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

Master’s Thesis

Fedor Turbabin

DETECTION OF DIABETIC RETINOPATHY LESIONS FROM SPECTRAL RETINAL IMAGES

Examiners: Lasse Lensu D.Sc. (Tech.)
Rifkhat Aleev D. Sc. (Phys.-Math.)

Supervisors: Lasse Lensu D.Sc. (Tech.)
Rifkhat Aleev D. Sc. (Phys.-Math.)
ABSTRACT

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

Fedor Turbabin

Detection of Diabetic Retinopathy Lesions from Spectral Retinal Images

Master’s Thesis

2017

56 pages, 18 figures and 3 tables.

Examiners: Lasse Lensu D.Sc. (Tech.)
Rifkhat Aleev D. Sc. (Phys.-Math.)

Keywords: diabetic retinopathy, lesion detection, retinal imaging, hyperspectral imaging, medical image analysis, convolutional neural networks, deep learning

Long-term diabetes may lead to a diabetic retinopathy, an eye disease, which is one of the most frequent causes of blindness. Its early symptoms can be detected from the photos of retina, thus, making the investigation into the automatic lesion detection method to be a valuable undertaking. Hyperspectral images provide additional information about the characteristics of the imaging target, which may improve accuracy in comparison to simple color images, but at the same time introduce problems of higher dimensionality. The purpose of this work is to develop an effective algorithm for detecting diabetic retinopathy lesions and find out whether hyperspectral images are beneficial for diabetic retinopathy diagnosis in comparison to colour images. The experiments provide evidence that proper algorithmic use of spectral images allows to achieve significantly better results.
CONTENTS

1 INTRODUCTION 7
  1.1 Background ......................................................... 7
  1.2 Objectives and delimitations ................................. 8
  1.3 Structure of the thesis ......................................... 9

2 DIAGNOSIS OF DIABETIC RETINOPATHY 10
  2.1 Eye and retinopathy ............................................. 10
  2.2 Fundus image processing for retinopathy diagnosis .......... 12
  2.3 Retinal image databases ......................................... 13

3 LESION DETECTION AND HYPERSPECTRAL IMAGE PROCESSING 16
  3.1 Lesion detection and severity grading ....................... 17
  3.2 HSI feature extraction and classification ................. 19
  3.3 Summary .......................................................... 21

4 NEURAL NETWORKS 23
  4.1 Challenges and motivations .................................. 23
  4.2 General algorithm ............................................... 24
    4.2.1 Fully-connected networks .................................. 24
    4.2.2 Convolutional networks .................................... 24
  4.3 Activation functions ............................................ 25
  4.4 Loss functions .................................................. 27
  4.5 Methods for parameters update ............................... 28
  4.6 Regularization ................................................... 29
  4.7 Normalization ................................................... 31

5 EXPERIMENTAL SETUP 33
  5.1 Network architecture .......................................... 33
  5.2 Sampling strategy ............................................... 35
  5.3 Data preprocessing and augmentation ......................... 38

6 EXPERIMENTS AND RESULTS 41
  6.1 Performance on hyperspectral images ....................... 41
  6.2 Performance on RGB images .................................... 46
    6.2.1 Evaluation on DiaRetDB2 database ...................... 46
    6.2.2 Evaluation on BristolDB database ...................... 47

7 DISCUSSION AND FUTURE WORK 49
7.1 Discussion of the results ........................................ 49
7.2 Future work ...................................................... 50

8 CONCLUSION ..................................................... 51

REFERENCES ....................................................... 52
# ABBREVIATIONS AND SYMBOLS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>Batch normalization</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional random fields</td>
</tr>
<tr>
<td>D-SIFER</td>
<td>Derivative-Based Scale Invariant Image Feature Detector With Error Resilience</td>
</tr>
<tr>
<td>DiaRetDB1</td>
<td>Diabetic retinopathy image database 1</td>
</tr>
<tr>
<td>DiaRetDB2</td>
<td>Diabetic retinopathy image database 2</td>
</tr>
<tr>
<td>DRIVE</td>
<td>Digital Retinal Images for Vessel Extraction</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>EER</td>
<td>Equal error rate</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian process</td>
</tr>
<tr>
<td>HE</td>
<td>Hard exudate</td>
</tr>
<tr>
<td>HM</td>
<td>Haemorrhage</td>
</tr>
<tr>
<td>HSI</td>
<td>Hyperspectral imagery</td>
</tr>
<tr>
<td>HSV</td>
<td>Hue, saturation, value</td>
</tr>
<tr>
<td>LOF</td>
<td>Local outlier factor</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a-posteriori</td>
</tr>
<tr>
<td>MAs</td>
<td>Microaneurysms</td>
</tr>
<tr>
<td>MESSIDOR</td>
<td>Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov random fields</td>
</tr>
<tr>
<td>NFL</td>
<td>Nerve fiber layer</td>
</tr>
<tr>
<td>NPDR</td>
<td>Non-proliferative diabetic retinopathy</td>
</tr>
<tr>
<td>NN</td>
<td>Neural network</td>
</tr>
<tr>
<td>PC</td>
<td>Principal component</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PDR</td>
<td>Proliferative diabetic retinopathy</td>
</tr>
<tr>
<td>PR</td>
<td>Precision-recall</td>
</tr>
<tr>
<td>PS</td>
<td>Precision</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified linear unit</td>
</tr>
<tr>
<td>RGB</td>
<td>Red, green, blue</td>
</tr>
<tr>
<td>RGS</td>
<td>Region growing segmentation</td>
</tr>
<tr>
<td>RLs</td>
<td>Red lesions</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>RPE</td>
<td>Retinal pigment epithelium</td>
</tr>
<tr>
<td>SEs</td>
<td>Soft exudates</td>
</tr>
</tbody>
</table>
SFSOM  Semi-fuzzy self-organizing map
SGD    Stochastic gradient descent
SN     Sensitivity
SP     Specificity
STARE  STructured Analysis of the Retina
SURF   Speeded-up robust features
SVM    Support vector machine
SYM    Symmetry point
1 INTRODUCTION

1.1 Background

Diabetes is a group of diseases caused by the inability of human organism to digest the sugar in the blood. This is caused either by the absence of the insulin or the improper reaction of the organism on insulin produced. According to [1], 422 million people have this illness as of 2014. Moreover, it took more than 1.5 million lives only in 2012, which makes it one of the leading causes of death worldwide [1].

Without appropriate treatment, diabetes may also lead to a number of not lethal, but nevertheless serious complications, one of which is the diabetic retinopathy. In the majority of the cases, such complication results from the prolonged diabetes. It damages the blood vessels in the retina, which in turn make the fluid leak out and form lesions on the retina surface. Depending on the type and amount of lesions, as well as their location in the retina, retinopathy may lead to vision impairment of different severity including total blindness. According to [2], the diabetic retinopathy is the leading cause of blindness among the people between 20 and 64 years old worldwide, and 25% of persons with diabetes have considerable visual impairment.

Success of the retinopathy treatment highly depends on early diagnosis. However, in its early stages, retinopathy does not exhibit any noticeable effects. In particular, the first stage, non-proliferative diabetic retinopathy (NPDR) occurs without symptoms and may be detected only in the images of the retina produced by the technology of fundus photography [3]. In such images, lesions may be distinguished based on their colour, shape, location and other characteristics. Due to the high complexity of the task, this diagnosis is usually carried out by a professional ophthalmologist.

Unfortunately, high spread of the disease and the absence of qualified specialists in some areas often make such diagnostics impossible. Therefore, there is a need for an automated way of detecting abnormalities in the retina from the fundus images. With modern technologies, this task may be carried out by a computer with an appropriate algorithm implemented. A considerable amount of work has been done in the fields of digital processing and machine learning, which constitutes a solid foundation for the investigation of such algorithms. In particular, the question of extracting information about the diabetic retinopathy has been studied by many researchers in the recent years [3]. Undoubtedly, in order to proceed in the investigation, one should study what was already tried out and
achieved in the field.

In the majority of the works about the retinopathy diagnostics, the performance of a proposed algorithm was evaluated using one of the publicly available databases of fundus images. Most of them contain information in the form of red-green-blue (RGB) images, and thus, the developed algorithms are specialized on and restricted by this type of information. In RGB images, each pixel is represented as a vector with 3 elements, which contain information about red, green and blue colour components. In contrast, pixels in hyperspectral images are vectors of arbitrary length with each element containing information about the incident light of a certain wavelength. Studies in the related areas provide evidence that hyperspectral imagery (HSI) may contain richer information about the object of interest.

At the same time, along with its benefits HSI undoubtedly also brings difficulties, which should be taken into account. Namely, even though RGB images are considered to be problematic objects in machine learning due to their high dimensionality, HSI are yet more high-dimensional. Another problem is to find the effective way of combining information from different bands of HSI images, without which their usage appears to be pointless. Thus, one can consider HSI processing to be a special area to study on its own.

1.2 Objectives and delimitations

Taking into account the aspects discussed above and regarding the effective method of lesion detection as the target, the objective of this work is to give an overview of the works in the fields of retinopathy detection and HSI image processing and propose a new method for diabetic retinopathy lesion detection from HSI images. It describes the existing methods with their performance evaluation, makes a comparison, suggests a new technique for lesion detection and investigates into the benefits of HSI for retinal imaging.

