work are needed. To obtain a model, which could perform the diagnosis process, usually learning is required. One of the possible ways is to use the active learning approaches [39].

Active learning uses the idea of the online machine learning [40] where the training data is represented as an ordered sequence during the training. The main element of the active learning algorithms is an active query function by using which one can perform efficient selection or sampling of an object from an unlabeled dataset to a desired training set [39]. The basic principle of these methods is that the learning process could be partially controlled and an acceptable performance could be achieved with a small training set (an ideal situation in this case is to use the one-shot learning [41], where a model learns information on the object from one image or from a small set of images).

2.5.1 Pool-based and stream-based active strategies

The construction of the query function is the primary task in the active learning implementation. There are several different ways how to perform it. Based on the survey conducted by Burr Settles in 2010-2012 [39], there are pool-based and stream-based strategies. In the former case, a small labeled dataset and a large unlabeled set are presented. The query requests sample from the static pool (the unlabeled dataset) by evaluating an informativeness of all or most of the samples in the set. Before the actual sampling, the ordering of all the samples depending on their informativeness (ranking) is performed. In the stream-based sampling strategies, no ranking is assumed and during the query one sample at a time is considered. A scheme of the active learning strategies based on [39] is shown in Fig. 6. The pool-based approaches are often applied in practice in tasks such as medical diagnosis [7] [8], text classification [42], image classification [43]. However, in the cases where a device computational and memory resources are limited, stream-based strategies could be more applicable.

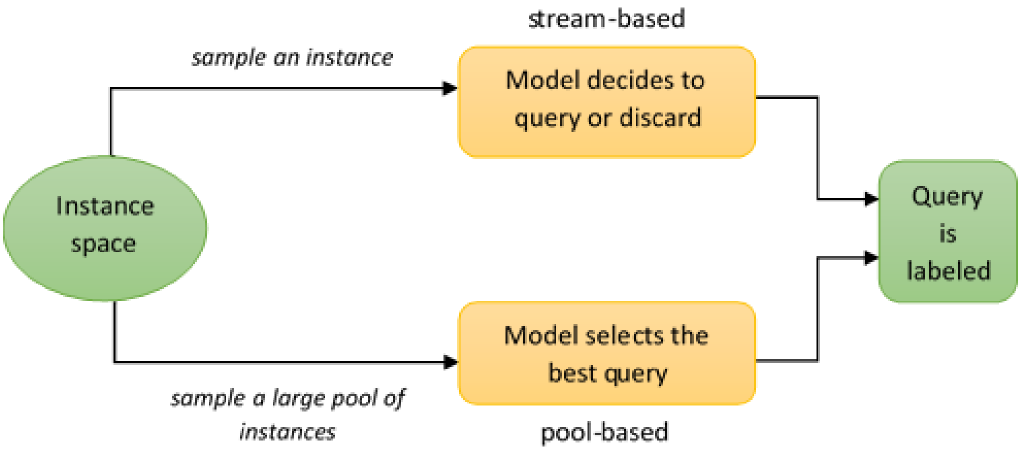

stream-based
sample an instance
Model decides to query or discard
Instance space
Query is labeled
Model selects the best query
sample a large pool of instances
pool-based

Figure 6. Active learning strategies [39].

2.5.2 Active query functions

Whether the pool or stream based sampling is used, a query selection strategy is required. One of the most popular query strategies is uncertainty sampling which was firstly proposed in 1994 by Lewis and Gale [44]. Here an active function queries a pool to find a sample on which classifier produces the most uncertain result. By using this simple technique to solve the problem of text classification, the authors could achieve successful results with less than 1000 samples in the training set whereas by using random sampling it required no less than 100 000 examples to achieve the same success rate.

Another popular technique called query-by-committee (QBC) was proposed by Seung et. al. in 1992 [45]. The idea was to apply not a single classifier as it is in the uncertainty sampling, but an ensemble of classifiers and choose the sample on which there is the greatest disparity in the classification results.

Other methods are focused more on the prediction error of the model, like expected error reduction [46], variance reduction [47] [48], or on how to change the model parameters in the most suitable way, like expected model change algorithms [49].

2.5.3 Application of the active learning

In the recent works, the convolutional neural networks [50] were successfully used as a part of the active learning algorithm. In [51], Geifman and El-Yaniv proposed a pool-based active learning approach. They tested their algorithm on the MNIST [52] (60000 training samples of images of handwritten digits) and CIFAR [53] (25000 and 150000 images of 10 different classes) image databases. The authors used a pre-trained model and improved the initial performance of it by applying a proposed querying function.

Another way to apply neural networks was proposed by Gal and Ghahramani in [43], where Bayesian Convolutional Neural Networks (BCNN) [54] [55] were used to query the pool. The key idea and the main difference from [51] is that their query functions use advantages of Bayesian modeling and convolutional networks to represent the uncertainty of the samples. Authors used for training their BCNN 1000 labeled samples from MNIST database. Several acquisition functions were considered and the training set was reduced to approximately 300 samples having the error rate 5%.

One of the theoretical studies of the stream-based strategy was proposed in [56] by El-Yaniv and Wiener, where the selective classification [57] case was considered. The authors used a binary classification case and emphasized the realizable setting. The active learning reduction to the perfect selective classification was presented. Other recent works devoted to this type of active strategy can be found in [58], [59], [60].

2.6 Active learning of the ground truth for medical images

One way to collect the ground truth of medical images is to use the public on-line platform where medical experts around the world could easily annotate the images. Such approach was described in [61], where the platform CrowdFower was used to diagnose different diseases [62].

The annotation process is quite resource-consuming and another solution is to use the active learning algorithms which could at least help the expert to select the most informative samples from the database. One of the research where machine learning was applied was conducted by Albarqouni et al. in 2016 [63]. Authors used convolutional neural network to aggregate the annotations of breast cancer histology images during the training process.

Active learning approaches for reducing the amount of the training set applied in the various medical tasks are considered in [15]. This research investigates the impact of the case selection during the classification. Authors trained a set of classifiers on the breast masses images database and proposed several case selection methods.

In the papers [7] and [8] devoted to the retinopathy diagnosis, C.I. Sánchez et al. proposed the pool-based uncertainty sampling, Query-by-Committee sampling and compared their performance to the random sampling. The main goal of their classification algorithm was to predict the exact type of the abnormality: drusen, hard and soft exudates. The authors preprocessed the input images by using several filters based on Gaussian derivatives and then applied the kNN classifier. The dataset they used contained normal images and damaged images as well. To evaluate the performance of the classifier the Receiver Operating Characteristic (ROC) curve was used. Based on this criteria, their active learning uncertainty sampling strategy outperformed simple random sampling with area under the ROC curve around 0.88 in the former case against 0.84 in the latter. The QBC results were similar to the random sampling performance.

Based on the conducted review, it was noticed that the most popular solution is to apply the pool-based uncertainty sampling in comparison with the random selection of the samples. Such approach will also be applied to the retinal blood vessel segmentation methods in this research.

3 RETINAL IMAGE SEGMENTATION WITH ACTIVE LEARNING

Two segmentation methods were selected from the ones considered previously to examine the effectiveness of different active learning approaches. The first one is the method proposed by Soares et al. [33] (the *Soares method* or the *Soares model*). This method is based on the Gabor filtering and is followed by Bayesian classification. The CNN called *U-Net* [38] with active learning extention proposed in [64] was selected as the second method to research. The both mentioned methods are supervised and require labeled dataset available. The active learning can be applied to reduce the training set size. One of the simplest and often used methods is the poll-based uncertainty sampling. An application of this active query function to the segmentation methods will be considered in this Chapter.

3.1 Blood vessel segmentation based on Gabor features and supervised Bayesian classification

This method can be divided into two major stages. The first one is the preprocessing part where the Gabor wavelet transform is applied to an inverted green channel of a retinal image. The implementation of the wavelet transform through the Fourier transform can be defined as follows [33]:

$$T_\psi(\mathbf{b}, \theta, a) = C_\psi^{-\frac{1}{2}} a \int \exp(j\mathbf{k}\mathbf{b}) \hat{\psi}^*(a_{-\theta}\mathbf{k}) \hat{f}(\mathbf{k}) d^2\mathbf{k} \tag{1}$$

where ψ is the wavelet, $\mathbf{k}$ is the wave vector, ψ^* is the complex conjugate of ψ, C_ψ is the normalizing constant, a is the scale parameter, $\mathbf{b}$ is the displacement vector, θ is the rotation angle, $j = \sqrt{-1}$, $\hat{f}$ and $\hat{\psi}^*$ - the Fourier transform.

The response of the filter is supposed to be a feature description of the input image. The two-dimensional Gabor wavelet is presented as follows [33]:

$$\psi_G(\mathbf{x}) = \exp(j\mathbf{k_0}\mathbf{x}) \exp(\frac{1}{2}|A\mathbf{x}|^2) \tag{2}$$

where $A = \operatorname{diag}[\epsilon^{-\frac{1}{2}}, 1]$, $\epsilon \geq 1$ is a 2×2 matrix which contains the filter anisotropy information. The frequencies $\mathbf{k_0}$ are another parameter of the method. The transform is then calculated pixel-wise for the different scales and angles in the range of $[0°..170°]$ with the step of $10°$. The maximum modulus of these values from all the considered rotations is computed to form the feature:

$$M_\psi(\mathbf{b}, a) = \max_\theta |T_\psi(\mathbf{b}, \theta, a)|. \tag{3}$$

The obtained pixel features are then normalized with the mean and standard deviation values:

$$v_i = \frac{v_i - \mu_i}{\sigma_i} \tag{4}$$

where v_i is the ith feature, μ_i and σ_i is the mean value and standard deviation of the ith feature.

After the filter response is formed, the classification process based on the obtained feature description can be started. In this research, the Gaussian Mixture Model (GMM) classifier was selected [65]. This method is based on the Bayes decision rule:

$$C = \begin{cases} C_1, p(\mathbf{x}|C_1)P(C_1) > p(\mathbf{x}|C_2)P(C_2) \\ C_2, p(\mathbf{x}|C_1)P(C_1) \leq p(\mathbf{x}|C_2)P(C_2) \end{cases} \tag{5}$$

where $P(C_i)$ is the prior probability of the class C_i, $p(\mathbf{x}|C_i)$ the class-conditional probability density function. $P(C_i)$ is calculated as $P(C_i) = \frac{N_i}{N}$, N is the size of the training set and N_i is the number of samples of the class C_i.

One of the important parameters of the GMM classification is the number of Gaussians k_i corresponding to class i. The parameters of each Gaussian estimated by using the Expectation-Maximization (EM) process [65]. The number of iterations for the EM process is another key parameter.

As a result of the *Soares method*, a probability map can be obtained, where for each pixel there is a probability value of being a vessel. By using thresholding, for example, all values less then 0.5 are set to 0 and otherwise to 1, a segmented image is constructed.

Based on the output probability map p with size $N \times M$, the overall image logarithm of the likelihood (loglikelihood, LL) can be calculated as follows:

$$LL = \sum_{\substack{0<i<M \\ 0<j<N}} \log p(i,j). \tag{6}$$

By using the LL value to characterize the segmented image the uncertainty estimation can be performed which allows to apply the active learning to train the model.

3.2 Supervised classification based on the U-Net CNN architecture

A convolutional neural network architecture called *U-Net* [38] was proposed by O. Ronneberger et al. in 2015. This network is one of the commonly used approaches for biomedical segmentation. The main purpose was to apply it to the task of image segmentation. The feature of the *U-Net* is that it is supposed to be trained on a small training set with preliminary data augmentation. In the conducted research, the authors used datasets with the size from 20 to 40 annotated images. The obtained testing results outperformed the previous successful solutions.

