FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Tuukka Bogdanoff

FUNDUS IMAGE ANALYSIS THROUGH DIMENSION REDUCTION USING UNSUPERVISED LEARNING

Bachelor's Thesis
Degree Programme in Computer Science and Engineering
October 2020

Bogdanoff T. (2020) Fundus Image Analysis through Dimension Reduction Using Unsupervised Learning. University of Oulu, Degree Programme in Computer Science and Engineering, 31 p.

ABSTRACT

The fundus of the human eye can be affected by many vision-threatening diseases. Early detection of such diseases is crucial to prevent damage to eyesight. Fundus photography is considered one of the most important and least invasive methods for this purpose. However, detecting signs of the diseases in fundus photographs is a laborious and time consuming process, which motivates the development of machine learning based tools to assist medical professionals in the diagnostic process.

This thesis presents machine learning based methods for fundus image analysis proposed in literature with a focus on the detection of signs related to diabetic retinopathy, which is one of the leading causes of preventable blindness in the world. In addition, a new method for fundus image analysis based on dimensionality reduction using unsupervised machine learning methods, autoencoder and the UMAP algorithm, is proposed.

The literature review presented in this thesis indicates that machine learning based methods for fundus image analysis can learn to discriminate between healthy and diseased cases with high accuracy, holding much potential for assisting healthcare professionals in a screening setting. The experiments conducted on the proposed framework indicate that a feature vector produced by a model based on unsupervised learning can retain relevant information for the task of fundus image analysis and can provide healthcare professionals with easily interpretable visualizations. However, the proposed approach was unable to benefit from training on relevant image data, which suggests that the framework requires further development to achieve a suitable solution to the problem based on unsupervised learning.

Keywords: feature extraction, diabetic retinopathy, UMAP, random forest, macular edema, exudate

Bogdanoff T. (2020) Silmänpohjakuvien analysointi dimension pudotuksella käyttäen ohjattua oppimista. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 31 s.

TIIVISTELMÄ

Ihmisen silmänpohjaan voivat vaikuttaa useat näköä uhkaavat sairaudet. Näiden sairauksien aikainen havaitseminen on tärkeää näön vahingoittumisen estämiseksi. Silmänpohjan valokuvausta pidetään yhtenä tärkeimmistä ja vähiten invasiivisista menetelmistä tähän tarkoitukseen. Tautien merkkien havaitseminen silmänpohjakuvista on kuitenkin työläs ja aikaa vievä prosessi, mikä motivoi koneoppimiseen perustuvien työkalujen kehittämistä terveydenhuollon ammattilaisten avuksi diagnosoinnissa.

Tässä tutkielmassa esitellään kirjallisuudessa esitettyjä koneoppimiseen perustuvia menetelmiä silmänpohjakuvien analysointiin keskittyen diabeettiseen retinopatiaan, joka on yksi suurimmista vältettävissä olevan sokeutumisen aiheuttajista maailmassa. Lisäksi esitetään uusi ohjaamattomaan oppimiseen perustuviin dimension vähentämismenetelmiin, autoenkooderiin ja UMAP-algoritmiin, perustuva silmänpohjakuvien analysointimenetelmä.

Tutkielmassa esitetty kirjallisuuskatsaus osoittaa, että koneoppimiseen pohjautuvat menetelmät silmänpohjakuvien analysointiin voivat oppia erottamaan terveet ja sairaat tapaukset toisistaan korkealla tarkkuudella, minkä perusteella menetelmillä on paljon potentiaalia olla avuksi terveydenhuollon ammattilaisille seulontaympäristössä. Esitetyllä lähestymistavalla suoritetut kokeet osoittavat, että ohjaamattomaan oppimiseen perustuvan mallin tuottama piirrevektori pystyy säilyttämään silmänpohjakuvan analysoinnin kannalta oleellista tietoa ja voi tarjota terveydenhuollon ammattilaisille helposti tulkittavia visualisointeja. Esitetty lähestymistapa ei kuitenkaan hyötynyt opetuksesta olennaisella kuvadatalla, minkä perusteella lähestymistapa vaatii kehitystä pidemmälle ongelmaan sopivan ohjaamattomaan oppimiseen perustuvan ratkaisun aikaansaamiseksi.