Although the investigation sheds light on the perspectives of spectral imaging for DR diagnosis, it is worth noting that it is restricted both in terms of the algorithm and the data. That is, a thorough comparison between the colour imagery and HSI for the purpose is not considered in this work. All the conclusions are based on the obtained results, but it should be taken into account that the experiments were carried out on one particular database and with one particular method.
1.3 Structure of the thesis

The remaining part of the work is organized as follows. Section 2 gives wider perspective on the problem of the retinopathy diagnosis. It contains information about the inference, which may be performed based on the retinal images, different types of lesions as well as an overview of existing databases with fundus images. Section 3 provides a survey of the existing methods for the lesion detection and HSI processing. Section 4 explains the reasons for choosing convolutional neural network (CNN) as the algorithm for the lesion detection, gives a general description of neural networks and related methods for efficient learning. In Section 5, utilized CNN architecture and other details about the experimental setup are given, and Section 6 provides the results. In Section 7, the obtained results and limitations of the study are discussed. Section 8 summarizes and concludes the thesis.
2 DIAGNOSIS OF DIABETIC RETINOPATHY

2.1 Eye and retinopathy

In Figure 1, scheme of the human eye is shown. The outermost part of the eye is the cornea which is transparent, and its main purpose is to refract the incident light. The focus of the cornea is fixed, and the lens is responsible for the focus adjustment. In order to transmit light the lens is also mostly transparent, but it absorbs light on the boundaries of the visible spectrum, and the absorption range changes as person becomes older [4]. The cornea is continuous with the sclera, a white and opaque layer which protects the eye. Consisting of elastic fiber and collagen, it allows the eyeball to maintain its form. Certainly the organs of the eye need a source of oxygen supply, which is the function of choroid and central retinal artery. The choroid provides blood for the outer parts of the eye and the retinal artery brings it to the inner retinal layers. The space between the lens and retina is filled with transparent vitreous body, which occupies about 2/3 of the eyeball volume [5].

![Figure 1. Scheme of the human eye and its main constituents [5].](image)

The retina consists of several layers each of which serves its own function. The nerve fiber layer (NFL), covered with the limiting membrane, transmits the signal from photoreceptor cells to the central nervous system and carries neural impulses to the optic disc. The neural layer consists of photoreceptor cells, bipolar cells, horizontal cells and ganglion cells. The main function of this ensemble is to convert the energy of incoming photons into neural signals, perform their preprocessing and pass the signals to the brain [6]. The retinal pigment epithelium (RPE) is the innermost layer of the retina, it absorbs a significant
amount of incoming light, which was not absorbed by the photoreceptors, thus protecting the eye from photo-oxidation [5].

Four examples of fundus image are shown in Figure 2. On the top left image one can see a healthy retina with three main components:

- macula, a darker area which is the only part of the retina where there are more cones than rods responsible for the acuity of vision;
- blood vessels are those of the central retinal artery mentioned earlier and the central retinal vein;
- optic disc is a point where the ganglion cell axons converge and form the optic nerve, which goes to the brain; as seen in the image, blood vessels also come together near the optic disc.

Figure 2. Fundus images with healthy retina and different stages of diabetic retinopathy. MAs - microaneurysms, EXs - exudates, HMs - haemorrhages [7].
In the course of DR development, two stages may be distinguished: non-proliferative and proliferative. In the NPDR stage, fluid seeps out of the damaged blood vessels and form lesions on the retina surface. Usually the first visible sign of the DR is small red microaneurysms (MAs), shown in Figure 2b. Their size and colour, similar to that of the blood vessels, causes difficulties in their accurate detection, and many researchers utilize some method to remove vessels before detecting the MAs. After a certain amount of time, MAs break and their content leak out, forming haemorrhages (HMs) that are also red, but bigger than MAs (Figure 2d). At the same time bright yellow hard exudates (HEs) become observable (Figures 2c and 2d), they are caused by the leakage of the lipids from the blood vessels. HEs, if they appear near the macula, may lead to the complete blindness [7]. Sometimes in the context of lesion detection, MAs and HMs, are grouped together and referred to as red lesions (RLs) [8]. Analogous to the HEs, smaller lesions of the same colour are referred to as soft exudates (SEs) [9]. Depending on the amount and location of the MAs, HMs and HEs, specialists distinguish mild, moderate and severe stages of NPDR. Without proper treatment, NPDR may further develop in advanced proliferative stage. In this case, the disease causes formation of new blood vessels, especially near the optic disc. These new vessels are highly vulnerable and cause further leakage and new lesions [7]. Most of the systems developed until now concentrate on the detection of lesions and, thus, on the NPDR stage diagnostics.

2.2 Fundus image processing for retinopathy diagnosis

Having a retinal image as input, a diagnostics algorithm may follow different strategies of evaluation. Arguably one of the most straightforward approaches would be to classify an image as a whole to one of the several predetermined classes of severity. As a basis for such classification, medical taxonomy of NPDR (with its stages) and PDR may be used. A more advanced option would be to explicitly locate all the lesions in the photo, thus providing extra information about the amount and location of the abnormalities. Therefore, such an approach would mean binary classification of each pixel in the image. The difficulty here is that, as mentioned in Section 2.1, different types of lesions often have different characteristics of features, such as size and colour. Furthermore, information about the type of a lesion is useful for determining the stage of the DR. Hence, the next step is to distinguish each lesion as either MA, HM or HE [9]. Currently the trained eye of a specialist ophthalmologist is considered as the ground truth for the DR diagnostic algorithms. However, often in real-life cases, a doctor cannot define each part of the image as healthy or belonging to one of the lesion classes. Instead, the most accurate ground-truth would rather contain probabilities of each pixel to be part of the abnormal
area, and the desirable algorithm would be able to produce such information as its output. Obviously, each of these approaches includes and extends all the previous ones in terms of information useful for the accurate diagnosis.

By their nature, MAs and HMs are of the same colour as the blood vessels, which makes their detection more challenging. The experiments also show that macula, being darker region on the light background, is also often confused with lesions by classification algorithms. Moreover, yellow optic disc may be mistaken for a region with HEs. Since these objects have their own features, which make them different from the lesions, it is a common practice to remove them from the image before performing lesion detection and classification. Thus, in the task of DR diagnostics there is a subtask of blood vessels, macula and optic disc detection and removal [7]. The most widely-used approach is to use digital image processing operations, such as morphological opening and closing, edge detection and bank of filters [8], although solutions based on machine learning algorithms also exist [10]. Concrete methods for solving this task will be discussed in detail in Section 3.1.

2.3 Retinal image databases

STARE (STructured Analysis of the Retina) [11] is one of the oldest databases and has become a de facto standard in the field. This project is concentrated not directly on the lesion detection, but on the retina segmentation. The full set of raw RGB images amounts to 400, however, only 40 of them are labeled with ground-truth for vessel segmentation and 80 for the optic disc. The baseline method and results are presented in [11].

DRIVE (Digital Retinal Images for Vessel Extraction) [12] is another project focused on the retina vessels segmentation. It provides the database containing a set of 40 images with hand-labeled vasculature, divided into two halves for training and testing. For the testing part there are two segmentations available: the gold standard made by a professional ophthalmologist and alternative one made by an instructed and trained human. These RGB images are of size 768 by 584 pixels and stored in JPEG format. They were randomly chosen from a larger image set of the DR screening program, in which fundus photos of 400 patients aged 25-90 were taken.

DIARETDB1 (Diabetic retinopathy image database 1) presented in [13] was designed to facilitate the research in DR lesion detection. It contains 89 RGB images of the retina, 84 of which have retinopathy abnormalities. All the images with lesions were coarsely
annotated independently by four medical experts. These annotations are stored in XML files, which describe an infected region, representative point and degree of certainty from 1 to 3. What is more, the annotations contain information about the lesion type and severity of the disease for each image as a whole. All the photos were taken with a ZEISS FF 450 plus fundus camera with 50 degree field-of-view and, thus, share the same characteristics. As with the DRIVE database, train and test subsets are predetermined and baseline methods and results are provided in the aforementioned paper.

BristolDB (Bristol retinal image dataset) is a non-public collection of RGB retinal images created by the Bristol Eye Hospital [14]. It contains 147 pixel-wise annotated images acquired with 45 degree field-of-view Canon CR6-45 Non-Mydriatic fundus camera with a modified digital back-unit. Unfortunately, the ground truth is available only for exudates, however, its advantage is that unlike the ground truth for the other databases, it is assumed to be pixel-wise accurate.

To the knowledge of the author, MESSIDOR (Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology) [15] database is currently one of the largest publicly available set of fundus images. It contains 1200 colour images collected from 3 ophthalmological departments. For each image two types of diagnostics are provided: retinopathy grade 0-3 and risk of macular edema 0-2. Grading of both characteristics have formal definitions in terms of the distance of lesions from the macula and their type. For all the photos, Topcon TRC NW6 non-mydriatic retinal camera with a 45 degree field of view was used.

E-ophtha [16] is one of the latest appeared databases in the field. It combines two sets of images: one with exudate type of lesion marked and another one with MAs. The former set contains 82 images, of which 47 are with lesions and the rest are healthy, for the latter one this ratio is 381/148. All the pixel-wise annotations made by ophthalmology experts are binary, indicating whether a pixel corresponds to infected area.