The scheme of the network is presented in Fig. 7. The *U-Net* is consisted of 23 convolutional layers, where a convolution of an input data with filters of a specific size is performed. This CNN implements the encode-decode architecture [66], which means that it encodes an input image to a feature map and then decodes it to a desired output. This CNN performs dowsampling by convolution with 3×3 filters and then upsampling. The ReLU (a rectified linear unit) function is using as activation function:

$$f(x) = \max(x, 0) \tag{7}$$

where x is a neuron output value. The downsampling operation is carried out by a max-pooling operation, the upsampling is implemented through a convolution with a 2×2 filter. At each downsampling layer the number of features is doubled. The last layer is a 1×1 convolutional layer which maps the output features to the target number of classes. As it can be seen in Fig. 7, this network does not have fully connected layers, as it is in the typical CNN architectures.

One of the possible ways to apply active learning here is to use the neurons activations to calculate the informativeness of the sample and based on it select the next one. As it was proposed in [67] by A. Kendall et al., having such an encode-decode network architecture the pixel-wise uncertainty can be estimated by using a dropout. Usually, the dropout is performed by random deactivation of the network activations during the training process. The authors proposed to apply the Monte Carlo Dropout during the testing to calculate the uncertainty. This technique allows to approximate the weight distribution $q(\mathbf{W})$ by minimizing the Kullback-Leibler divergence [68] between the full posterior distribution $p(\mathbf{W}|\mathbf{X},\mathbf{Y})$ and the approximating one:

$$KL(q(\mathbf{W})||p(\mathbf{W}|\mathbf{X},\mathbf{Y})) = \sum_i q(W_i) \log \frac{q(W_i)}{p(W_i|(\mathbf{X},\mathbf{Y})}. \tag{8}$$

In [67], authors used the Bernoulli distribution to approximate the weights. The most uncertain sample is selected based on computing the variance for each pixel for the different predictions. The research aimed to biomedical image segmentation and using this technique was proposed by M. Górriz Blanch in 2017 [64].

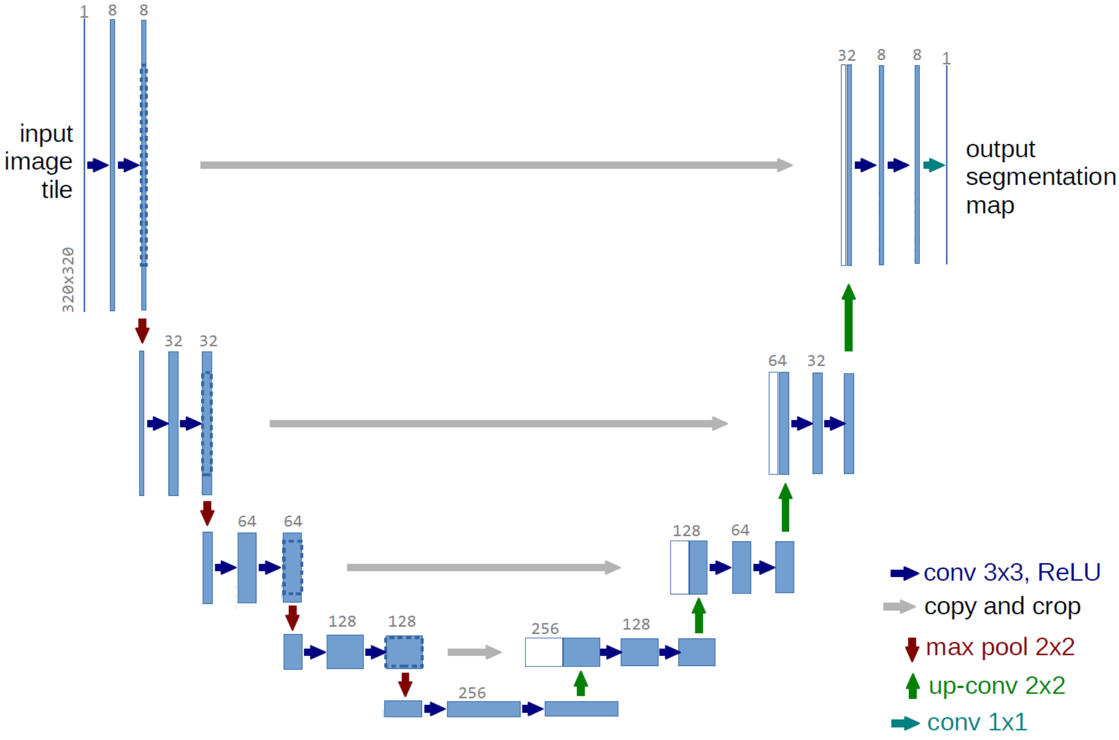


Figure 7. The scheme of the *U-Net* [38]. A blue box is represent a multi-channel feature map. The number on top of the box corresponds to a number of channels. White boxes denote copied feature maps.

4 EXPERIMENTS AND RESULTS

Retinal image databases containing the annotations in the form of segmentation maps are required to test the performance of the selected segmentation methods. The appropriate evaluation criteria selection is also needed. The essential step is the method parameters tuning.

In this Chapter the testing results of the considered segmentation methods in case of the standard training and the active learning application are presented.

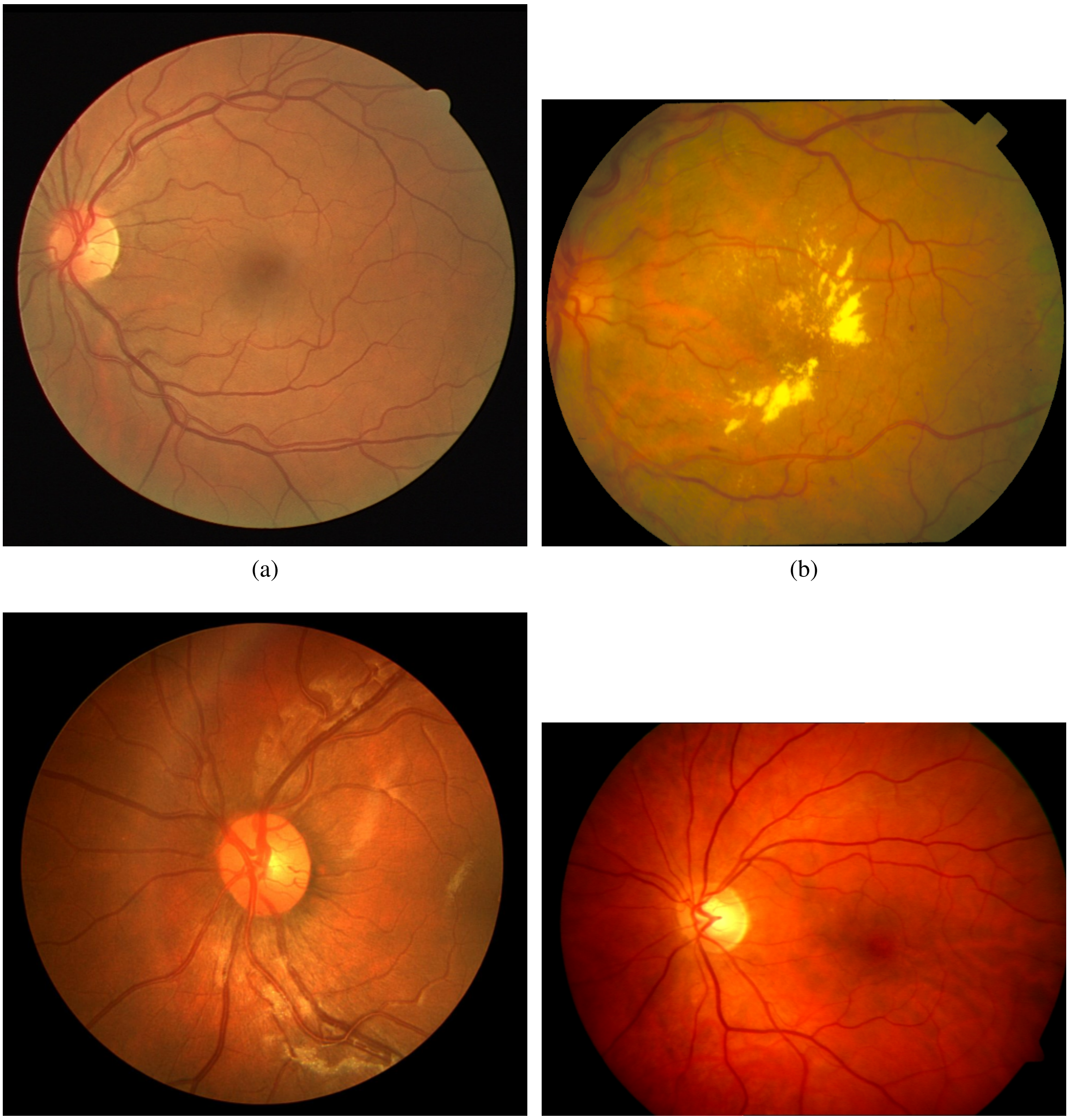
4.1 Databases of the retinal images

From the list of publicly available databases several datasets of RGB retinal images were selected. Their characteristics can be found in Table 2. Ground truth information is presented in the form of binary segmentation maps. The spatial resolution of the images from the CHASEDB1 dataset were decreased two times when using the *Soares method*. Sample images from each considered dataset are presented in Fig. 8.

Table 2. Selected databases of RGB retinal images with ground truth information, N_t - the total number of images in the dataset, N_{seg} is the amount of the segmented images (GT), N_{GT} is the amount of GT sets per each image

Database	N_t	N_{seg}	**Resolution**	N_{GT}
STARE [24]	400	20	700×605	2
ARIADB [27]	143	143	768×576	1
DRIVE [25]	40	40	584×565	2
CHASEDB1 [26]	28	28	500×480 (999×960)	1

Example images and the corresponding segmentation map from DRIVE are shown in Fig. 9. These images were cropped in order to exclude the black border and focus the segmentation model on the retina.



(a)

(b)

(c)

(d)

Figure 8. Sample images from the DRIVE (a), STARE (b), CHASEDB1 (c) and ARIADB (d) datasets.