Avainsanat: diabeettinen retinopatia, UMAP, satunnaismetsä, makulaturvotus, eksudaatti

TABLE OF CONTENTS

FOREWORD

This Bachelor's thesis was composed in the center for machine vision and signal analysis (CMVS) in the University of Oulu for the purpose of evaluating the suitability of a machine learning framework based on unsupervised learning to fundus image analysis. I would like to thank my supervisor professor Olli Silvén for the opportunity of working at the research center and Riku Hietaniemi for suggesting the topic of the thesis. I would also like to thank Seyyedjavad Hosseininia for his invaluable assistance in developing the framework proposed in the thesis.

Oulu, October 13th, 2020

Tuukka Bogdanoff

LIST OF ABBREVIATIONS

AMD	age-related macular degeneration
AUC	area under the receiver operating curve
CNN	convolutional neural network
DME	diabetic macular edema
DR	diabetic retinopathy
FOV	field of view
FPI	false positive detections per image
FPR	false positive rate
HE	hard exudate
kNN	k nearest neighbors
NPDR	nonproliferative diabetic retinopathy
OD	optic disk
PDR	proliferative diabetic retinopathy
MA	microaneurysm
MSE	mean square error
RDR	referable diabetic retinopathy
RF	random forest
ROC	receiver operating characteristics
SE	soft exudate
UMAP	Uniform Manifold Approximation and Projection

1. INTRODUCTION

The fundus of the eye can be affected by a number of different vision-threatening conditions. Some of the most prominent are diabetic retinopathy (DR), macular edema and age-related macular degeneration (AMD). For each of these diseases, early diagnosis is crucial for preventing extensive damage to the retina. Fundus photography is an invaluable tool for screening patients at risk of being affected by these diseases. This presents the possibility of using a machine learning based approach to assist in the screening process as such approaches have been found to perform well on medical image analysis (Litjens et al., 2017).

This thesis will focus on two common vision-threatening diseases of the fundus: diabetic retinopathy and diabetic macular edema. A literary review of machine learning based approaches for the detection of these diseases is provided. In addition, a new approach for fundus image analysis based on dimension reduction through unsupervised machine learning is proposed and evaluated.

Diabetic retinopathy is a vascular disease caused by diabetes that affects the fine vessels of the retina (Gargeya & Leng, 2017; Kauppi et al., 2007). It is one of the most common causes of preventable blindness in the world as approximately 40-45% of diabetics will likely suffer from DR at some point in their life (Gargeya & Leng, 2017). The number of people with diabetes is expected to rise from 451 million in 2017 to 693 million by 2045 (Cho et al., 2018). Given the increasing prevalence of diabetes and the high risk of DR in diabetics, the number of patients suffering from DR can be expected to rise significantly.

Diabetic retinopathy is divided into three stages of nonproliferative diabetic retinopathy (NPDR) and one stage of proliferative diabetic retinopathy (PDR). The earliest detectable signs of DR are usually microaneurysms (MAs) (Antal & Hajdu, 2012).

When MAs are the only visible abnormalities in the retina, the stage of DR is classified as mild nonproliferative diabetic retinopathy (Wilkinson et al., 2003). When the DR progresses from this stage, more MAs appear in the retina, and some of them can rupture and leak blood onto the retina, which appears as red lesions called hemorrhages (Orlando, Prokofyeva, del Fresno, & Blaschko, 2018). In the presence of other signs of DR besides MAs, the stage of DR is classified as moderate nonproliferative retinopathy as long as it does not fulfill the definition of severe nonproliferative retinopathy (Wilkinson et al., 2003).

As the retinopathy advances, lesions called hard exudates (HEs) and cotton wool spots may start appearing as bright yellow spots in the fundus. Cotton wool spots are also called soft exudates (SEs) although this is a misnomer as these lesions are not exudates (Schmidt, 2008). "Bright lesions" is a common term that refers to exudates, cotton wool spots as well as drusens, which are associated with AMD (Niemeijer, van Ginneken, Russell, Suttorp-Schulten, & Abramoff, 2007). Cotton wool spots appear in the fundus because of microinfarcts occurring due to the obstruction of retinal blood vessels (Kauppi et al., 2007). Hard exudates are formed by lipids leaking out of the weakened blood vessels as the disease advances (Kauppi et al., 2007).