All the mentioned databases contain colour RGB images. Due to the HSI acquisition process being more sophisticated, it is not so widely used in ophthalmology, consequently the material for a database with such information is limited. In [17], Styles et al. describe the acquisition of HSI fundus images in detail and use the described technique to obtain images of 35 subjects, half of which were discarded on the post-processing stage. This method was further improved in [4] and used in the DiaRetDB2 dataset creation. DiaRetDB2 contains images of persons aged from 21 to 81, 55 of which were diagnosed with different stages of DR and the other 17 are healthy. Generally 30 narrow-band illu-
minations were used, however in some cases several bands at the shorter wavelength were
discarded because of the absence of significant information. Three ophthalmologists in-
dependently performed annotation, which includes type of lesion and degree of certainty
on the scale uncertain - somewhat certain - certain. Besides, the overall assessment of the
DR severity was also given for each image.
3 LESION DETECTION AND HYPERSPECTRAL IMAGE PROCESSING

This work concentrates on the DR lesion detection from hyperspectral retinal images. This topic has been investigated directly by very few researchers, most probably due to the absence of the publicly available datasets, which are necessary for evaluating algorithm performance. However, this topic may be considered as intersection between the areas of HSI image processing and DR diagnostics from retinal images. In these domains, a significant amount of work has been done. An overview and analysis of this work may give an idea about the way to combine findings in the two fields in order to introduce an effective method for the DR detection. The following characteristics are used for measuring algorithm performance: sensitivity (SN) and specificity (SP) defined as

$$SN = \frac{TP}{TP + FN}, \quad SP = \frac{TN}{TN + FP},$$

where $TP$ are true negative samples (background classified as background), $FP$ false positives (background classified as lesion), $TP$ true positives (lesion classified as lesion), and $FN$ false negatives (lesion classified as background), and equal error rate (EER) defined

$$EER = 1 - SYM,$$

where $SYM$ - symmetry point on the receiver operating characteristic (ROC) curve, where $SP = SN = SYM$. The ROC curve is a line, which is obtained by plotting sensitivity as a function of false-positive rate (FPR), defined as

$$FPR = 1 - SP,$$

for different thresholds. As a corresponding numerical measure, area under the ROC curve (AUC) is often utilized. Another metric, which appears to be interesting in the context of lesion detection is precision:

$$PS = \frac{FP}{TP + FP},$$

which intuitively characterizes probability of the detected positive sample to be true positive. Analogously to the ROC curve, precision-recall (PR) curve describes the dependency of PS on SN by varying the decision threshold.
3.1 Lesion detection and severity grading

Since each type of lesion has its own unique colour and shape properties, morphological operations on images can be used to segment them. Different combinations of operations have been proposed. In one of the earliest works [18], Spencer et al. propose a strategy based on the combination of preprocessing, bilinear top-hat transformation, matched filtering and binarization for MAs detection. In the preprocessing stage, uneven illumination was corrected and shade-correction was performed. In the next stage, image was opened with linear structuring element in different orientations. In order to remove the blood vessels and keep the lesions, the length of the element was manually chosen to be just greater than the largest circular feature in the images. Further, the top-hat transformation was completed by subtracting the resulting image from the original. Finally, 11x11 MAs-matching Gaussian cross-section filter was applied to the image to refine the results. The performance of the algorithm has been evaluated on 4 images by comparing its results with diagnosis from 4 ophthalmologists, 82% sensitivity have been reported.

Frame et al. [19] further developed the described approach by extracting features of detected regions and performing classifying them into MAs and spurious objects. In this work, three types of classifiers were used: empirically-derived logical rule, linear discriminant analysis and artificial neural network. The logical rule showed the best results in this experiment with the sensitivity and specificity equal to 84% and 85% respectively.

In [20], these two techniques were combined and extended. First, on the preprocessing stage bright lesions are removed to avoid their confusion with MAs in the classification stage. Second, new features were introduced by adding to the feature vector colour information and pixel responses to Gaussian filters with different $\sigma$ values. Finally, kNN with classifier was used to distinguish MAs from non-MAs. Sensitivity and specificity of 100% and 87% have been reported.

Kande et al. [21] detected red lesions utilizing the modification of the approaches described. The goal of the work is to detect the red lesions. Since they contrast more with background in the green channel, but brighter in the red one, information from these channels were combined with histogram matching and the result was used for further processing. Blood vessels were removed and candidates for lesion regions were segmented in the same was as it was done in [20]. The support vector machine (SVM) was used to classify the candidate regions as either red lesion or non-red lesion. For performance evaluation, researchers randomly took 80 images from Clemson, DiaRetDB0, DiaRetDB1 databases and divided them into the training subset of 30 images and the test subset of 50 images.
Sensitivity and specificity have been reported to be 96.22% and 99.53% respectively.

Probabilistic models were also used to tackle this problem. T. Kauppi [9] presented the Gaussian mixture model (GMM) as a baseline classification algorithm for the DiaRetDB1 database. Classification was performed both on the pixel level as a binary task: "lesion type present" against "lesion type absent". Four types of lesions were considered: HA, MA, HE and SE. GMM was supplied with pixel colour information only and its parameters were optimized with the Figueiredo-Jain algorithm. The mean EER among all types of lesions has been reported to be 33%. It should be taken into account that plain colour information is clearly not enough for efficient detection, and the author explicitly states that the baseline method is not supposed to be state-of-the-art.

In [7], the authors also utilized GMM for lesion classification and managed to achieve relatively high results. As in other works, the images were first preprocessed for background elimination, noise removal, correction of illumination and contrast enhancement. After that, blood vessels, macula and optic disc were removed by means of morphological closing, Hough transform and clustering of dark pixels respectively. The resulting image is thresholded and the remaining areas are considered as candidates for DR lesions. A number of hand-crafted features is extracted for each region and fed into the GMM-based bayesian classifier. Parameters for mixture components are fitted with EM algorithm. The sensitivity and specificity have been reported to be 97.2/98.32 on the STARE database and 95.6/98.2 for MESSIDOR database. It is worth noting that researchers also tried to use SVM instead of GMM for classification, while leaving all the other components of the algorithm unchanged, but got less accurate results. The authors point out to the fact that the overall performance may be improved with more accurate algorithm for macula detection.

An alternative way for retinal image segmentation was proposed in [10]. In this work, researchers utilized MLP for vessel detection. The MLP was supplied with several first principal components (PCs) and edges detected on the first PC. The optic disc was located as an area with highest variation in adjacent pixels, and fovea was detected with matching correlation method. For finding HEs, region growing segmentation (RGS) was utilized. In the essence, RRGS algorithm is based on segmenting the original image depending on the similarity of the adjacent pixels, substituting the median intensity of each pixel in a region for the median intensity and binarizing the image by thresholding. For the RLs detection, a newly introduced 'Moat Operator' was applied and the resulting image was thresholded. Detection of HEs was reported to provide 88.5% sensitivity and 99.7% specificity. For RLs detection, sensitivity was 77.5% and specificity was 88.7%.
The study [8] proposed to use an ensemble of GMM and m-Mediods algorithms for the purpose. In the preprocessing stage, blood vessels were segmented with Gabor wavelet with multilayered thresholding, and the optic disc was detected as a circular region with high intensity values. Images were further enhanced with Gabor filter bank and thresholded for candidates extraction. For each candidate a set of 16 hand-crafted features was taken for classification. These feature vectors were passed to GMM and m-Mediods separately, and the final result was derived as a weighted sum of their outputs. M-Mediods method is based on finding \( M \) representative points, called mediods, with semi-fuzzy self-organizing map (SFSOM) in the training stage and in the test stage infer the class for each sample depending on the class of \( k \) closest mediods. The performance was evaluated on the combination of DRIVE, STARE, DiaRetDB, and MESSIDOR datasets. For RLs, sensitivity and specificity were 97.83% and 98.36% respectively, for HEs they were 97.39% and 98.02%.

3.2 HSI feature extraction and classification

L. Laaksonen [4] evaluated performance of several classification methods on the HSI retinal images of DiaRetDB2 database and compared the performance against the same methods supplied with simple RGB images. Generally, only for SVM the improvement was significant (which had low results on RGB), and if one would only consider the best result across all the methods the improvement of HSI data over RGB appears to be relatively small. This suggests that either multispectral images do not provide information useful for the purpose or that this information was not effectively extracted.

HSI is extensively used in the field of remote sensing, and in this context, a significant amount of methods have been introduced for feature extraction and image classification, which take into account special characteristics of multispectral images. In particular, one of the main tasks, when characterizing a pixel, is to effectively encode information about both its spatial neighbourhood and spectral dimension. For performance evaluation in this area, datasets with HSI images of the ground taken from the aircraft are used, and the purpose is to assign each pixel to one of predetermined classes of earth surface.

Another observation, which can be made from [4] is that high-level feature extractor SURF [22] outperformed all the other features both for the task of HSI bands alignment and classification. This provides evidence that by utilizing more effective feature extraction, algorithms it is possible to significantly improve HSI classification results. In [23], researchers propose D-SIFER (Derivative-Based Scale Invariant Image Feature Detector
With Error Resilience) algorithm, which was designed to take into account both spatial and spectral dimensions of a hyperspectral image. D-SIFER is based on filtering the image with the 10-th order Gaussian derivative. For comparison, SURF feature extractor utilizes second order Gaussian derivative. According to the paper, this difference makes D-SIFER more robust with respect to many kinds of image noise. Furthermore, the authors state that 10-th order derivative minimizes Heisenberg’s uncertainty both in scale and space, as shown in Figure 3. In the experiments, described in the paper, D-SIFER showed generally better performance than SURF. Thus, it might be useful both for the purpose of band alignment and more accurate classification of image areas. The drawbacks of the approach are its slower performance as compared to SURF and the absence of open-source implementation.