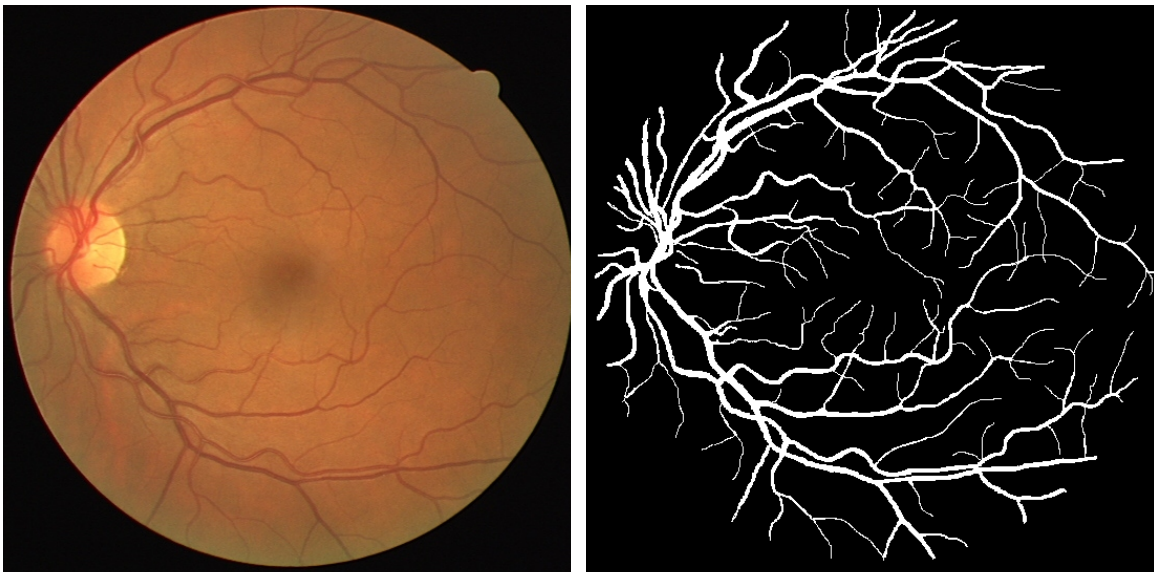


Figure 9. Example image of DRIVE [25]: (a) RGB retinal image; (b) the ground truth in the form of segmentation map.

4.2 Evaluation criteria

For the evaluation of the segmentation results, the standard techniques such as the Dice Similarity Coefficient and Accuracy were selected. The Dice Similarity Coefficient (DSC, DC) [69] is defined as

$$DC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{9}$$

where $DC \in [0, 1]$, A and B are sets which are the image segmented by a model and the the image segmented by an expert in case of the segmentation task, $|A|$ means the number of elements in the set.

Accuracy (Acc) measurement is performed as follows [70]:

$$Acc(A, B) = \frac{TP + TN}{TP + FN + TN + FP} \tag{10}$$

where TP is the number of true positive and FN is the number of false negative classifications, TN is the number of true negative and FP is the number of false positive classifications.

4.3 Parameter selection

Both the considered methods require specific parameters tuning and also additional preliminary configuration of the datasets. The parameter selection and dataset preparation are described in the following sections.

4.3.1 Parameters of segmentation based on Gabor features and supervised Bayesian classification

For the preprocessing step, three wavelet levels should be specified. The optimal wavelet levels were studied in [71] and can be found in Table 3. Firstly, the *Soares model* was trained on the full available labeled training dataset and then tested. The model obtained after training is called the *fully trained model*. From the ARIADB dataset 40 images were selected, 20 of them were considered as the training set and the other 20 - as the testing set. The results of training the *Soares model* on the full available training set are presented in Table 3 [71].

Table 3. Results of the *Soares method* with the model trained on the full training set, where N_{train} and N_{test} are sizes of training and testing sets respectively.

Database	N_{train}	N_{test}	**Wavelet levels**	**DC**	**Accuracy**
STARE [24]	10	10	[2, 3, 6]	0.75	0.95
ARIADB [27]	20	20	[2, 5, 6]	0.61	0.93
DRIVE [25]	20	20	[2, 3, 5]	0.76	0.95
CHASEDB1 [26]	14	14	[3, 8, 9]	0.68	0.92

4.3.2 U-Net segmentation parameters

The *U-Net* architecture is implemented in Python by using high-level neural network API Keras [72]. The implementation used during the research was based on the research [73]. The experiments on the network training were conducted on GPU NVIDIA GeForce TITAN Black, 6 Gb RAM, Intel Xeon CPU E5-2680, 128 Gb RAM. To satisfy these memory resource conditions the optimal image resolution was selected to be 320×320. All the images used in the training or testing process were reduced in the resolution to the selected one. Also the sizes of the network layers were reduced taking into account the memory restrictions. The selected values can be found in Fig. 7.

During the experiments the following steps were taken for each selected dataset:

- Train the network on the fully annotated training set;
- Evaluate the fully trained model performance on the testing set;
- Train the network with active learning and with selected active iterations one training epoch per each iteration;
- Evaluate the active trained model performance in each active iteration.

The training parameters, which are to be set according to these experiments, can be found in Table 4. The size of the initial labeled set was selected according to the size of the overall training set available.

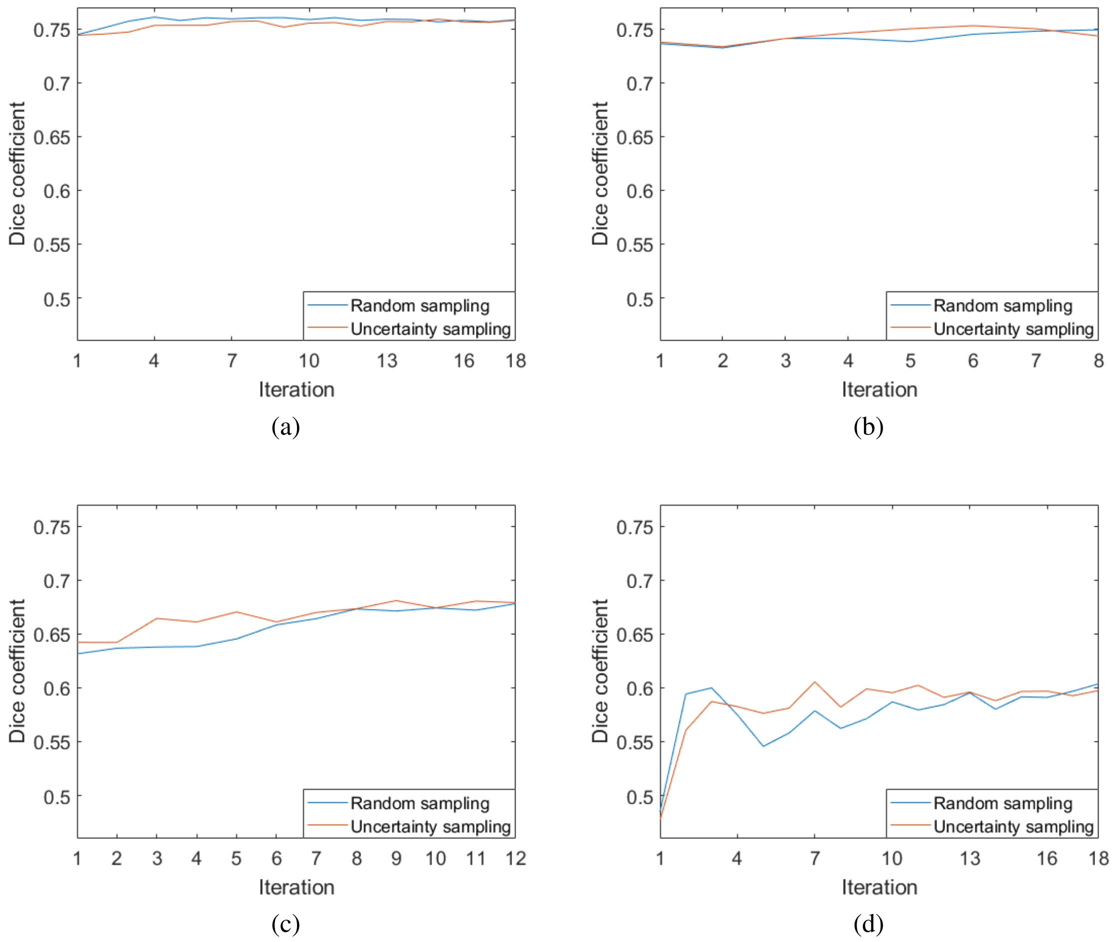
Table 4. Training parameters of the *U-Net* for each considered dataset. N_{train} is the size of the training set, N_{test} is the size of the testing set, E_{f} is the number of the full training epochs, N_{i} is the size of the initial labeled set, I_{a} is the number of active iterations, E_{a} is the number of training epochs per each active iteration.

Database	N_{train}	N_{test}	E_{f}	N_{i}	E_{i}	I_{a}	E_{a}
DRIVE [25]	30	10	200	5	10	25	6
STARE [24]	10	10	200	2	10	8	6
CHASEDB1 [26]	18	10	200	4	10	14	6
ARIADB [27]	30	10	200	5	10	25	6

4.4 Results

4.4.1 Active learning with Gabor features and Bayesian classifier

At first the model was trained on two images from the training set and then retrained with each new frame selected based on the active query. The selection of the next image to label and the following model training occur in one iteration of active learning called the *active iteration*. The results of the active learning process in comparison with random



(a)

(b)

(c)

(d)

Figure 10. Comparison of the DC for the uncertainty and random sampling (the *Soares model*): (a) DRIVE; (b) STARE; (c) CHASEDB1; (d) ARIADB.

sampling are shown in Fig. 10 for the DC and in Fig. 11 for the Acc. All the presented results oh the *Soares method* evaluation contain information of a single run.

From these results it can be noticed that the Acc and DC change in the same manner for the both random and uncertain sampling. After 2-4 active iterations these criteria remain on the same level for all considered datasets.

The results of the *Soares method* presented in the Fig. 12 indicate that the model learns features quite fast and reaches the performance of the fully trained model after the 4-6 active iterations depending on the database.

Comparison of the active learning performance and the performance of the model trained on the full available training set for the *Soares method* is presented in Fig. 12. The sample results of the *Soares model* performance for the DRIVE dataset can be found in Fig. 13.