The most progressed stage of NPDR, severe nonproliferative diabetic retinopathy, is characterized by the presence of venous beading and intraretinal microvascular abnormalities in the retina (Wilkinson et al., 2003). If DR progresses to this stage,

it poses a high risk for developing PDR (Wilkinson et al., 2003). DR is defined as PDR if any neovascularization, vitreous hemorrhages or preretinal hemorrhages are present on the retina (Wilkinson et al., 2003).

Diabetic macular edema (DME) is defined as the thickening of the retina, which is caused by the fluid leaking onto the retina from damaged blood vessels (Wilkinson et al., 2003; Giancardo et al., 2012). Determining the thickness of the retina definitively requires three-dimensional evaluation of the fundus, which is often relatively difficult due to the lack of appropriate equipment or experienced personnel (Wilkinson et al., 2003). This is why DME screening often relies on determining the presence of hard exudates on the retina as they usually appear in association with significant macular edema (Wilkinson et al., 2003; Ciulla, Amador, & Zinman, 2003; Giancardo et al., 2012).

1.1. Data Sets

In recent years, the scientific research surrounding digital fundus image analysis has been driven by the increased number of publicly available data sets of fundus images. The publicly available fundus image data sets vary significantly in purpose and comprehensiveness of the provided ground truth information, which makes choosing the most appropriate data set for a given task an important step in developing an algorithm for fundus image analysis. The public availability of fundus image data sets with lesion level annotations has been especially crucial to the development of new approaches based on deep learning.

The Messidor data set contains 1200 eye fundus color images that were captured at 3 ophthalmologic departments in France. The images were captured at a 45 degree field of view (FOV) at resolutions of 1440x960, 2240x1488 or 2304x1536 pixels with 8 bits of depth per color plane. Pupil dilation was used when acquiring 800 of the images, and 400 images were collected without dilation. The data set contains a medical diagnosis for each image but no manual annotations for structures such as lesions or vasculature (Decencière et al., 2014).

The DIARETDB1 database consists of 89 color fundus images of which 84 show signs of at least mild NPDR and 5 are free of any signs of diabetic retinopathy according to all experts who contributed to the evaluation of the data set. The images were taken at the Kuopio university hospital in Finland. The images were captured at a 50 degree field of view with varying imaging settings. The data set contains lesion level annotations for microaneurysms, hemorrhages, hard exudates and soft exudates created by medical experts for each image. The data set is also provided with a predefined split into a training set of 28 images and a test set of 61 images (Kauppi et al., 2007).

The e-optha database consists of fundus images collected during the years 2008 and 2009 through the OPHDIAT teleophthalmology network. The images resulted from 25702 examinations, each containing at least four images as well as additional contextual information. Some of these images were used to develop two publicly available data sets called e-optha EX and e-optha MA which contain manual annotations for exudates and microaneurysms respectively. The first data set, e-optha EX, contains 47 images with a total of 12278 exudates in addition to 35 images of

healthy fundi. The second data set, e-optha MA, contains 148 images with 1306 microaneurysms and 233 healthy images (Decencière et al., 2013).

The Indian Diabetic Retinopathy Image Dataset (IDRiD) consists of 516 retinal fundus images obtained at an eye clinic in Nanded, (M.S.), India. The images were captured with a 50 degree field of view at a resolution of 4288x2848. Pupil dilation was used when capturing the images. Pixel level annotations of the optic disk and lesions typical to DR were provided for 81 of the images. In addition, all the images have been graded for the severity of DR and DME. The center coordinates for the optic disk and fovea are also provided for each image (Porwal et al., 2018).

2. RELATED WORKS

A number of machine learning based approaches have been proposed for the detection of DR and DME in literature. The approaches usually concentrate on detecting one type of sign of a disease, although more general approaches have also been proposed. In recent years, methods based on deep learning have become prominent due to an increasing amount of relevant publicly available data. Deep learning based models called convolutional neural networks, which are known to perform well in analyzing image data, represent the current state of the art in automated fundus image analysis.

Niemeijer, van Ginneken, Staal, Suttorp-Schulten, & Abramoff, 2005 proposed a method for detecting red lesions in fundus images by employing a three-stage framework involving preprocessing, detecting lesion candidates and classifying candidates as positive or negative findings of red lesions. They preprocessed their data to compensate for the gradually varying background intensity in fundus images and removed bright lesions from the pictures, as their focus was on red lesions only. For detecting potential red lesion candidates, they combined a mathematical morphology method based on previous research with their own pixel classification method based on a k nearest neighbors (kNN) classifier.