![Figure 3. Heisenberg’s uncertainty in space/time and scale for different orders of Gaussian derivative [23].](image)

Another possible way to capture information in two dimensions is to use a combination of two methods, each accounting for one characteristic. Such an approach was proposed in [24]. Namely, this paper combines Gaussian processes (GP) to account for spectral dimension with conditional random fields (CRF) for spatial dimension in one GPCRF framework. GP provides advantages of being both the kernel method (like, e.g., SVM) and at the same time a Bayesian probabilistic model. CRF can be viewed as an advanced version of Markov random fields (MRF), which does not require data to obey the Markov property. In the context of CRF, an image is viewed as a graph with pixels as its nodes and dependencies between them as edges. The joint distribution over labels, given observations, is modeled as a linear combination of association potential and interaction potential. During the training stage, coefficients of the linear combination and parameters for both potentials are fitted. The combination of two models is implemented by substituting the GP latent function in association potential as a function which maps the observation on the feature vector. In the experiments, GPCRF showed to be superior to classification
only with GP.

As discussed in Section 3.1, the most widely-used feature selection approach in DR detection is a set of hand-crafted features. Taking into account the recent trends of learning features from the data, authors of [25] used an ensemble of autoencoders to effectively capture the useful information from HSI images. It is worth noting that instead of the common approach of learning compressed representation of data, this work instead used a mapping to the overcomplete feature space, because such representation showed to be more effective for classification purpose. The ensemble of autoencoders consisted of the encoder for mapping the spectral representation of a pixel and several encoders for different scales in spatial dimension. After training the encoders, new representation for a pixel were obtained by stacking together the output of all the autoencoders in the one feature vector. In the experiments, feature vectors were further fed to the SVM for classification, and the results have been reported to outperform the state-of-the-art methods.

3.3 Summary

Summarizing sections 3.1 and 3.2, we can say that the general flow of all the DR lesion detection algorithms may look as follows:

- Preprocessing, including:
  - Illumination correction and noise removal.
  - Detection and segmentation of optic disc, blood vessels and macula.
  - Contrast enhancement.
  - Background elimination.
- Candidate lesion detection.
- Feature extraction.
- Classification.

By concretisation of each step with the precise target (e.g. classification between classes of lesions or binary classification lesion against non-lesion) and method (e.g., SVM for classification and top-hat transformation for vessel segmentation), it is possible to get each of the described algorithms from this scheme. Generally, it is difficult to draw strict conclusions about the most promising methods because they were often evaluated on different datasets, most of which are relatively small.
Variability of the algorithms in the field of HSI processing and feature extraction is significantly higher and can hardly be put under one common scheme. However, they all designed to effectively capture the information from spatial and spectral domains. Successful application of HSI in remote sensing suggests that it might also be appropriate in other areas, such as fundus imaging, although this latter application is not yet well studied.
4 NEURAL NETWORKS

4.1 Challenges and motivations

The main goal of this work is the investigation into efficient algorithm for lesion detection from hyperspectral images by utilizing the DiaRetDB2 database. During the course of the work, this investigation faced a number of challenges due to the nature of the data as well as specific characteristics of the task and the dataset. These factors played a major role in the decision to use neural networks (NN) for the purpose.

One of the most important difficulties is high dimensionality of the data. This property is undoubtedly due to the nature of hyperspectral images and every research dealing with HSI has to consider this problem in order to benefit from large amount of information. Big size of a dataset is usually considered as a compensation for high dimensionality. However, both acquisition and labeling of retinal images are time-consuming processes. Aggravated by privacy issues, this leads to labeled datasets with retinal images containing small number of samples. DiaRedDB2, consisting of 55 labeled images is not an exception from this rule. Thus, an algorithm for the lesion detection has to work in difficult conditions of scarce data and high dimensionality.

Another problem with lesion detection is highly skewed distribution of class labels. If detection is considered as a segmentation task, where the algorithm should classify each pixel as either belonging to abnormal area or healthy area, it is clear that usually there are much more healthy pixels than infected ones. This makes lesion detection more challenging than segmentation of images in remote sensing or blood vessels detection, where the difference of proportions of different classes is usually much smaller. Furthermore, lesion markings are not only more difficult to produce, they are also more subjective and ambiguous. For this reason, DiaRetDB2 contains ground truth from three experts and markings of two types: coarse regions and representative areas. Although this diversity in markings provides more information from experts, it also raises questions about proper training and evaluation procedures. It should be noted that none of the available markings are considered to be pixel-perfect [4], and according to the experiment described in [4], this can highly affect performance of the detection algorithm.

Considering all this challenges, one can conclude that an algorithm utilized for the purpose should be powerful enough to extract low-level features from high-dimensional data, resilient to inaccurate markings and flexible to provide means for tuning in order to com-
pensate scarcity of the data. Taking into account recent achievements in the field of neural networks and deep learning and a great number of different techniques developed for their efficient training, it was decided to use a convolutional neural network (CNN) for detection of diabetic retinopathy lesions. The rest of this section is describing the mechanism of neural networks and major aspects related to it.

4.2 General algorithm

4.2.1 Fully-connected networks

The concept of fully-connected neural networks is based on the representation of complex nonlinear function as a directed acyclic graph. The network consists of nodes-neurons, which are connected between each other with edges. Neurons are grouped into layers, on which operations are performed. Neurons of one layer are connected to neurons of the next and preceding layers, but they are not connected to the other neurons of the same layer. The training procedure consists of two stages: forward pass and backpropagation. During the forward pass, training data is fed to the first layer and then passed through all the layers in the network being transformed with multiplications by coefficients corresponding to edges and nonlinear activation functions corresponding to the network layers. Thus, a neuron on layer $n$ will contain $\sum_{i=1}^{k} a_{i}^{n-1} w_i + b_n$, where $k$ is amount of neurons on layer $n - 1$, $a_{i}^{n-1}$ activation of $i$th neuron of layer $n - 1$, $w_i$ - weight of edge between the two neurons and $b_n$ is the intercept term. After the forward pass, the last layer contains the output, which is compared with the ground truth by means of a loss function. The goal of the training stage is to minimize the loss function by taking gradient of it with respect to each of the weights in the network and utilizing iterative gradient descent.

4.2.2 Convolutional networks

Convolutional neural networks were introduced by Yann LeCun [26]. The main goal of their design is to efficiently extract information from images and decrease the amount of parameters in the network. The mechanism is based on convolution operation, when a matrix of size $h \times h \times n$, which is called the kernel, is used as a sliding window to convolve an input tensor. Mathematically for two-dimensional discrete case this operation
is defined as follows:

\[ S(i, j) = (I \ast K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n), \]

where \( I \) is an input and \( K \) is a kernel. Coefficients in the kernel matrices are parameters of the network, and differentiability of the convolution enables us to update them with backpropagation and gradient descent.

Apart from the aforementioned advantages of CNNs, it has also an important property of local connectivity. Unlike neurons of fully-connected networks, neurons of a CNN are not connected to all the neurons of neighboring layers. Instead, activation of one neuron depends only on the activations of neurons under the sliding window at one position. However, since result of the convolution may be convolved again, layers higher in the hierarchy have a larger field of view than those which are lower. It means that different layers may extract different types of features. In application to image processing, the first layers usually extract low-level image features such as corners and edges, while the higher layers are responsible for more complex concepts such as different classes of objects. Pictorially, convolution operation is shown in Figure 4.

![Figure 4. Illustrative representation of two-dimensional convolution [27].](image)

### 4.3 Activation functions

After receiving a linear combination of neurons’ activations of the previous layer, each neuron computes a non-linear activation function of the received value. This is necessary in order to enlarge network’s representational power. Without the activation function, all the transformations in the network would be affine, and thus, the algorithm would be able to learn only affine functions.
Historically, the first proposed activation function for neural networks was sigmoid, which has the following form:

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$  

This function takes a real value and squashes it between 0 and 1. It has a plausible property of being interpretable as a probability, however, due to discovered drawbacks, it is no longer used in modern architectures. The main issue with sigmoid is that it is prone to vanishing gradients problem. The gradient of the sigmoid is close to zero both near 0 and 1. If a neuron has a small local gradient, it will be multiplied by the incoming global gradient and eventually block the gradient flow for all the neurons down the stream.

In order to resolve this problem, Krizhevsky et al. [28] proposed rectified linear unit (ReLU) function. It is defined as

$$f(x) = \max(0, x).$$  

Disregarding its simplicity, it was shown to significantly improve convergence time and mitigate problem of vanishing gradients described above. What is more, it requires much less computations than sigmoid, since it only involves simple thresholding operation. Currently, ReLU is the default recommendation for any neural network.

Nevertheless, some researches emphasize that ReLU still prone to gradient problem since it zeros all the negative inputs. As a result, if the learning rate is set to high many neurons in the network will end up never activating. To address this issue several new activation functions were proposed by researchers. One of them is leaky ReLU [29]:

$$f(x) = 1(x < 0)(\alpha x) + 1(x \geq 0)(x),$$  

where $\alpha$ is some small constant and $1(f)$ is logistic function, which equals to 1 if $f$ is true and 0 otherwise. From the formula one can see that for the positive values Leaky ReLU gives the same result as ReLU, but it also allows small activations for negative inputs. Although in theory, this property appears to be plausible, there is no strong evidence of its benefits in practice. Another variation of ReLU called Parametric ReLU treats $\alpha$ in the above formula as a learnable parameter [30]. This provides the opportunity to adjust activation functions, but also doubles the amount of parameters in the network, which leads to longer training and danger of overfitting.
ReLU activation functions are good for hidden layers because they allow efficient gradient flow in the network. However, for the output layer interpretability is a desirable property. For example, for the classification task it would be handy to have probabilities of an input sample to each of the predefined classes as output from NN. Sigmoid function is one option for this purpose, however it is limited to binary classification. The softmax function is the usual choice for multi-class classification. It is defined as

\[
\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}.
\]

If this function is applied to all neurons of one layer, their activations sum up to 1, and the result may be interpreted as probabilities associated with multinoulli distribution [31].