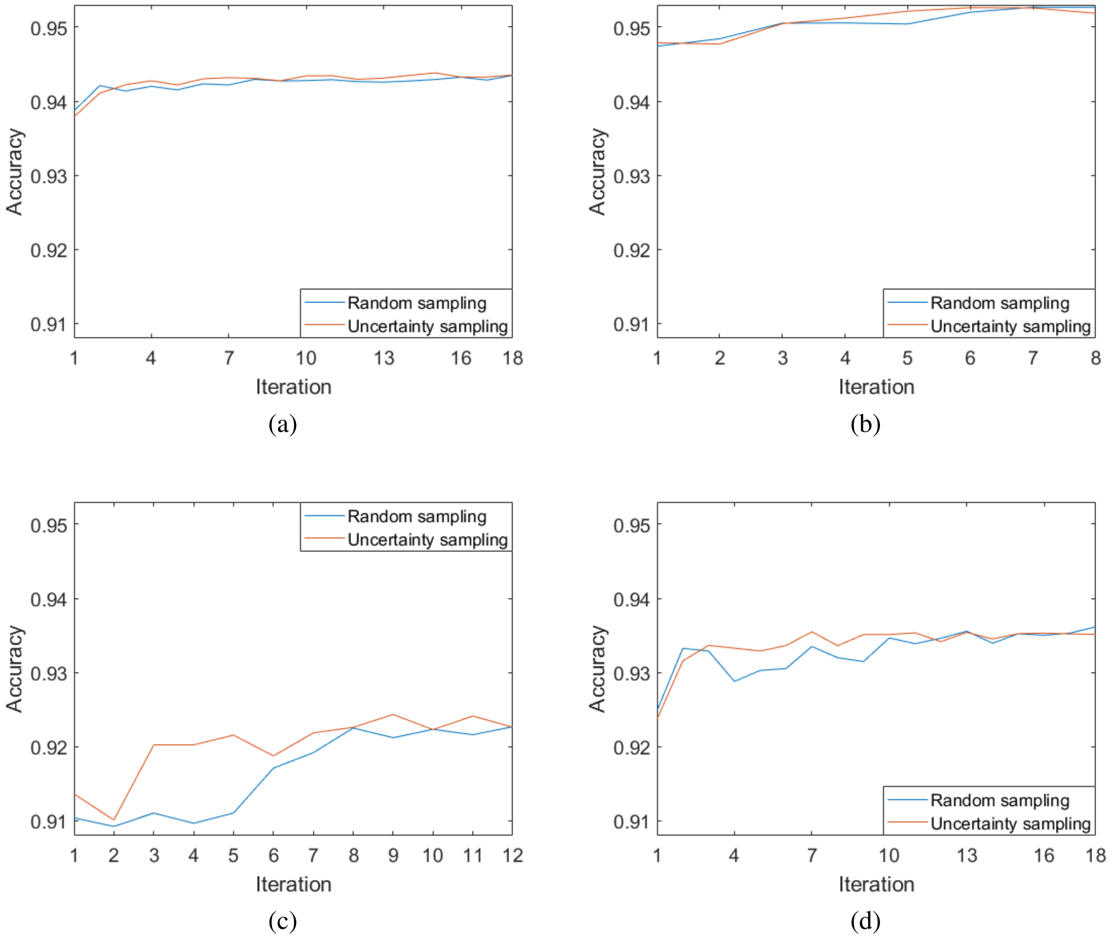
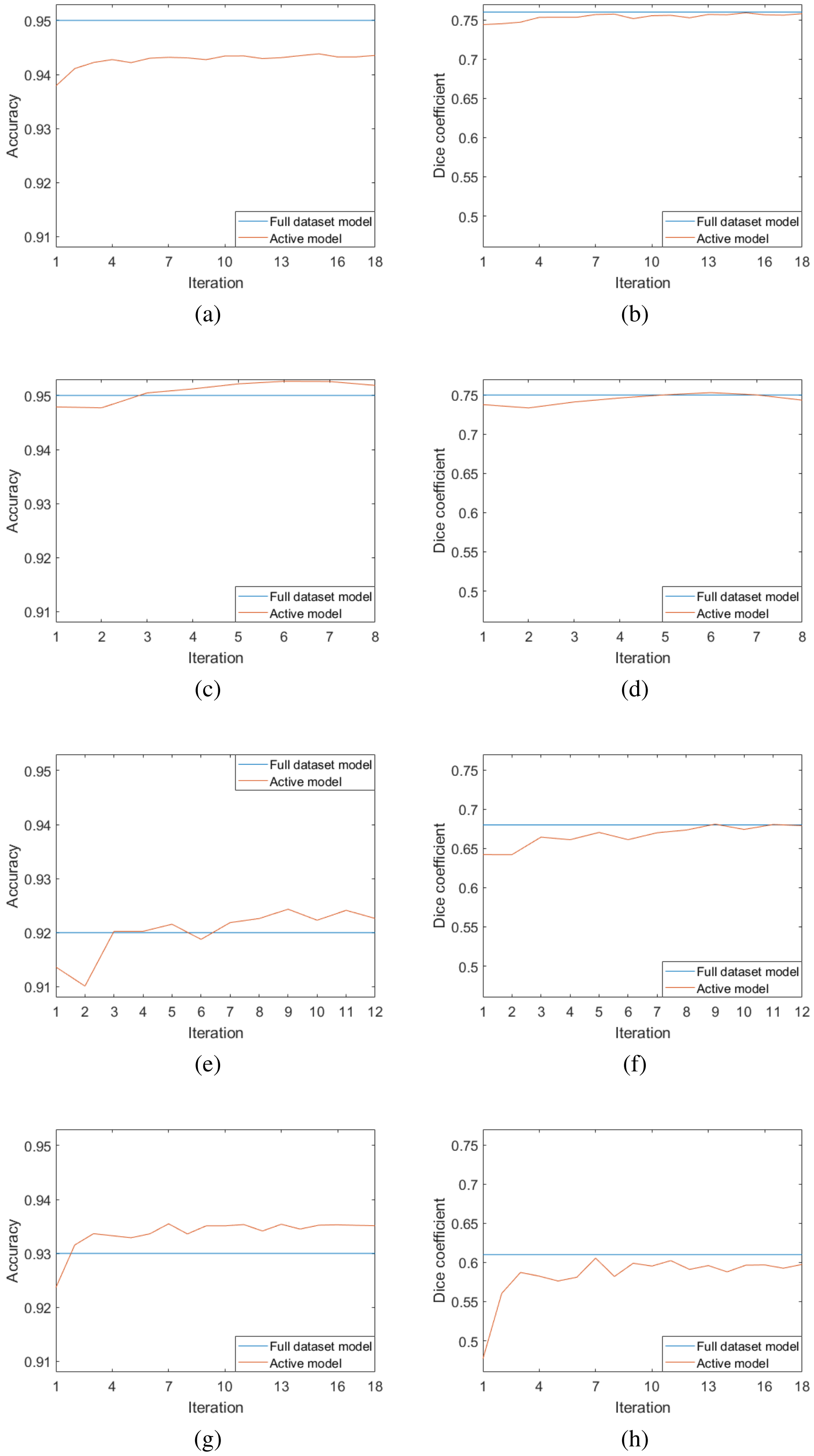


Figure 11. Comparison of the Acc value for the uncertainty and random sampling (the *Soares model*): (a) DRIVE; (b) STARE; (c) CHASEDB1; (d) ARIADB.

Having the similar results for the both tested query functions, the visual characteristics of the segmentation results differ. On the segmented images from the DRIVE in Fig. 13i and Fig. 13j obtained during the testing of the random sampling the elements of the retina round border can be seen. At the same time, the results of the uncertainty sampling in Fig. 13g and Fig. 13h are close to the fully trained model results already on the fourth active iteration.

Being the uncertainty measure for the *Soares method*, the LL was calculated for all the segmented images in the unlabeled set in each active iteration. The minimal LL value for the both considered query functions is presented in Fig. 14.

In the case of the uncertainty sampling, as it is presented in Fig. 14 the minimal LL decreases with each new selected frame whereas in the case of random sampling this does not happen.



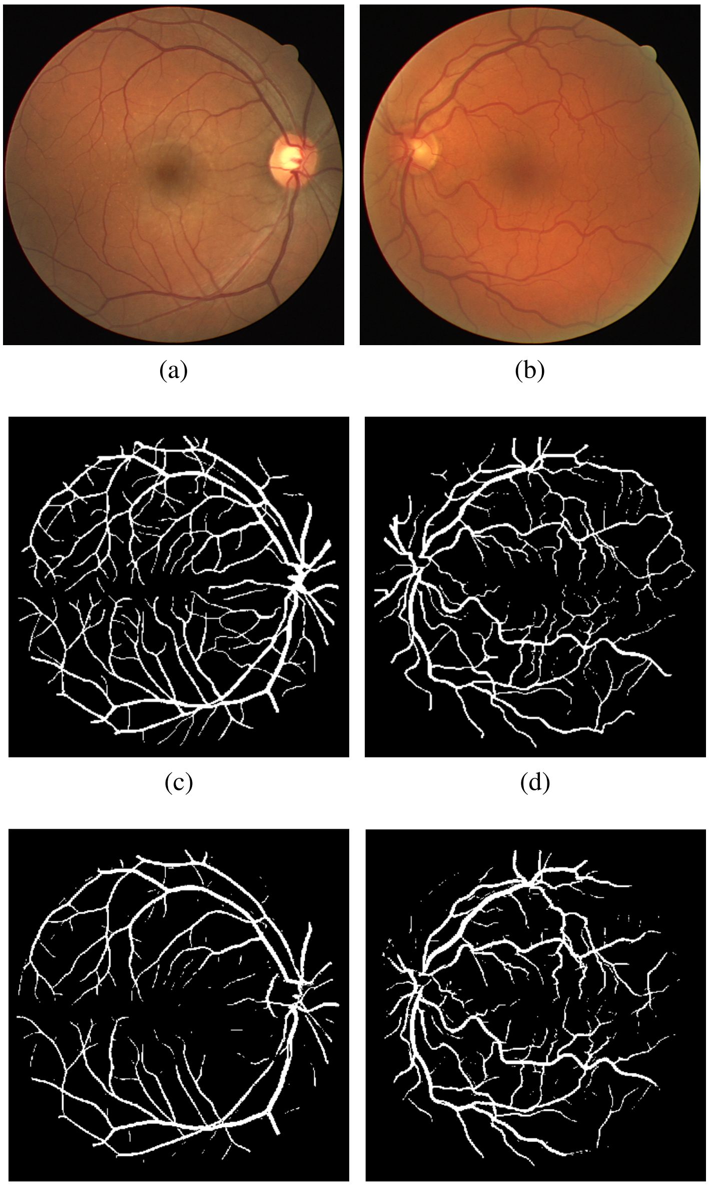
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 12. Comparison of the active learning performance and the performance of the model trained on the full available train dataset for the *Soares model* (the first column - for the Acc, the second - for the DC): (a) and (b) DRIVE; (c) and (d) STARE; (e) and (f) CHASEDB1; (g) and (h) ARIADB.

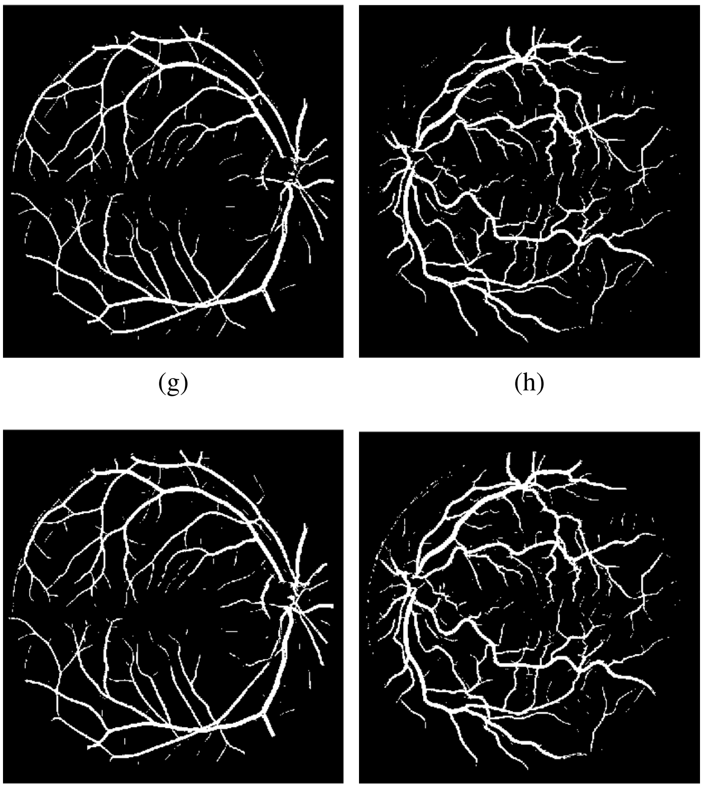


(a) (b)

(c) (d)

(e) (f)