To classify the extracted candidates as containing red lesions or not, they extracted a set of 68 hand-crafted features which were used to train another kNN. To train the pixel classifier for candidate detection, they used a publicly available data set called DRIVE (Staal, Abramoff, Niemeijer, Viergever, & van Ginneken, 2004). To train the completed system, they combined 26 fundus images from a DR screening program with 74 images obtained from a referral hospital to form a data set of 100 images. An ophthalmologist annotated this data set by marking all pixels they considered to be part of red lesions. This data was randomly split into a training and test set of 50 images each.

The authors report a sensitivity of 100% at a specificity 87% for the task of determining whether an image contains red lesions on the test set. The sensitivity of the system on a per-lesion basis was reported to be 30%.

The result is a promising proof of concept that demonstrates the effectiveness of preprocessing and utilization of domain knowledge in red lesion detection. However, the system suffers from a lack of scalability due to the use of kNN classifiers which perform poorly with a large amount of data which is necessary for a generally applicable solution. The small amount of data used in producing the result also represents only a small portion of the world's population. The algorithm would have to be trained and evaluated using more diverse data and adjusted accordingly to make the algorithm more generally applicable.

2.1. Deep Learning Approaches

Deep learning is a discipline of machine learning based on the utilization of neural network models. Unlike traditional machine learning approaches which rely largely on hand crafted features, a neural network automatically infers a set of discriminative features for a given task from the data it is trained on. Neural network models of a specific type called convolutional neural network (CNN) have been especially

successful in image analysis, including a variety of medical applications. In recent years, convolutional neural network have become one of the most powerful and popular approaches for automatic fundus image analysis. In a 2015 competition for the detection of diabetic retinopathy organized by Kaggle, the majority of the 661 participating teams applied deep learning and four teams utilizing end-to-end CNNs achieved above human level performance (Litjens et al., 2017).

Gulshan et al., 2016 developed an algorithm for detecting referable diabetic retinopathy (RDR), which is defined as moderate and worse DR, and referable, i.e. clinically significant, diabetic macular edema by training a deep convolutional neural network. The CNN was based on the Inception-v3 architecture (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). The output of the network consisted of multiple binary predictions, predicting the presence of referable DME in addition to detecting moderate or worse and severe or worse cases of DR.

The network was trained using a total of 128 175 fundus images obtained from 3 eye hospitals in India and EyePACS in the USA. In addition, the Messidor-2 data set (Decencière et al., 2014; Quellec et al., 2008) was used for validation. All images in the data sets were graded by a panel of licensed ophthalmologists for the severity of DR and image quality. The performance of the algorithm was tested on two separate validation sets that were not used in training. One consisted of 9963 images from the EyePACS-1 data set, with the other one being the Messidor-2 data set.

The authors analyzed the performance by assessing the receiver operating characteristics (ROC) curve, reporting an area under the receiver operating curve (AUC) of 0.991 on the EyePACS-1 validation set and an AUC of 0.990 on Messidor-2 for the task of detecting RDR. They also selected two operating points from the curve with one being selected for high specificity and the other for high sensitivity. At the operating point chosen for high specificity, they report a sensitivity of 90.3% and specificity of 98.1% on the EyePACS-1 validation set and a sensitivity of 87.0% and specificity and 98.5% on Messidor-2. At the other operating point chosen for high sensitivity, they report a sensitivity of 97.5% and specificity of 93.4% on the EyePACS-1 validation set and a sensitivity of 96.1% and specificity of 93.9% on Messidor-2.

The results are promising and demonstrate that a CNN can learn a set of discriminative features for detecting RDR. However, this algorithm was developed to detect moderate and worse cases of DR and cannot detect the earliest stages of nonproliferative DR which are the most difficult to determine even for professionals.

Orlando et al., 2018 proposed an automated method for detecting red lesions in fundus images based on combining hand crafted features with the features learned by a CNN. The first step in their framework for red lesion detection is an unsupervised candidate detection phase based on morphological operations. The pixels around the retrieved candidates are then used to train a CNN to learn a set of features. These features are augmented with hand crafted ones based on shape and intensity. Finally, the resulting feature vectors are used to train a random forest classifier to discriminate between false candidates and true lesions.