### 4.4 Loss functions

The result of the forward pass is the output of the last layer. In order to train a neural network with gradient descent, it is necessary to define the loss function, which should measure how far are the obtained results from the desired ones. In the supervised learning framework, the loss function usually compares the output of a network with the ground truth labels. Two popular choices of loss function for multi-class classification are hinge loss and cross-entropy.

Multi-class hinge loss is defined as

\[
L_i = \max_{j \neq p_i}(0, t_j - t_{p_i} + \delta),
\]

where \(\delta\) is constant value for desired margin between the activation corresponding to the true class and all the others, \(t_c\) - activation corresponding to class \(c\), \(j\) is the correct class and \(p_i\) is the class corresponding to the \(i\)th neuron in the output layer. This loss function, thus, tries to force the network to produce answers such that output activation corresponding to the true class would be larger than all the other activations by some constant value \(\delta\), which is usually taken to be equal to 1. It means that hinge loss does not interpret the input activations and their actual values do not matter as long as the margin is preserved.

In contrast to hinge loss, categorical cross-entropy treats input activations as probabilities and tries to minimize Kullback-Leibler divergence between the true distribution of
labels and the probabilities predicted by the network. The mathematical formulation is as follows:

\[ L_i = -\sum_j t_{i,j} \log(p_{i,j}), \]

where we sum over all activations. Since this loss function relies on the probabilistic interpretation of the inputs, it is usually used after applying the softmax activation.

### 4.5 Methods for parameters update

After the forward pass with the loss function, one can obtain the gradient, which is then used for updating the weights of the network. The most straightforward strategy for these updates is the stochastic gradient descent (SGD), which makes steps in the direction of the negative gradient which is calculated on a mini-batch of training data. Thus, after a forward pass each weight in the network is updated according to the following formula:

\[ w^{i+1} = w^i - \alpha \frac{\partial L}{\partial w}, \]

where \( \alpha \) is the hyperparameter defining the learning rate, \( w \) is the target weight and \( L \) is the loss function. It was shown by the researchers that SGD is prone to long convergence since it is highly sensitive to noise in mini-batches [31].

Several alternative strategies were proposed to address this problem. One of the important ideas is to interpret minimization of the cost function as physical rolling of a particle (current point in parameter space) from a hill. The newly computed gradient is interpreted as momentum which affects velocity, and velocity in turn affects the displacement of the particle [32]. The formulas for parameters update are as follows:

\[ v^{i+1} = \mu v^i - \alpha \frac{\partial L}{\partial w} \]

\[ w^{i+1} = w^i + v^{i+1}, \]

where \( \mu \) is constant corresponding to friction in the system. The momentum update usually converges faster than the SGD, however, it still requires tuning of the learning rate as a hyperparameter.
Several algorithms, which automatically tune the learning rate were also proposed. One of the most recent ones is called Adam [33], its name is derived from "adaptive moments". This algorithm substitutes calculated gradient with its moving average and makes each update inverse proportional to the moving average of squared gradients. This allows for parameters which usually receive small gradients have larger coefficients than the others, and, thus, balance the overall update for all the parameters. The update is performed in three steps:

\[
G_{w}^{i+1} = \beta G_{w}^{i} + (1 - \beta) \frac{\partial L}{\partial w},
\]

\[
H_{w}^{i+1} = \delta G_{w}^{i} + (1 - \delta)(\frac{\partial L}{\partial w})^2,
\]

\[
w_{i+1} = w_{i} - \frac{\alpha G_{w}^{i+1}}{\sqrt{H_{w}^{i+1}} + \epsilon},
\]

where \( \beta \) and \( \delta \) are hyperparameters, for which the recommended values are 0.9 and 0.999 respectively [33].

4.6 Regularization

Overfitting is a concern for any machine learning algorithm. Neural networks are especially prone to this problem due to the large model complexity and large number of parameters even in architectures of modest depth. The newly proposed methods for avoiding overfitting played an important role in the current wave of interest to the neural networks.

One of the most widely-used regularization strategy is to add extra term to the loss function, which would impose a constraint to the values of network parameters. The most common approach is called L2-regularization. The idea is to add the sum of squared weights of the network to the loss function. Thus, the overall loss becomes

\[
L = \frac{1}{N} \sum_{i} L_{i} + \frac{1}{2} \lambda \sum_{j} \sum_{l} w_{j,l}^2,
\]

where \( \lambda \) is regularization coefficient, and in the second term we sum over all the weights in the network. Theoretically inference after learning parameters with L2-regularization may be interpreted as maximum a-posteriori (MAP) estimation with Gaussian prior over parameters. Intuitively L2-regularization forces the network to have small activations for all the neurons rather than concentrate large values in a few neurons.
Another widely-used type of regularization is L1-regularization. It is similar to the previous method, the only difference is that, instead of the sum of squared weights, we add the sum of absolute values of weights to the loss function:

$$L = \frac{1}{N} \sum_i L_i + \frac{1}{2} \lambda \sum_j \sum_l |w_{j,l}|.$$  

Although the formula looks similar, the interpretation and effect on the weights are different. From the probabilistic point of view, using L1-regularization is the same as setting a Laplace prior over the network parameters. This regularization makes the network have many parameters equal to zero, which means that many input features will be discarded and only the most relevant ones will be used.

For any machine learning algorithm, it is important to have relatively large amount of representative data, from which it would be possible to extract information. Scarce data often leads to issues with overfitting. Thus, increasing the amount of data may be viewed as a regularization method. When it is not possible to get more data, it is often still possible to generate new valid samples from the training set. A set of techniques for this purpose is called data augmentation. One of the simplest and most widely used approaches for image classification are flipping, rotating, scaling and adding small translations. If the goal is classification then the training samples are usually invariant to such transformation, meaning that the transformed images are still valid members of their original classes. It is also not uncommon to utilize domain-specific data augmentation or adding small random noise to samples to make the network robust to noisy real-world data [34].

Initialization of weights and architectures may have significant effect on the training process, and it is often hard to decide on the best training setup. Many researches reported that using ensemble of several networks for prediction is beneficial in comparison to using only one. However, both training and prediction become much more computationally expensive. Dropout [35] was proposed as simple and effective method to prevent overfitting in neural networks. Its principle is such that during the training, each neuron is "turned off" with some predefined probability as shown in Figure 5. Dropout has the interpretation of training many different networks and averaging result over their predictions. Unlike model ensembles, this benefit comes for almost no cost, since dropping involves only a few simple operations. Usually dropout is applied to fully connected layers, and it is widely-used in many modern successful architectures.
4.7 Normalization

In order to speed up and facilitate the learning process, input data is usually normalized with one of several commonly-used transformations. In most cases, standardization is utilized to center the data and set its standard deviation to 1:

\[ x_{\text{centered}} = x - \frac{\sum_i x_i}{N} \]

\[ x_{\text{normalized}} = \frac{x_{\text{centered}}}{\text{Var}[x]} . \]

There are two options with this normalization: either standardization is performed along the feature dimension or along the sample dimension. The former one provides the opportunity for each feature to have equal effect on the final result, while the latter makes samples more comparable.

Another common strategy is to scale all the input values such that they will all in one interval \([x_{\text{min}}, x_{\text{max}}]\). To achieve this, from each of the values we subtract the minimum and then divide the result by the value span:

\[ x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} . \]

This method is somewhat close to standardization in data dimension, but it is handy when exact minimum and maximum values are needed.

The data samples in the training set are usually correlated. However, it has been shown that machine learning algorithms work better if there is no correlation between the input samples. Principal component analysis (PCA) may be utilized in order to decorrelate the data. PCA finds an orthogonal basis with components sorted by the amount of ex-
plained variation. By projecting the data onto this basis, one can achieve decorrelation. Furthermore, if the decorrelated data is divided by the variance in the corresponding component, the resulting dataset becomes standard Gaussian. This process is often referred to as whitening.

The normalization techniques described above are effective for normalizing the input data. However, during the training, the activations in each layer of the network often become correlated. Since each hidden layer may be viewed as input for all the successive layers, for them the correlation is an undesirable property for the same reason as it is for the data. Batch normalization (BN) algorithm proposed in [36] was shown to effectively remove the covariance shift during the network training. Batch normalization can be viewed as another network layer. It takes activations of the previous layer as input and transforms them as:

$$BN_{\gamma,\beta}(x) = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta,$$

where $\epsilon$ is some small constant to prevent division by 0, $\gamma$ and $\beta$ are learnable parameters scaling and shifting, and $\mu$ and $\sigma$ are the mean and variance computed on the input batch. It is easy to see that batch normalization essentially performs standardization on the activation for each data batch. What is more, differentiability of this operation makes possible backpropagation through these layers and the learnable parameters regularize the impact of normalization. For example, the network may learn the both parameters to be equal to 0 and completely discard the layer if this would be beneficial for the training.
5 EXPERIMENTAL SETUP

5.1 Network architecture

The general neural network model utilized for lesion detection in this work, DiaRetNNet, is shown in Figure 6. It takes a patch of size XxYxZ around the target pixel as input and returns probabilities of this pixel to belong to each of three classes: background, haemorrhage and exudate. The task of detecting red small dots is undoubtedly of high importance and interest, but it is not considered in this work. Unfortunately, currently available ground truth labels for red small dots in DiaRetDB2 contain only markings of single pixel for each region, which makes proper training and evaluation impossible.