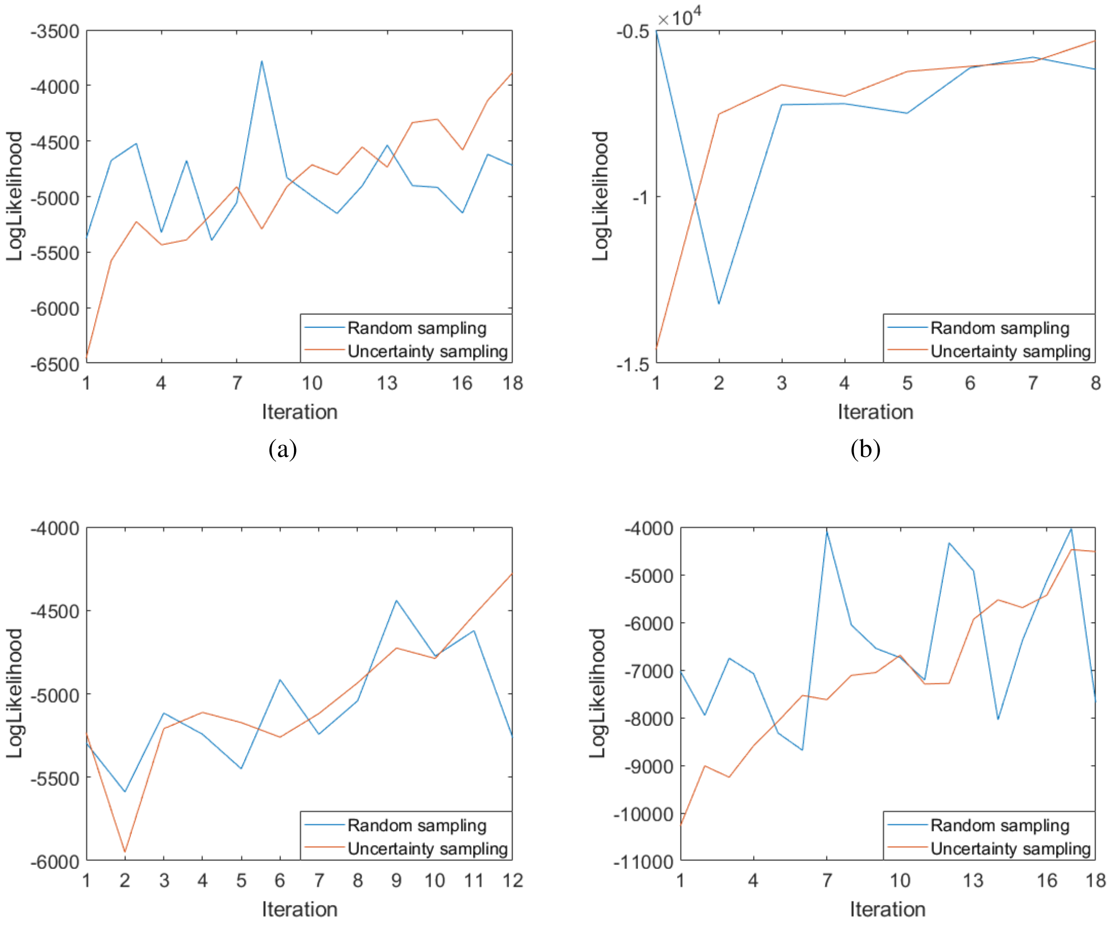
Figure 13. Example testing results of the *Soares model* on the DRIVE: (a) and (b) the input retinal images; (c) and (d) the manually segmented images; (e) and (f) the fully trained model.



(g) (h)

(i) (j)

Figure 13. (continued) Example testing results of the *Soares model* on the DRIVE: (g) and (h) the active trained model with the uncertainty sampling; (i) and (j) the active trained model with the random sampling.



(c) (d)

Figure 14. The minimal LL value among the segmented images from the unlabeled set after each iteration of the *Soares model* training: (a) DRIVE; (b) STARE; (c) CHASEDB1; (d) ARIADB.

4.4.2 Active learning with U-Net

In the case of the small amount of training data the preliminary pretraining on the patches from known annotated images can improve learning performance. *U-Net* is a fully convolutional network, which allows to pretrain it on the set of small patches and then train it on larger images without changes in the architecture. For these purposes, set of 100 patches with spatial resolution 96 × 96 pixels was made based on 5 images and their annotations in the DRIVE. Examples of these patches can be found in Fig. 15.

To obtain an expected performance level for evaluation active learning of the model, the *U-Net* was trained on the available training images for each considered dataset. The training was made with 200 epochs. The rate of the DC after each epoch can be found in Fig. 16. The trained model was tested on the testing set for each database. The average DC value for both 200 and 300 pretraining epochs can be found in Table 5. There were

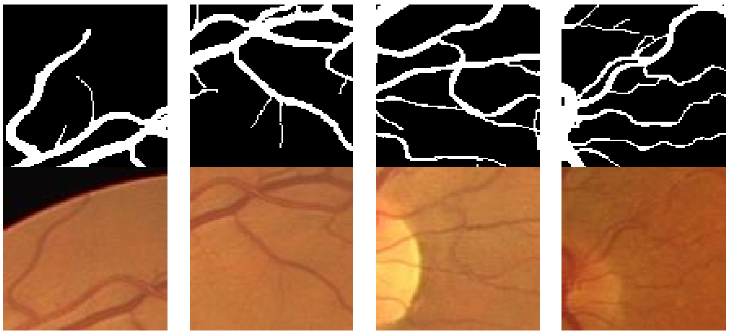


Figure 15. Example patches for pretraining of the *U-Net*.

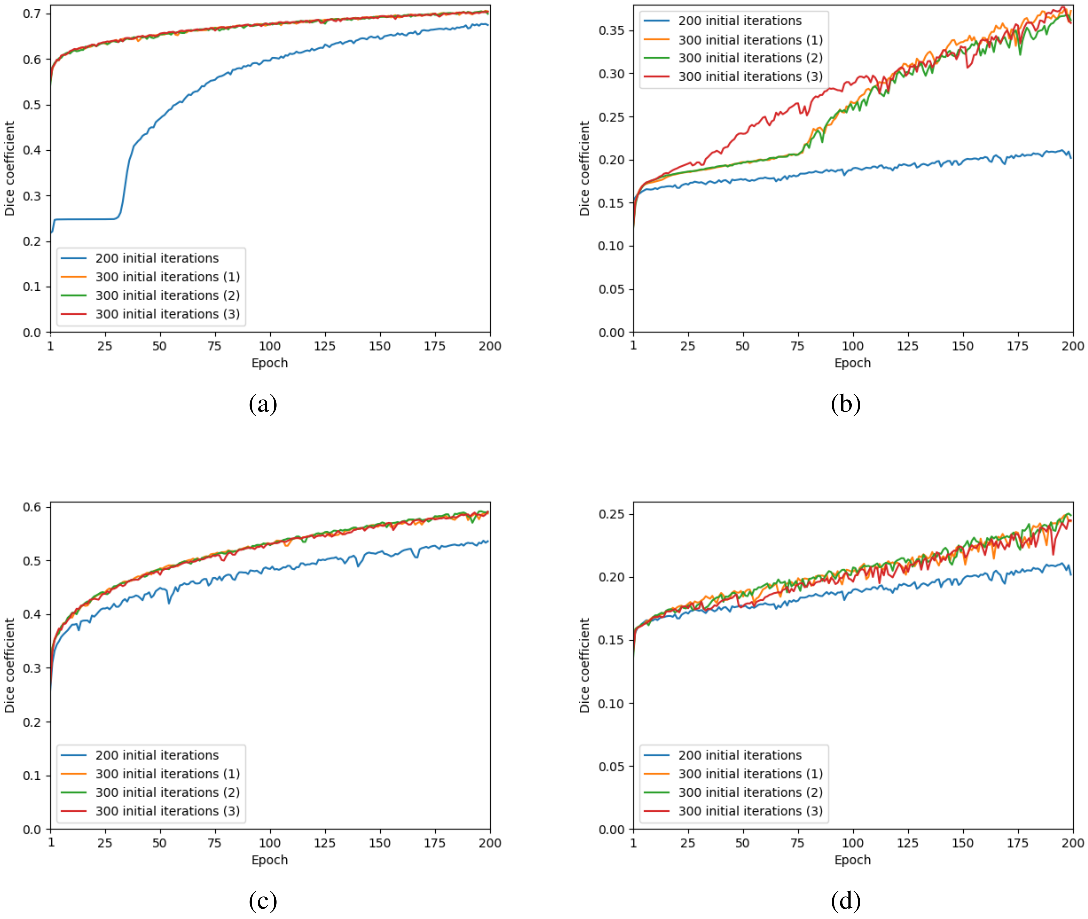
four training runs for the both standard and active training to verify the results: one run for the 200 pretrained epochs option and three runs for the 300 pretrained epochs option.

Table 5. The average DC obtained during the testing of the fully trained *U-Net* model for 200 and 300 pretraining epochs.

Database	**DC (200)**	**DC (300)**
DRIVE [25]	0.63	0.64
STARE [24]	0.10	0.10
CHASEDB1 [26]	0.24	0.35
ARIADB [27]	0.19	0.18

For the active learning, the pretrained model was used. Two options were examined: 200 and 300 pretraining epochs. The active learning performance was assessed for the both cases. For the latter option, three runs of the model training were conducted. After each active iteration one new frame was annotated and added to the train set. The number of active iterations a training epochs per each iteration can be seen in Table 4. The DC during the training in comparison with the average testing performance of the fully trained model can be found in Fig. 17.

The *U-Net* performance evaluation was conducted according to the parameters presented in the Table 4. From the plots in Fig. 17, it can be seen, that in the first four active iterations during the training on DRIVE, the DC already reaches the full model expected value, but further extension of the training set leads to it's decrease until the 14th iteration. On the other hand, during the training the model on ARIADB, the performance reaches its peak value in the middle of the active iterative process and then only decreases till the end. For the rest datasets, the DC remains almost in the same level for all the active



(a)

(b)

(c)

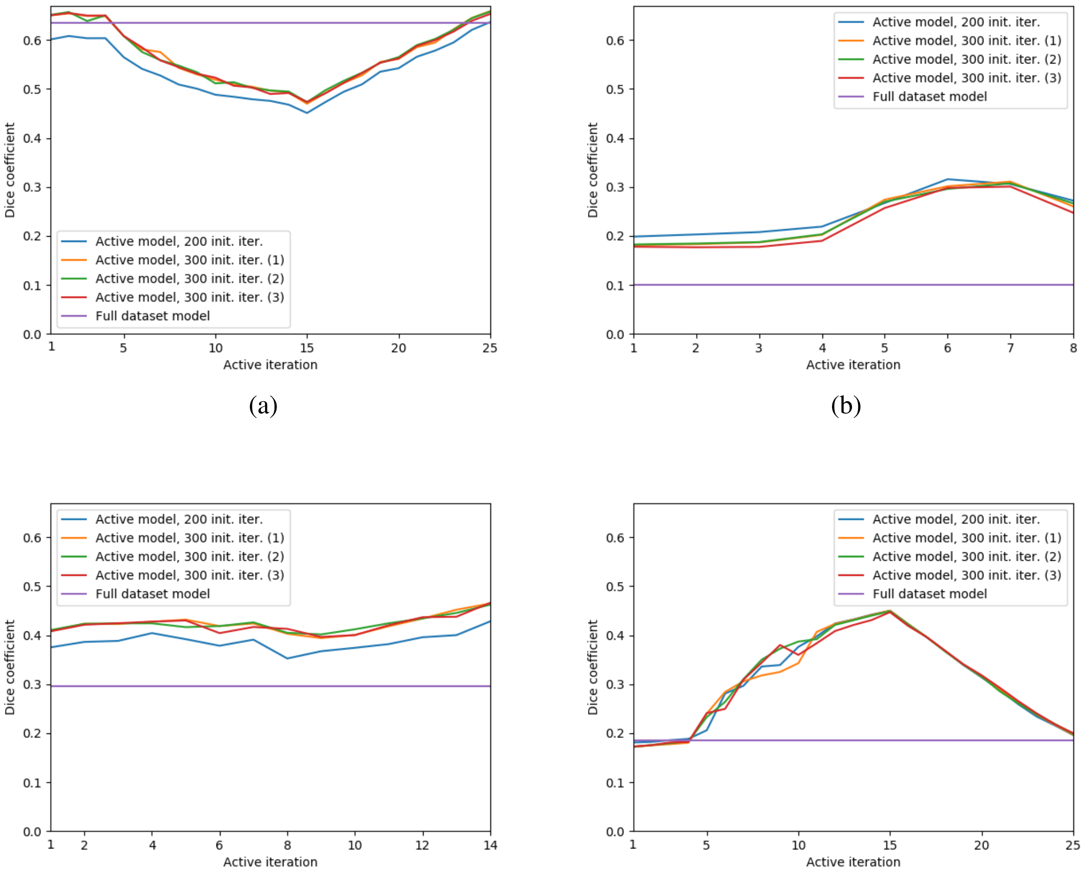
(d)

Figure 16. Changes in the DC value during training of the *U-Net* on the full available training set: (a) DRIVE; (b) STARE; (c) CHASEDB1; (d) ARIADB.

iterations.