The performance of the method was evaluated in two experimental setups. In the first experiment, the model was trained on the training set from DIARETDB1 consisting of 28 images. This model achieved an AUC of 0.8932 and a sensitivity of 0.9109 at a specificity of 50% in the task of determining whether a fundus image contains any signs of DR on the MESSIDOR data set. For detecting RDR on MESSIDOR, the

model achieved an AUC of 0.9347 and a sensitivity of 0.9721 at a specificity of 50%. The per-lesion performance of this model was tested on the DIARETDB1 test set. In this setting, the model achieved a sensitivity of 0.4883 for detecting individual red lesions when the number of false positive detections per image (FPI) was 1.

In the second experiment, the model was trained on the training set from DIARETDB1 and the data set provided in the Retinopathy Online Challenge (Niemeijer et al., 2009). This model achieved a per-lesion sensitivity of 0.3680 at an FPI of 1 for the task of detecting red lesions on the e-optha data set.

The method provides promising results and reportedly outperforms previous approaches in the tested settings. The result demonstrates the efficiency of CNNs in medical image analysis and the apparent importance of employing domain knowledge for the task of detecting red lesions in fundus images. However, the specificity achieved by the method is still rather low at acceptable levels of sensitivity for a screening setting, which suggests that there is also much room for further improvements.

3. IMPLEMENTATION

This chapter proposes a feature extraction framework utilizing an unsupervised neural network model called convolutional autoencoder for the purpose of analyzing fundus images. The purpose of the experiments is to determine whether a neural network model based solely on unsupervised learning can learn to extract a useful representation for detecting signs of diseases in fundus images.

In order to evaluate the usefulness of the feature extractor, the feature vectors were visualized with bright lesion ground truth information using a dimensionality reduction algorithm called Uniform Manifold Approximation and Projection (UMAP). A random forest classifier trained using the features is used to evaluate the usefulness of the feature extractor in a classification framework.

The motivation behind using an unsupervised feature extractor instead of relying entirely on supervised learning is that an unsupervised model does not need be trained on data with detailed ground truth annotations which are expensive and laborious to generate. Due to not relying on the ground truths generated by humans, an unsupervised feature extractor does not learn to represent the biases or mistakes of individuals, which could potentially allow it to produce more generally applicable results.

3.1. Materials

The experiments were conducted on three openly available data sets of fundus images: Messidor, DIARETDB1 and e-optha EX. The images from the Messidor data set were used for training the convolutional autoencoder neural network architecture. To ensure a more consistent set of fundus images for this purpose, some images were manually excluded from the Messidor data set, leaving only 1010 of the original 1200 images to be used in the experiments. The images from DIARETDB1 and e-optha EX were used for evaluating the representation learned by the autoencoder because they contain detailed ground truth information for multiple signs related to DR and DME.

As the data set did not contain the necessary FOV masks for the fundus images, the masks were generated automatically using the method described by Orlando et al., 2018. In this method the FOV masks are extracted by converting the RGB images to CIELab format and thresholding the luminosity channel of the resulting image. After this, the resulting binary mask is processed with a median filter using square windows of side 5 to reduce the effects of noise and only the largest connected component is preserved. To obtain the FOV masks used in these experiments, the values of the luminosity channel were normalized to a range between 0 and 1, after which the channel was thresholded at an empirically determined value of 0.05 for the images in Messidor and 0.02 for the images in DIARETDB1 and e-optha EX. The thresholds differ from those used by Orlando et al. for the same images due to different preprocessing steps taken.

3.1.1. Subimage Extraction

In order to gain a more consistent set of images for training the autoencoder model, a set of smaller subimages was extracted from each fundus image instead of training the model on the entire fundus images. The width of the subimages extracted from a given image was 0.25 times the width of the FOV mask of the image and the height of the subimage was half the width of the subimage. An algorithm was developed to choose the locations of the subimages randomly inside the FOV mask so that the intersection over union between any two images would not exceed 0.5 and no subimage would overlap with the optic disk. A total of 24851 subimages were extracted from the fundus images in Messidor, 2002 were extracted from the images in DIARETDB1 and 1118 from the images in e-optha EX. An example result of subimage extraction is presented in Figure 1.

To obtain the set of samples used for training the autoencoder model, the images were normalized and resized to a resolution of 192x96. In addition, only the green channel of the images was used as this channel gives the highest contrast for the lesions associated with most diseases of the fundus.