![DiaRetNNet convolutional neural network architecture.](image)

The decision to utilize patches around pixels as samples instead of whole images or individual spectra was made for the following reasons. First of all, as discussed in Section 3.2, both the spectral and spatial information are crucial in HSI image segmentation. This clearly applies to retinal HSI as well. One good example of it is haemorrhage detection:
due to their nature, haemorrhages have the same spectra as blood vessels. Consequently, haemorrhages would be indistinguishable from the vessels for the algorithm, which would utilize only spectral information. This is why it is important to incorporate into the input information the neighborhood of a target pixel. Another point is that dataset of size 55 images is extremely scarce. Thus, a successful algorithm should be able to make use of available information as effectively as possible. If the CNN would operate on the whole image to map it into mask with three classes, significant amount of information would be lost in the pooling layers and the effect of small areas with lesions would be negligible in comparison to the large proportion of the background. Instead, when each patch is mapped directly to the class probabilities, the inference algorithm can leverage all the relevant spatial and spectral information in order to give the exact answer to the question of interest as accurately as possible.

Conceptually the model may be viewed as consisting of three components. The first block of three layers with 1x1 convolutions is meant for dimensionality reduction. The input images have high spectral dimensionality. However, presumably only the fraction of this information is relevant for lesion detection and the rest would only increase the number of parameters in the network, making it prone to overfitting and slowing down the learning process. Several of the algorithms described in Section 3.1 [21] utilized only one or two layers of retinal RGB images, which also can be viewed as a manual way of dimensionality reduction, and PCA is widely used for dimensionality reduction both in retinal imaging [10] and HSI processing [25]. The dimensionality reduction block enables us to learn the most suitable transformation instead of resorting to predefined methods without introducing many parameters to the model. Three layers with 16, 8 and 3 filters total to 659 parameters in this block.

The purpose of the second block is extraction of low-level features. It consists of two sub-blocks with two 3x3 convolutional layers followed by 2x2 maxpool layer with stride 2. Convolutional layers in the first sub-block have 64 filters and 128 in the second one. This part of the network is identical to the first two blocks of the VGG19 [37] architecture, and weights of this model pretrained on ImageNet database [38] were utilized and fine-tuned for DiaRetDB2. This motivates the decision of reducing the image dimensionality to three with the first block: VGG19 was designed for and trained on the RGB images with three channels. ImageNet contains images with objects of high conceptual level and great variability. However, objects in DiaRetDB2 may be characterized with low-level features and have much less intraclass variability. Thus, there is no need in very deep architectures such as VGG19, which would lead to overfitting due to scarce data and large amount of parameters. It is well-known that first convolutional layers in deep neural
networks effectively detect low-level features, and for this reason the first two VGG19 blocks were used in DiaRetNNet. These layers provide in total 260160 parameters to the network.

The last block can be viewed as feature processor. It contains two convolutional layers with 64 5x5 and 32 3x3 filters respectively. Both of these layers use zero-padding such that the output tensor size would be the same as input and have 2x2 maxpool layers with stride 2 after them. Output of the last maxpool layer is flattened and passed to the fully-connected layer with 500 neurons, which, in turn, connected to output layer with three units corresponding to probabilities for three classes. The third block contains 289331 parameters.

DiaretNNet makes use of batch normalization by placing a layer of this type after each convolutional layer except those in the second block, due to the fact that no batch normalization was used in the original VGG19 architecture. One batch normalization layer was also placed after the input layer in order to perform normalization among samples, which are possibly taken from different images, from the beginning of forward pass. Weights of all the layers are regularized with L2-regularization with coefficient $\lambda$ equal to 0.01 in order to prevent overfitting as much as it is possible with such complex data. For the same purpose dropout layer was placed between the fully connected and output layers. For all the experiments, the probability of dropping each neuron $p$ was taken to be 0.5. The learnable parameters of batch normalization layers add 612 more parameters to the architecture, which contains in total 550762 weights.

Rectified linear unit was used as activation function everywhere in the network except the output layer, where softmax function was utilized to get proper probability distribution in the output. For those layers, which are followed by batch normalization, activation is placed after the BN layer. During the training stage output of the last layer is passed to cross-entropy cost function, which is being optimized. The contribution of samples to the cost function value is weighted depending on the sample’s class label. Further details and reasons for this will be described later. After each batch is processed with forward and backward pass, parameters of the network are updated with Adam.

### 5.2 Sampling strategy

Arguably the most straightforward way for sampling new batches for training would be to take all the patches from a training and, after feeding them all to the network, switch
to the next image. Such strategy, however, is not applicable in the described setting for several reasons. First, images in DiaRetDB2 are of size 1024x1024 pixels. This means that one image contains 1048576 samples. In practice, this number is always smaller and varies from image to image due to masks which eliminate empty edge areas, but it is still large enough to make sequential processing of all of them infeasible. What is more, due to variations in illumination and highly individual properties of human retina, images are significantly different from each other. Therefore, a batch consisting of samples taken from single image would bias the network towards properties of that image. Processing images sequentially in this way would make the network starting to relearn each time a new image is taken, and thus, it would substantially increase convergence time or make the training fail to converge. With this in mind, this work draws samples for new batches with equal probability for all valid pixels of each class in the training set of images, making the probability of taking pixel from a particular image dependent only on the amount of valid pixels in this image.

As stated in Section 4.1, one of the most important issues with the data is highly skewed class distribution. In such setting, randomly taking samples leads to sampling only background almost all the time. In this way, the CNN does not have a chance to learn the properties of underrepresented classes and classifies all the data as belonging to the largest class. To avoid this problem, the algorithm increases probability of drawing samples of underrepresented classes. More precisely, for each batch the probability of drawing sample $x_i$ is

$$
p(x_i) = \frac{p(y_i)}{\sum_j 1(y_j = y_i)},
$$

where $p(y_i)$ is probability of taking sample of class $y_i$, which is set as hyperparameter, and in the denominator the sum is over the whole training set. In other words, concentrations of samples of each class in a batch are set by hand and evenly distributed among their members, such that the probabilities inside class $y_i$ sum up to $p(y_i)$ and the sum over the whole dataset would be equal to 1, defining a proper probability distribution.

As stated above, by considering each pixel as a sample, it turns out that even a single image contains many samples. In fact, such perspective poses a problem of large amount of data, which is infeasible to process on the computer used for the research. In terms of the training process, it means that size of the background class is too large, and the network cannot see all its samples during the training. Thus, it is desirable to perform sampling from this class as effective as possible. First of all, it is important to draw new
samples for each batch, that is, select the ones which were not seen before. For this purpose, after each sampling the probabilities for drawn samples are set to zero and the distribution is renormalized to sum up to 1.

By using the strategy described above, the sampling algorithm is guaranteed to produce samples unseen before. However, with such immense amount of data, it is also important to take the most "interesting" samples, that is the ones which are difficult to classify correctly. Returning to the example about the distinction between the haemorrhages and blood vessels, the spatial dimension provides information which makes this distinction possible, but if the CNN will not see enough samples of the blood vessels, it is unlikely that it will learn to properly utilize this information. Unfortunately, currently there is no available expert ground truth for the blood vessels in DiaRetDB2, and it is not possible to explicitly sample pixels of vessels. To address this issue, the "interesting" background was defined as those spectra which are close to the spectra of other classes in Euclidean space. The idea is to calculate the distances between the available background samples and the samples of haemorrhages and exudates as

\[ d(b, l) = \|b - l\|_2^2, \]

where \(b\) and \(l\) are sampled spectra of the background and lesions respectively, and add to the batch those background samples, for which \(d\) is larger. In practice, there are several considerations about this method. One issue is that it is infeasible to calculate distances for all samples. Because of this, only a subset of all the samples is considered. The size of this subset should still be kept large enough for the method to be meaningful. Another consideration is that taking the hardest samples might lead to overfitting due to outliers in the dataset. This is why, instead of the hardest samples, semi-hardest ones are taken. One can compare this strategy to the one utilized for Facenet [39] embeddings learning. In this case, however, the goal is not to learn embeddings for the classification but perform direct classification. Therefore, the sampling algorithm constructs not triplets for the batches but batches directly.

Another type of background, which may be considered as interesting, is the outliers. If we would again perform a comparison with the triplet learning in [39], the outliers are hard positive samples of the background class. The intuition behind this is that the outliers are those samples which are rare and very different from the other samples. This implies that there is a relatively small probability to draw them randomly, as well as that the transformations learned on the other samples might not be applicable to them. This is
why it is important to incorporate the outliers of the background class into each training batch. For the outlier detection, this work utilized local outlier factor (LOF) [40]. This method detects the outliers by performing local density estimation around each candidate sample and comparing densities of neighboring samples, the ones for which it is smaller are considered as outliers. The density estimation is also based on $k$ nearest neighbors of each sample. Slightly simplifying the notation from [40], density can be written as

$$\text{lrd}(p) = \frac{1}{|N(p)|} \sum_{o \in N(p)} \text{rd}(p,o),$$

where $p$ is target sample, $N(p)$ is $k$-neighborhood of $p$, $|N(p)|$ is its cardinality and $\text{rd}(p,o)$ is reachability distance between $p$ and $o$. The LOF is, then, defined as follows:

$$\text{LOF}(p) = \frac{\sum_{o \in N(p)} \text{lrd}(o)}{|N(p)|} \text{lrd}(p).$$

In this work, LOF was calculated on a subset of samples from the background class and then $M$ samples with the largest LOF were added to the batches. Finally, considering all the described components, the diagram of constructed batch is shown in Figure 7. The hyperparameters for sampling methods discussed here will be given in Section 6.