The *U-Net* model obtained after every active iteration was tested on the training set formed for each considered database (the parameters can be found in Table 4).The testing run was repeated 10 times. An average DC and its standard deviation within the testing run in comparison with the average testing performance of the fully trained model are shown in Fig. 18. As it can be seen in Fig. 18, the DC reaches the expected value faster in the case of 300 pretraining epochs than in the case of 200 epochs.

The sample testing results from the fourth active model are presented in Fig. 19 in comparison with the testing results of the full trained model. The active model segmentation maps Fig. 19e and Fig. 19f have no elements of the round border. The thin vessels are almost absent, but the wide vessels are better segmented and connected. As it can be seen in Fig. 19, segmentation results in Fig. 19g and Fig. 19h are slightly better in the thin



(a)

(b)

(c)

(d)

Figure 17. Changes in the DC during active training of the *U-Net* (init. iter. - initial iterations): (a) DRIVE; (b) STARE; (c) CHASEDB1; (d) ARIADB.

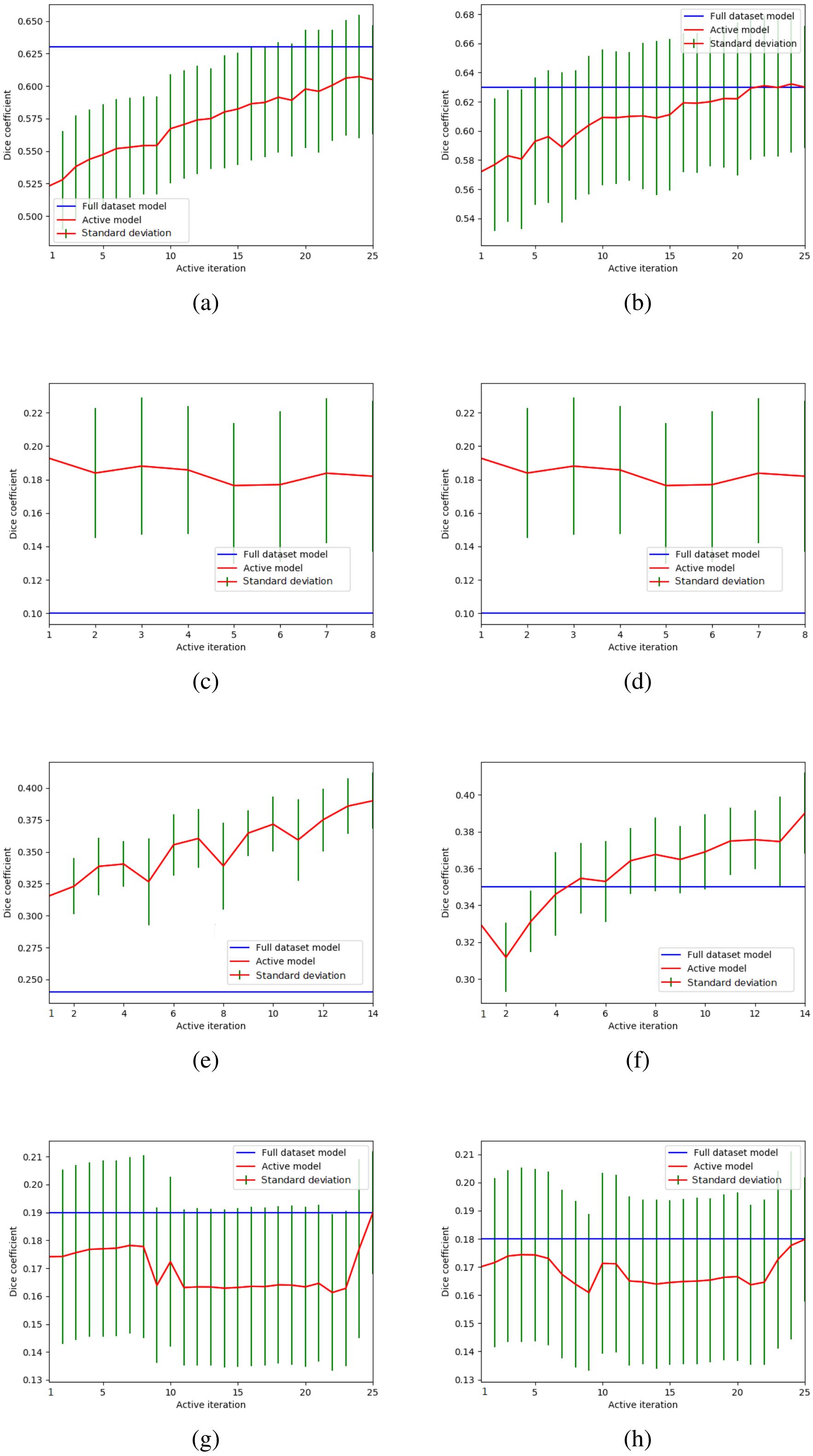
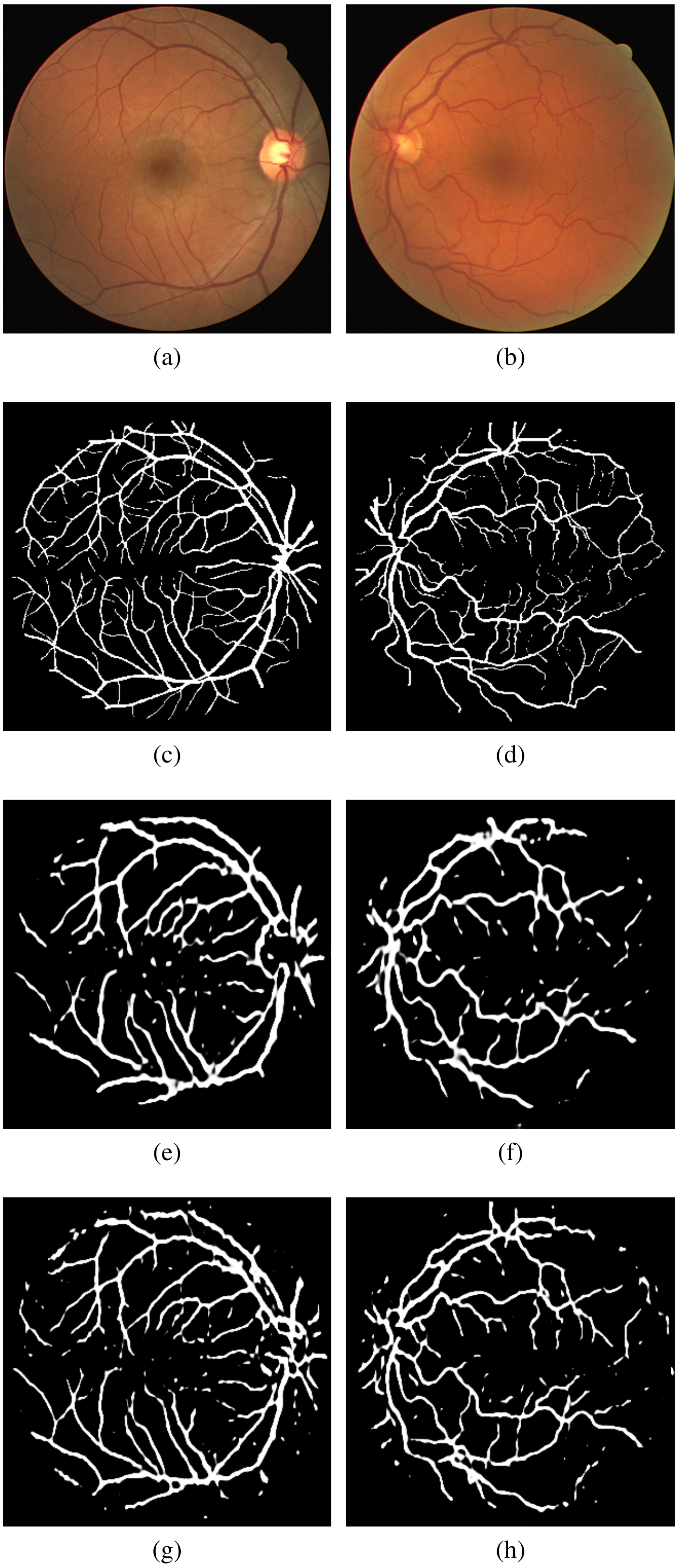


Figure 18. The average DC value and its variation during testing on of each active *U-Net* model for 200 and 300 pretraining epochs: (a) 200 and (b) 300 for DRIVE; (c) 200 and (d) 300 for STARE; (e) 200 and (f) 300 for CHASEDB1; (g) 200 and (h) 300 for ARIADB.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 19. Example testing results of the *U-Net* mode after four active training iterations and the model trained on the full available DRIVE training set (30 images) with 200 epochs: (a) and (b) the input retinal images; (c) and (d) the manually segmented images; (e) and (f) the active trained model; (g) and (h) the fully trained model.

vessel segmentation, but they also contain parts of the round edges in the border.