5.3 Data preprocessing and augmentation

For the aforementioned reasons, spectral retinal images in DiaRetDB2 exhibit significant variations. A successful algorithm for the lesion detection should be able to overcome this difficulty. Generally, there are two ways to address such problem: perform normalization which would remove the diversity irrelevant to the task (e.g., correction of uneven illumination) or force the learning algorithm to generalize over this diversity (e.g., use images with many different illuminations). In this work, both of these approaches were utilized.
First of all, it is important that all the values would be comparable. As a general normalization step, minmax scaling in interval [0; 1] was performed for each channel separately on the whole dataset. Although this eliminates the problem of variance in absolute values, inside the interval values may still be biased towards different values. For example, from the histograms corresponding to the same wavelength for several images in DiaRetDB2 (Figure 8), it is easy to see that they have different means. This is property is undesirable due to the fact that the overall image content is the same, and hence difference in moments is most likely caused by factor irrelevant to the task. With this in mind, it was decided to subtract from all values the mean over the corresponding band after the minmax normalization.

![Histograms](image)

**Figure 8.** Histograms for the band corresponding to the wavelength 583 nm of four images from DiaRetDB2.

One of the sources of variation between images is inter-person variation due to high uniqueness of human retina. In [9], an attempt was made to account for this property by taking anchor points on retinal images, choosing one reference image and fitting transformation from anchor points of other images to those of reference image. After that, all the images were transformed according to found transformations. In this work, due to high flexibility of CNN and large number of parameters, it was assumed that the network should be able to generalize over this type of variation with proper data augmentation method. For this purpose, several new sets of images were produced by histogram matching. Histogram matching is a general-purpose method for transferring colour properties of one (reference) image to another (target) image by forcing the target to have the same
colour distribution as the reference. In the case of hyperspectral images, one have to work with several bands, and it was decided to apply such transformation for each band separately, that is, by histogram matching of the hyperspectral image to another one, it is meant to perform matching between each of their corresponding bands.

Because of the complexity of the acquisition process, spectral images are prone to different sources of noise, for which it is difficult to create good model and, thus, it is difficult to account for it. One of the common ways to address the problem in neural network community is to add some noise to the data, which is regarded as a way of data augmentation making the network more robust to noise. Although this method was shown to be effective in some cases [34], it is desirable to add variation which is known to occur in the data instead of random noise. A technique to achieve this, Fancy PCA, was proposed in [28] in application to the ImageNet dataset classification. With this method, PCA was run over all the spectra in the training set, and during the training, spectra of each patch were altered by adding the following quantity:

\[
[p_1, ..., p_{30}] [\alpha_1 \lambda_1, ..., \alpha_{30} \lambda_{30}]^T,
\]

where \(p_i\) is \(i\)-th principal component, \(\lambda_i\) is corresponding amount of explained variance and \(\alpha_i\) is random variable drawn from Gaussian distribution with zero mean and standard deviation 0.1.

An attempt was also made to make the network more robust to variations in illumination. Illumination model proposed in [41] was fitted to the value channel of training images, which were previously converted from HSI to RGB and then to hue, saturation, value (HSV) color space. The fitted illumination fields were then used to correct illumination in all spectral bands. During the training procedure, one of a set of model parameters for one of the images was picked randomly, varied by adding noise from \(\mathcal{N}(0, 0.1)\) and produced illumination fields was applied to all the images. The algorithm was switching between the model parameters for different training images in the beginning of a new epoch and on each sampling with probability 1%.

Apart from the described approaches for the data augmentation, more classical ones were also applied. Namely, each patch was rotated by 0°, 90°, 120°, 270° with equal probabilities and flipped with left-right, up-down or both flips.
6 EXPERIMENTS AND RESULTS

All the experiments were run on computer with an Intel Xeon processor with the frequency 2.7 GHz and 8 cores. All the computations related to CNN training were performed on Nvidia Geforce GTX TITAN Black graphics processing unit with 6 gigabytes of memory through cuDNN library. Python 3.5 and Keras 2.0.2 library [42] with Tensorflow 1.0.1 [43] as its backend were used for the implementation. Besides, scikit-learn 0.19.dev0 [44], scikit-image 0.12.3 [45] and numpy 1.12.1 were utilized for miscellaneous data-processing tasks.

6.1 Performance on hyperspectral images

DiaRetDB2 dataset contains 55 labeled images of patients with different stages of diabetic retinopathy. From this set, 35 images were taken for training and the rest were left off for testing. During the creation of the dataset some bands at the edges of the utilized spectra were discarded in some images due to small amount of contained information. Since the size of the network input should be fixed, missing spectra were padded with zeros. For all the experiments the spatial size of the patch was chosen to be 35x35 due to the fact that this size is larger than the size of the largest blood vessel among the training set.

DiaRetDB2 contains two types of ground-truth: coarse markings and representative regions. The former ones usually enclose large areas containing lesions of specific type, while the latter contains regions for which an expert was confident that all its pixels belong to the lesion area. Two examples of markings are shown in Figure 9. It can be clearly seen that apart from infected areas, a coarse region also contains large amount of healthy background. Such setting gives rise to the important question about the correct way of creating ground-truth masks. Unfortunately, for the ambiguous areas lying between the representative region and coarse boundary, it is not possible to say whether those pixels should be considered as lesion or as background. For this reason, such areas were not considered both during training and evaluation.

Another important decision is the proper way of using the three ground truth markings available. During the examination of the dataset, it was found out that there is considerable level of disagreement between the three experts. In such a setting, fusion methods typically perform poorly and provide scarce masks. For comparison, markings of all the experts for two images from DiaRetDB2 are shown in Figure 10. In the images, one can
Figure 9. Ground-truth for haemorrhages on two images from DiaRetDB2. Coarse boundaries are marked with blue and representative regions are marked with red.

see that the first expert tends to mark misleadingly large coarse regions, while the third expert does not mark representative regions inside the coarse areas. When assessed visually, the markings of the second expert seem to be the most meaningful for the purpose of collecting the ground truth. For this reason, only the markings from the second expert were utilized in this work.

For the hard background, sampling subsets of size 100000 and 10000 were used for taking pixels which are similar to lesions and finding the outliers respectively. The difference is due to efficiency concerns: calculating Euclidian distances between several pixels with lesions and background samples requires less computations than calculating LOFs. For the same reason, the amount of neighbors was set to 10 and the contamination coefficient for LOF algorithm was constrained to 0.005%, that is, 50 outliers were added to each batch.

Data augmentation with histogram matching, described in Section 5.3, was performed by producing 3 extra sets of images by transferring the colour information from image \( n_{(i+k)\%55} \) to image \( n_i \), where \( i \in [1, ..., 55] \), \( k \in [1, ..., 3] \) and \( \% \) denotes the modulus operation. During the training procedure algorithm randomly switches between these four datasets. The training was performed for 50 epochs with 100 batches per epoch and 300 samples per batch.

In order to find out how different setups affect the algorithm performance, several runs were performed: with the full described setup except the illumination augmentation, the
Figure 10. Comparison of the ground truth for haemorrhages on two images from DiaRetDB2. Coarse boundaries are marked with blue and representative regions are marked with red.

same, but without hard background sampling and the full setup with augmentation by applying illumination fields with noisy parameters. The results are shown in Table 1. The following evaluation metrics, described in Section 3, are given: sensitivity, specificity, area under the ROC curve, and area under the PR curve. Besides, ROC and PR curves for the run with the full setup are shown in Figure 11 and Figure 12.

Table 1. Performance results of haemorrhage and exudate detection for three runs with different setups on hyperspectral images from DiaRetDB2. ACC - accuracy, ROC AUC - area under the receiver operating characteristic curve, SN - sensitivity, SP - specificity, PR AUC - area under the precision-recall curve.

<table>
<thead>
<tr>
<th>Type</th>
<th>ACC</th>
<th>ROC AUC</th>
<th>SN</th>
<th>SP</th>
<th>PR AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.96</td>
<td>0.97/0.99</td>
<td>0.82/0.89</td>
<td>0.96/0.99</td>
<td>0.15/0.18</td>
</tr>
<tr>
<td>No hard background</td>
<td>0.99</td>
<td>0.93/0.83</td>
<td>0.4/0.42</td>
<td>0.99/0.99</td>
<td>0.07/0.14</td>
</tr>
<tr>
<td>Full + illumination</td>
<td>0.99</td>
<td>0.9/0.96</td>
<td>0.42/0.37</td>
<td>0.99/0.99</td>
<td>0.03/0.05</td>
</tr>
</tbody>
</table>

From the table one can see that ACC, SP and ROC AUC are relatively high for all the runs. However, SP and PR AUC clearly show the advantage of the first setup. It appears that in general haemorrhage detection is more challenging than exudate detection, since
the metrics for it are lower in most cases. Although this is not true for all the cases, this is, probably, due to the fact that in DiaRetDB2 there are significantly more samples of haemorrhages than for exudates. Good, average and bad examples of haemorrhage and exudate detection from HSI are shown in Figure 14. The masks were chosen based on the PR AUC for each image.

![Figure 11](image1.png)

**Figure 11.** ROC and PR curves for haemorrhages detection on hyperspectral images from DiaRetDB2.

![Figure 12](image2.png)

**Figure 12.** ROC and PR curves for exudates detection on hyperspectral images from DiaRetDB2.

It appears to be interesting to compare dimensionality reduction performed by the network with a standard dimensionality reduction method such as PCA. One example of this comparison is shown in Figure 13. PCA was run on spectra of a random image from DiaRetDB2 to reduce its dimensionality from 30 to 3 and the same was performed by the first block of DiaRetNNet with the learned weights.
Figure 13. The first three PCA components of a random image from DiaRetDB2 and the same image after DiaretNNet dimensionality reduction. Blue boundaries show the markings for haemorrhages.