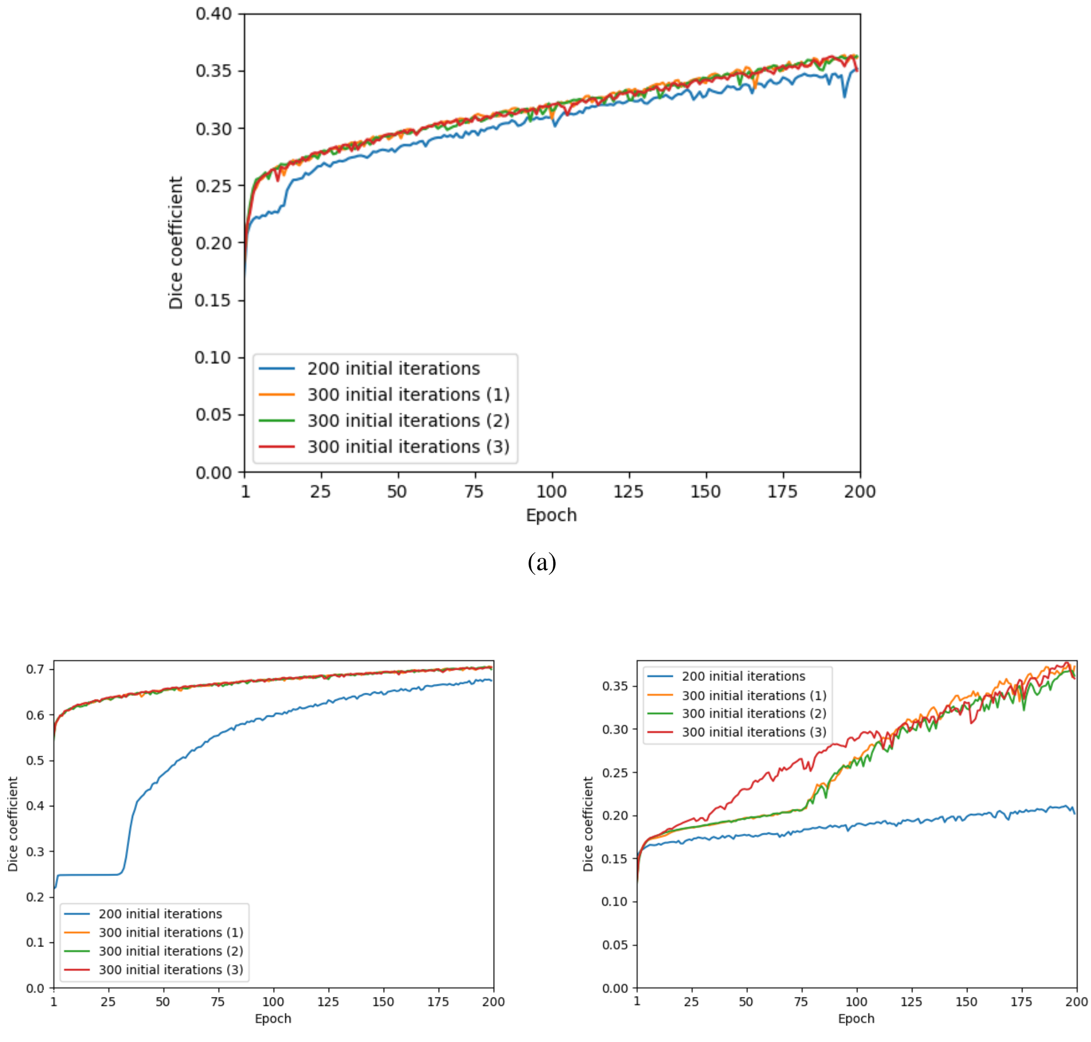
The performance level is different for all the considered databases. The only dataset training on which could provide the DC higher than 0.6 is DRIVE. In images from this database the retina is presented in a full round shape, the illumination is even and the blood vessels can be easily seen. Next, but with less acceptable performance was CHASEDB1. In this case the resolution was decreased almost three times compared to the original which means that there is a place to improve the segmentation accuracy. Also according to [26], the retinal images were collected from 10-year old children, hence there can be seen other vessels, not only the retina vessels. The retina in STARE and ARIADB images is cropped. The illumination in these images is brighter compared to the other two datasets. Also the STARE dataset contains only 20 images, 10 of them were used for training. Taking into account the before-mentioned factors, this amount of data may be insufficient for a model to learn the appropriate features.

4.4.3 Evaluation the U-Net active model performance on mixed datasets

In order to expand the training data, a set consisting of 30 images from DRIVE and 10 images from STARE was created. The set of 100 patches with the resolution 96×96 based on 5 images from the new mixed dataset was also used to pretrain the *U-Net* with two options: 200 and 300 training epochs with three runs for the latter one. Having these initial weights, the network was trained on the full dataset of 40 images with 200 training epochs. The DC value changes are shown in Fig. 20. In the following active learning there were 35 iterations. As a result, 35 trained active models were obtained. The DC during the training as well as the average testing performance of the fully trained model can be seen in Fig. 21.

The performance of the *U-Net* during the training on the mixed dataset presented in Fig. 21a has a similar trend as the performance on ARIADB (Fig. 21b). The overall performance is getting better during the training, but it is quite low on the testing even for the model trained on the full dataset: the DC does not exceed 0.30 (Fig. 22).

Each model performance was evaluated on the DRIVE testing set consisting of 10 images. As it was in the previous experiments, the testing run was repeated 10 times. The average DC value for the both active and fully trained models, as well as its standard deviation among the testing runs for the *U-Net* on each iteration are shown in Fig. 22.

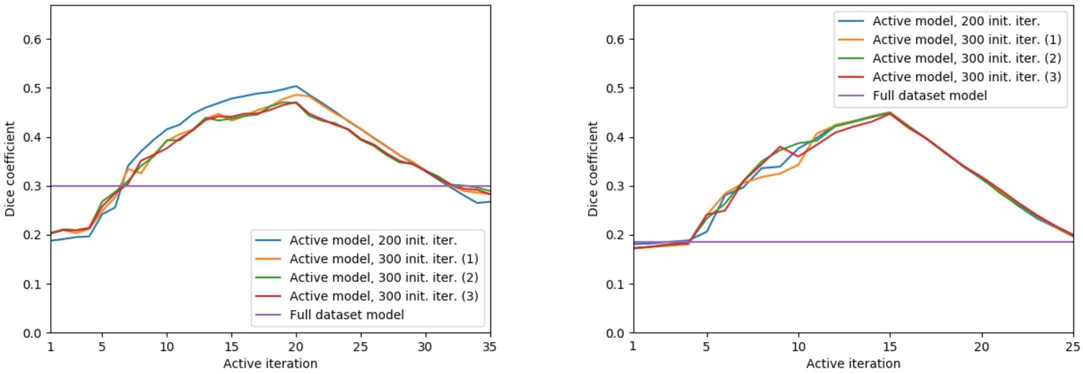


(a)

(b)

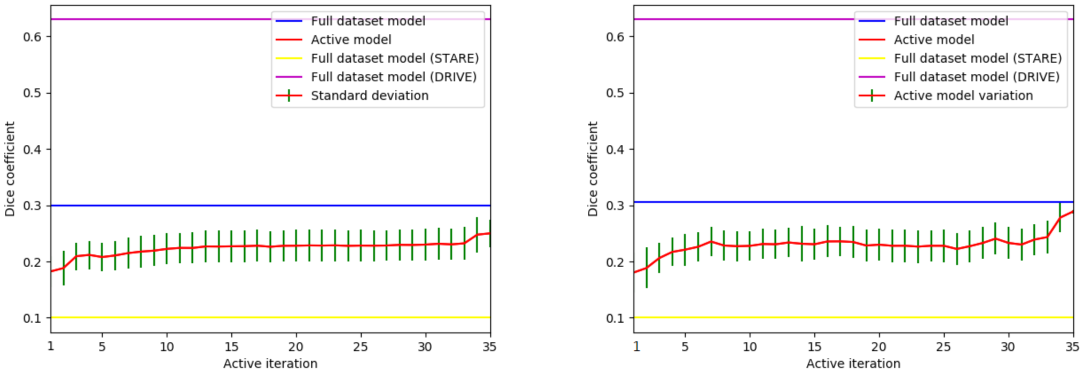
(c)

Figure 20. The DC during training of the *U-Net*on the mixed set (a) (40 images), DRIVE (b) and STARE (c) datasets.



(a) (b)

Figure 21. Changes in the DC value during active training of the *U-Net*on the mixed DRIVE and STARE dataset (a) in comparison with the performance on the ARIADB set (b).



(a) (b)

Figure 22. The average DC and its standard deviation during testing of each active *U-Net* model for the mixed DRIVE and STARE dataset: (a) 200 pretraining epochs; (b) 300 pretraining epochs.

5 DISCUSSION

In the research, the active learning methods for the ground truth collection in the task of the retina blood vessels segmentation were proposed. The goal was to achieve the success rate of the retinal image segmentation which would be similar to the standard approaches rate. Selection of the proper database which contains necessary amount of the annotations and appropriate image quality is one of the challenges.

5.1 Study results

Two segmentation methods were studied during the research. These are the *Soares method* and the CNN called *U-Net*. For the testing purposes, separation the training and validation set was conducted in all the considered databases. The both models were trained in the standard way having the fully labeled training set available. In order to have an expected model performance level, these models were evaluated on the validation set.

For the *Soares method*, the uncertainty sampling method was proposed and compared with the random sampling. The expected performance level was achieved in 4-6 active iterations depending on the dataset for the both active query functions. The DC value of the model trained on the DRIVE dataset becomes close the expected DC value first (in the fourth active iteration), however, when it comes to the Acc evaluation, the first one is the model trained on ARIADB (the second active iteration). The Acc value exceeds the expected rate while training the model on all the databases except DRIVE. Nevertheless, for all the considered datasets the DC value exceeds 0.6 and the Acc value exceeds 0.91 already within the first active iterations. The best performance and learning rate was achieved while training on DRIVE database.

According to the obtained results, it can be seen that differences between the uncertainty and random sampling application are minor. However, visible characteristics of the images segmented by the model trained with the uncertainty sampling are better. The minimal LL value among the segmented images from the unlabeled set decreased in this case, whereas in the case of the random sampling there was no noticeable dependency. This trend may indicate that the grounded selection of the next candidate to the training set has an impact on the model performance.

For increasing of the learning rate, the *U-Net* was also pretrained with the set of small patches, which were made of several images from the training set. The impact of the number of the pretraining epochs is noticeable, but the specific value should be carefully selected depending on the number of the labeled images available.

From the obtained results, it can be seen that there is no common trend in the active *U-Net* model performance during the training process on the different databases. When evaluating the testing performance (Fig. 18 on page 38), the DC in the certain iteration becomes close or exceeds the expected from the fully trained model value for all the datasets. However, based on the analysis of the differences between the images from the databases (Fig. 8 on page 25), it can be noticed that the rate and quality of the *U-Net* learning depend on the input image quality. The only result with the DC value exceeding 0.6 was achieved while training on the DRIVE database. The segmented images obtained during the active learning on this dataset do not contain the round edges as it is in the case of the fully trained model. This tendency indicates that for the *U-Net* model, as well as for the *Soares model*, the grounded and ordered selection of the next candidate to the training set also influences on the performance.

To examine the possibility to train the network on several databases, the mixed dataset was created based on DRIVE and STARE. All the tendencies observed during the previous experiments were noticed also in this case. Nevertheless, the average segmentation performance was quite low and the DC did not exceed 0.30. The images of STARE (Fig. 8b on page 25) contain the retina not in a full round shape, but in the cropped one, whereas in DRIVE (Fig. 8a on page 25) the eye fundus is presented in the full shape. These differences may have a significant impact on the network learning performance while training on the mixed set.

5.2 Future work

For the *U-Net* experiments, the initial spatial resolution of images was decreased, whereas for evaluating the *Soares method* performance, the original resolution was used. This can have an impact on the *U-Net* feature extraction, especially when it comes to the thin vessel segmentation. For the further research, testing the *U-Net* performance on the images of the original size would be promising since the current results are not so good. Also there is a possibility to change the network architecture, for example, to increase the sizes of the layers. It should be taken into account and examined more carefully.

In order to make the query function more accurate, one can use a mask image which can reduce the region of interest when calculating the uncertainty to the retina part of the image only. This mask can be presented in the form of a binary image where 1 corresponds to the retina and 0 to the background. The influence of such mask application on the active learning performance is a question to study further.

Another direction to a more careful research is combining several retinal image databases into a single one. It could increase the amount of information on which the model learns the features. The conducted experiments on the mixed dataset have shown the necessity to preprocess the images from different databases. Hence, in order to conduct this research, one needs to make the mixed dataset homogeneous, for example, correct the color, illumination and/or resolution.