Figure 14. Examples of binary classification masks obtained on hyperspectral images for exudates (top) and haemorrhages (bottom). The examples are sorted by the area under the precision-recall curve from the highest on the left to the lowest on the right. Coarse ground truth boundaries are marked with blue and representative regions are marked with red.
6.2 Performance on RGB images

For detecting lesions on RGB images, the set up was identical to those for hyperspectral images with the exception that the first block of the neural network for dimensionality reduction was removed. Such setting enables us to fairly compare the performance of the same network on RGB and HSI, and understand whether HSI provides any benefit for the lesion detection. In order to compensate the absence of the first block, an extra run was performed with one 1x1 convolutional layer with three filters before the first block. This allows the network to learn conversion to some other colour space more suitable for the purpose, and was shown to improve CNN performance on popular image datasets [46].

6.2.1 Evaluation on DiaRetDB2 database

Images for which the experiments are given in this section were acquired by converting DiaRetDB2 spectral images to the RGB color space. After the conversion, no corrections were performed other than the aforementioned channel-wise minmax normalization. The performance results for runs with the learned colour conversion and without it are given in Table 2 and Figures 15 and 16.

<table>
<thead>
<tr>
<th>Type</th>
<th>ACC</th>
<th>ROC AUC</th>
<th>SN</th>
<th>SP</th>
<th>PR AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain colour</td>
<td>0.99</td>
<td>0.96/0.81</td>
<td>0.45/0.24</td>
<td>0.99/0.99</td>
<td>0.08/0.05</td>
</tr>
<tr>
<td>Colour conversion</td>
<td>0.95</td>
<td>0.94/0.99</td>
<td>0.67/0.91</td>
<td>0.96/0.99</td>
<td>0.08/0.26</td>
</tr>
</tbody>
</table>

The results provide evidence that the colour conversion is beneficial for the purpose. Significant difference in performance between the two runs can be interpreted as the evidence that the colour correction is needed for the RGB images acquired by the conversion from HSI. In the second run, SN and PR AUC for exudate detection are the highest even in comparison to runs on the hyperspectral images. However, performance on the haemorrhages is still considerably lower, and the overall result on HSI appears to be better.
Figure 15. Receiver operating characteristic and precision-recall curve for haemorrhage detection from plain RGB images from DiaRetDB2.

6.2.2 Evaluation on BristolDB database

BristolDB database [14] contains 107 RGB images with ground-truth for exudates. Unlike markings for DiaRetDB2, the labels for BristolDB are assumed to be pixel-perfect. By comparing evaluating performance of the proposed CNN architecture on this database, it is both possible to understand how efficient the algorithm is by itself, and how much the quality of ground-truth affects its performance. Since BristolDB contains ground truth labels only for one class, classification task becomes binary. Nevertheless, multi-class softmax activation function was used on the output layer with two neurons in order to keep the setup similar to the previous experiments. For the same reason, the spatial size of input patch was kept to be 35x35, although the images in BristolDB are of size 536x540. Thus, the obtained experimental results, shown in Table 3 and Figures 17 and 18, are presumably suboptimal with respect to what could be achieved with proper adjustments.

Figure 16. Receiver operating characteristic and precision-recall curve for exudate detection from plain RGB images from DiaRetDB2.
Table 3. Performance results for runs on RGB images from BristolDB. ACC - accuracy, ROC AUC - area under the receiver operating characteristic curve, SN - sensitivity, SP - specificity, PR AUC - area under the precision-recall curve.

<table>
<thead>
<tr>
<th>Type</th>
<th>ACC</th>
<th>ROC AUC</th>
<th>SN</th>
<th>SP</th>
<th>PR AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain colour</td>
<td>0.98</td>
<td>0.98</td>
<td>0.87</td>
<td>0.98</td>
<td>0.52</td>
</tr>
<tr>
<td>Colour conversion</td>
<td>0.99</td>
<td>0.98</td>
<td>0.84</td>
<td>0.99</td>
<td>0.6</td>
</tr>
</tbody>
</table>

The RGB images from BristolDB seem to be of better visual quality than the images of DiaRetDB2. This may explain the small difference between the two runs on BristolDB: the images do not need any extra colour transformations for improving the exudate detection performance. What is more, the PR AUC is considerably higher for the both runs on BristolDB than for any run on DiaRetDB2. To some extent, this may be explained by the fact that the classification of three classes is a more difficult task, but undoubtedly the accurate ground truth markings played an important role in it.

Figure 17. Receiver operating characteristic and precision-recall curve for plain RGB images from BristolDB.

Figure 18. Receiver operating characteristic and precision-recall curve for RGB images with learned colour conversion from BristolDB.
7 DISCUSSION AND FUTURE WORK

7.1 Discussion of the results

Generally, it is easy to see that the area under the precision-recall curve is the most problematic characteristic for all the experiments. It is relatively low even with high values of sensitivity and with accurate markings. Thus, low precision can be viewed as the most important drawback of the proposed algorithm. In combination with high SN and SP, this means that in practice this method might be effectively utilized for highlighting possibly abnormal areas, but not for the final diagnosis.

Among the classification results obtained on DiaRetDB2 database, the one obtained with sampling of hard background and without augmentation through illumination fields appears to be the best. The reasons for its advantage over random background sampling were discussed in Section 5.2. Lower results of the last method might be due to the inappropriateness of the utilized illumination model or variations in the initial illumination fields in the bands, which were approximated by a single field fitted to the value channel of the corresponding colour image.

Figure 13 provides evidence that illumination is an important factor. One can see that the right parts of both images have different colour than the rest of the images, which is caused by imperfect illumination conditions. Another interesting observation from this figure is that some blood vessels are shrunk or cut. Such effect might possibly be caused by sampling of lesion-like background. Spectra of blood vessels are close to haemorrhages in the Euclidian space, hence they are often sampled, and the network tries to remove vessels since they are not of interest for the purpose. PCA version of dimensionality reduction highlights vessels and macula because they have high variations across the spectral bands. Unfortunately, although in the image produced by the network one can distinguish small spots corresponding to lesion markings, there are many other elements with high intensity. It means that the dimensionality reduction block did not manage to learn properly, possibly due to scarcity of spectra with lesions or inaccurate markings.

SP, accuracy and AUC for ROC curve are relatively high for all runs. These metrics along with SN are the ones, which are used the most frequently in machine learning literature including the works on lesion detection. However, in the case of pixel-wise classification, distribution of classes is highly skewed and in such setting these metrics may appear mis-
leading. In particular, SP does not show how many FPs were obtained with respect to the number of positive samples, but only with respect to the amount of negatives. Therefore, if positives are underrepresented, e.g., haemorrhages in comparison to the background, the large amount of FP will be negligible in the final value of the metric. To address this, AUC for PR curve was utilized, which shows the main shortcoming of the proposed method.

7.2 Future work

The most important issue of DiaRetNNet architecture and proposed training strategy appears to be the tendency to high false positive rate. This property persisted with all types of data, which was considered in this work, causing low AUC for PR curve. This might be improved by tuning hyperparameters, such as concentrations of each of the types of data in a training batch or different weights for samples of different classes in the cost function. Another way might be to change the architecture and map each input patch to class map instead of a probability distribution on classes. Such approach would not restrict the target pixel to be in the center of the patch and allow random shifting as a data augmentation method. What is more, the described approach would overcome another important shortcoming of the proposed algorithm: relatively demanding and slow processing of input images. With the current setup, the speed of processing of a single HSI image is about five minutes, mostly due to the fact that the patch around each pixel should be extracted and processed. A network mapping patches to class maps would be able to process images much faster. However, it is not clear whether such approach would improve the false positive rate since it would require more complicated inference with the same input.

The fact that all the performance metrics are considerably higher on hyperspectral images than on RGB suggests that HSI may be advantageous for the purpose of lesion detection. However, since no corrections were applied to RGB, such result may be an issue of the conversion algorithm. Further investigation would be needed to confirm HSI benefits.

Evaluation of the algorithm on the BristolDB showed that accurate ground truth is essential for training an efficient lesion detection algorithm. Thus, further work is required for enhancing the markings available in DiaRetDB2. Namely, the current coarse and representative regions could be utilized as input to some segmentation algorithm, which output may be then given back to experts for making it more accurate. Further, a fusion algorithm would be needed to produce single ground truth from multiple experts’ markings.
8 CONCLUSION

Diabetes is a group of widely spread diseases caused by the inability of a human organism to regulate the level of blood sugar. Prolonged diabetes may become a cause of various complications, one of which is diabetic retinopathy. In the early stages it can be diagnosed only with fundus photography by detecting lesions in the retina. Due to the high rates of this abnormality, the demand for such diagnostics is large and automatic methods are of interest.

In this work, an overview of existing techniques in the fields of DR lesion detection and HSI feature extraction was given. Further, they were compared and analyzed. A new method for lesion detection from hyperspectral images based on the convolutional neural network was proposed and evaluated on DiaRetDB2 dataset. Specific features of the task were discussed and reflected in the utilized methods for data augmentation and sampling strategy. The proposed architecture is easily adjustable for colour images, which allowed the comparison of performance on the lesion detection task between them and the spectral images. Results of the experiments suggest that information in hyperspectral images may be beneficial for lesion detection if it is effectively utilized.
REFERENCES


[38] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In In CVPR, 2009.