In some of the considered databases, annotations from several experts are presented. This feature was not used in the research, however, the fusion of the multiple experts annotations may have an impact on the informativeness of the ground truth.

6 CONCLUSION

For a medical expert, an image of the eye fundus is enough to diagnose many diseases. In order to accelerate the diagnosis process and also to help the medical doctor enable a wide screening, automatic image processing methods were applied. A large annotated dataset and proper ground truth images or other information are required while using these methods.

Since one of the important steps of the automated retina-based diagnosis is the blood vessels segmentation in this research two different supervised segmentation approaches was examined. The first approach (*Soares method*) was based on the Gabor filtering and Bayesian classification. The second approach involved the convolutional neural network *U-Net* to segment the retinal images. Active learning was studied to speed up the collection of the ground truth. As the query function, poll-based uncertainty sampling was selected. For the *Soares method* method, own uncertainty measuring procedure based on the probability map was proposed. In the case of the *U-Net*, Monte Carlo Dropout was used to calculate the informativeness of the image.

Four publicly available datasets containing the ground truth in the form of the segmentation map were selected for the evaluation purposes. For the additional experiments, a mixed dataset was created based on two considered retinal image databases. Having the accuracy and the dice similarity coefficient as a quality criteria the segmentation, algorithms were tested in two modes: training on the fully annotated dataset and training with the active uncertainty sampling. In was noted that the *Soares method* is less dependent on the input image quality than the *U-Net* model.

The conducted experiments have shown that by means of the active learning the compact representation of the training set based on the most informative images is possible. The initial results have shown that usage of active learning has an impact on the training process and allows to train the model better, hence it can be effectively applied in the retina blood vessels segmentation task.

REFERENCES

[1] Gordon Guyatt, John Cairns, David Churchill, Deborah Cook, Brian Haynes, Jack Hirsh, Jan Irvine, Mark Levine, Mitchell Levine, Jim Nishikawa, et al. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA*, 268(17):2420–2425, 1992.

[2] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

[3] Oliver Faust, Rajendra Acharya, Eddie Yin-Kwee Ng, Kwan-Hoong Ng, and Jasjit S Suri. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. *Journal of Medical Systems*, 36(1):145–157, 2012.

[4] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[5] Wong Li Yun, U Rajendra Acharya, Yedatore V Venkatesh, Caroline Chee, Lim Choo Min, and E Yin Kwee Ng. Identification of different stages of diabetic retinopathy using retinal optical images. *Information Sciences*, 178(1):106–121, 2008.

[6] Katia Estabridis and Rui JP de Figueiredo. Automatic detection and diagnosis of diabetic retinopathy. In *IEEE International Conference on Image Processing, 2007*, volume 2, pages II–445. IEEE, 2007.

[7] Clara I Sánchez, Meindert Niemeijer, Thessa Kockelkorn, Michael D Abràmoff, and Bram van Ginneken. Active learning approach for detection of hard exudates, cotton wool spots, and drusen in retinal images. In *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260, page 72601I. International Society for Optics and Photonics, 2009.

[8] Clara I Sánchez, Meindert Niemeijer, Michael D Abràmoff, and Bram van Ginneken. Active learning for an efficient training strategy of computer-aided diagnosis systems: application to diabetic retinopathy screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–610. Springer, 2010.

[9] Tomi Kauppi. *Eye fundus image analysis for automatic detection of diabetic retinopathy*. PhD thesis, Lappeenranta University of Technology, 11 2010.

[10] Picture of the eye. `https://en.wikipedia.org/wiki/Eye`. Online; accessed 25 January 2018.

[11] Richard F Spaide, James M Klancnik, and Michael J Cooney. Retinal vascular layers imaged by fluorescein angiography and optical coherence tomography angiography. *JAMA Ophthalmology*, 133(1):45–50, 2015.

[12] Muthu Rama Krishnan Mookiah, U Rajendra Acharya, Chua Kuang Chua, Choo Min Lim, EYK Ng, and Augustinus Laude. Computer-aided diagnosis of diabetic retinopathy: A review. *Computers in Biology and Medicine*, 43(12):2136–2155, 2013.

[13] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Nature Publishing Group, Scientific Reports 2016*, 6:24454, 2016.

[14] Ming Li and Zhi-Hua Zhou. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):1088–1098, 2007.

[15] Chisako Muramatsu, Kohei Nishimura, Takeshi Hara, and Hiroshi Fujita. Preliminary investigation on CAD system update: Effect of selection of new cases on classifier performance. In *Medical Imaging 2013: Computer-Aided Diagnosis*, volume 8670, page 86701T. International Society for Optics and Photonics, 2013.

[16] Mohamed F El-Bab, Nashaat Shawky, Ali Al-Sisi, and Mohamed Akhtar. Retinopathy and risk factors in diabetic patients from Al-Madinah Al-Munawarah in the Kingdom of Saudi Arabia. *Clinical Ophthalmology*, 6:269, 2012.

[17] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[18] Vladimir Vapnik, Steven E Golowich, and Alex J Smola. Support vector method for function approximation, regression estimation and signal processing. In *Advances in Neural Information Processing Systems*, pages 281–287, 1997.

[19] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[20] Tomi Kauppi, Valentina Kalesnykiene, Iiris Sorri, Asta Raninen, Raija Voutilainen, Joni Kamarainen, Lasse Lensu, and Hannu Uusitalo. Diaretdb1 v2.1 – diabetic retinopathy database and evaluation protocol. `http://www2.it.lut.fi/project/imageret/diaretdb1_v2_1/`. Online; accessed 25 January 2018.

[21] Tomi Kauppi, Valentina Kalesnykiene, Joni Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raijaand Voutilainen, J. Pietilä, Heikki Kälviäinen, and Hannu Uusitalo. DIARETDB1 diabetic retinopathy database and evaluation protocol. In *Proceedings of the Medical Image Understanding and Analysis*, pages 61–55. Aberystwyth, UK, 2007.

[22] Michael D. Abramoff. ROC project website. `http://webeye.ophth.uiowa.edu/ROC/`. Online; accessed 28 January 2018.

[23] Meindert Niemeijer, Bram Van Ginneken, Michael J Cree, Atsushi Mizutani, Gwénolé Quellec, Clara I Sánchez, Bob Zhang, Roberto Hornero, Mathieu Lamard, Chisako Muramatsu, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 29(1):185–195, 2010.

[24] M.D Michael Goldbaum. STARE project website (2003, july). `http://www.cecas.clemson.edu/~ahoover/stare`. Online; accessed 25 January 2018.

[25] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

[26] Christopher G Owen, Alicja R Rudnicka, Robert Mullen, Sarah A Barman, Dorothy Monekosso, Peter H Whincup, Jeffrey Ng, and Carl Paterson. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (CAIAR) program. *Investigative Ophthalmology & Visual Science*, 50(5):2004–2010, 2009.

[27] Damian JJ Farnell, FN Hatfield, P Knox, M Reakes, S Spencer, D Parry, and SP Harding. Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. *Journal of the Franklin Institute*, 345(7):748–765, 2008.

[28] EYEPACS, LLC. EYEPACS project website. `http://www.eyepacs.com/`. Online; accessed 28 January 2018.

[29] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Laÿ, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al.

Feedback on a publicly distributed image database: the MESSIDOR database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[30] The University of Birmingham School of Computer Science. CMIF project website. `http://www.cs.bham.ac.uk/research/projects/fundus-multispectral/`. Online; accessed 28 January 2018.

[31] Iain B Styles, A Calcagni, Ela Claridge, Felipe Orihuela-Espina, and JM Gibson. Quantitative analysis of multi-spectral fundus images. *Medical Image Analysis*, 10(4):578–597, 2006.

[32] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. Blood vessel segmentation methodologies in retinal images – a survey. *Computer Methods and Programs in Biomedicine*, 108(1):407–433, 2012.

[33] João VB Soares, Jorge JG Leandro, Roberto M Cesar, Herbert F Jelinek, and Michael J Cree. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, 25(9):1214–1222, 2006.

[34] Michal Sofka and Charles V Stewart. Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures. *IEEE Transactions on Medical Imaging*, 25(12):1531–1546, 2006.

[35] Uyen TV Nguyen, Alauddin Bhuiyan, Laurence AF Park, and Kotagiri Ramamohanarao. An effective retinal blood vessel segmentation method using multi-scale line detection. *Pattern Recognition*, 46(3):703–715, 2013.

[36] Peter Bankhead, C Norman Scholfield, J Graham McGeown, and Tim M Curtis. Fast retinal vessel detection and measurement using wavelets and edge location refinement. *PLOS ONE*, 7(3):e32435, 2012.

[37] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. Trainable COSFIRE filters for vessel delineation with application to retinal images. *Medical Image Analysis*, 19(1):46–57, 2015.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[39] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[40] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[41] Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[42] Rosa L Figueroa, Qing Zeng-Treitler, Long H Ngo, Sergey Goryachev, and Eduardo P Wiechmann. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, 19(5):809–816, 2012.

[43] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[44] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[45] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 287–294. ACM, 1992.

[46] Nicholas Roy and Andrew McCallum. Toward optimal active learning through Monte Carlo estimation of error reduction. *International Conference on Machine Learning, Williamstown*, pages 441–448, 2001.

[47] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[48] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[49] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296, 2008.

[50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[51] Yonatan Geifman and Ran El-Yaniv. Deep Active Learning over the Long Tail. *International Conference on Learning Representations*, 2018.

[52] Yann LeCun. MNIST database page. `http://yann.lecun.com/exdb/mnist/`. Online; accessed 10 May 2018.

[53] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

[54] Yarin Gal and Zoubin Ghahramani. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. In *4th International Conference on Learning Representations (ICLR) Workshop Track*, 2016.

[55] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[56] Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(Feb):255–279, 2012.

[57] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(May):1605–1641, 2010.

[58] Kaito Fujii and Hisashi Kashima. Budgeted stream-based active learning via adaptive submodular maximization. In *Advances in Neural Information Processing Systems*, pages 514–522, 2016.

[59] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203, 2014.

[60] Alexander Narr, Rudolph Triebel, and Daniel Cremers. Stream-based active learning for efficient and adaptive classification of 3D objects. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 227–233. IEEE, 2016.

[61] Antonio Foncubierta Rodríguez and Henning Müller. Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia*, pages 9–14. ACM, 2012.

[62] Figure Eight Inc. CrowdFlower platform. `https://www.figure-eight.com/`. Online; accessed 07 May 2018.

[63] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, 2016.

[64] Marc Górriz Blanch. Active deep learning for medical imaging segmentation. B.S. thesis, Universitat Politècnica de Catalunya, 2017.

[65] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[66] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[67] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *28th British Machine Vision Conference*, 2017.

[68] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[69] Th Sorenson. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *K Dan Vidensk Selsk Biol Skr*, 5:1–34, 1948.

[70] David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.

[71] Pavel Vostatek, Ela Claridge, Hannu Uusitalo, Markku Hauta-Kasari, Pauli Fält, and Lasse Lensu. Performance comparison of publicly available retinal blood vessel segmentation methods. *Computerized Medical Imaging and Graphics*, 55:2–12, 2017.

[72] Google. Keras: The Python Deep Learning library. `https://keras.io/`. Online; accessed 07 May 2018.

[73] Marc Górriz Blanch. Active deep learning for medical imaging segmentation. `https://github.com/marc-gorriz/CEAL-Medical-Image-Segmentation`. Online; accessed 07 May 2018.