For the DIARETDB1 and e-optha EX data sets per-subimage ground truth information was obtained by extracting subimages from the same coordinates in the corresponding consensus maps for DIARETDB1 or segmentation ground truth masks for e-optha EX. These per-lesion ground truths were also used to determine the label of each subimage. The ground truth labels could not be determined for the subimages extracted from Messidor as this data set does not contain lesion-based ground truth annotations.

As the experiments presented in this thesis only focus on the detection of bright lesions related to DR and DME, only the ground truth labels for these types of lesions had to be determined. For the subimages from DIARETDB1, a $\geq$75% level of agreement was required when determining the labels for HEs and SEs. The subimages from e-optha EX were labeled as containing exudates if the corresponding ground truth segmentation map contained any annotations for exudates. 285 of the 2020 subimages extracted from DIARETDB1 and 217 of the 1118 subimages from e-optha EX were labelled as containing bright lesions as described.

3.1.2. Locating the Optic Disk

In order to achieve as balanced a set of images as possible, the subimages were extracted so that the optic disk (OD) would not appear in them. This choice was made because representing the OD poses a difficult task for the autoencoder and this area is not of much interest for the conducted experiments. If the OD was to be included in the data set of subimages, the OD would only appear in a small portion of the resulting samples, leading to an unbalanced data set, which would make it more difficult for the autoencoder to learn a useful representation.

As the ground truth information of the location of the OD was not provided for the data sets used in the experiments, the OD locations had to be determined by annotating them manually or inferring the locations automatically. The location was determined

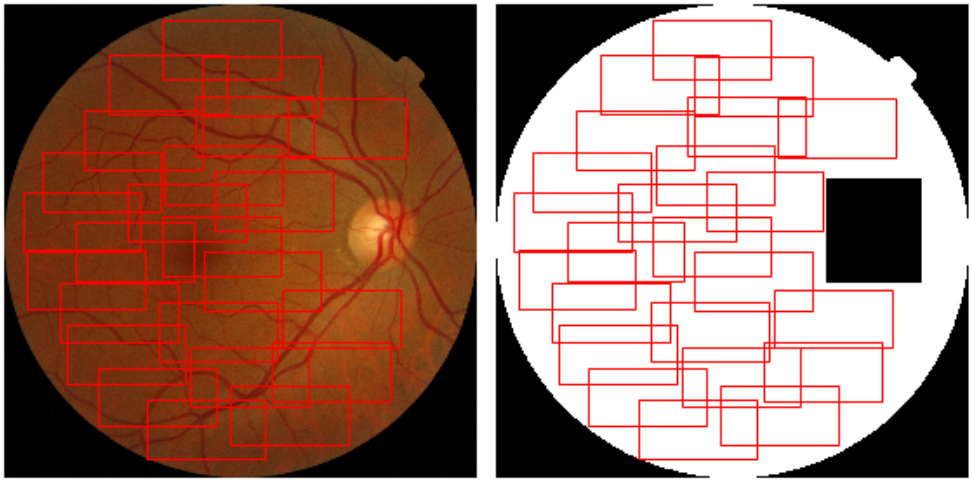

Figure 1. Example coordinates of subimages extracted from a fundus image in the Messidor data set with the image on the right illustrating the subimage coordinates with regard to the area that was considered usable during subimage extraction.

as a bounding box surrounding the OD, as accurate segmentation of the OD was not considered necessary in this demonstrative experimental setup.

In order to determine the location of the OD, a convolutional neural network of the YOLOv5 architecture was trained for this purpose (*ultralytics/yolov5: v3.0*, s.a.). Although this network is designed for the purpose of simultaneously detecting and classifying objects, thus containing unnecessary functionality for the simpler task of locating the OD, the architecture was chosen because it is well documented and efficient even for the required task. The trained YOLOv5 neural network attempts to detect the OD from a given image, and outputs the bounding boxes surrounding the detections.

In order to form a set of images for training the YOLOv5 model for detecting the OD from fundus images, the bounding box surrounding the OD was manually determined for the 89 images in the DIARETDB1 data set. In addition, the 81 images in the IDRiD data set for which OD segmentation ground truth was available were used. These images were combined to form a data set of 170 images which was further split into a training set of 145 images, a validation set of 13 images and a test set of 12 images. All the images were resized to a resolution of 416x416 for use with the YOLOv5 architecture. Data augmentation randomly consisting of vertical and horizontal flipping and rotation operations between -30° and 30° was performed on the training set to form an augmented data set of 431 images.

The network was trained on the augmented training data set for 200 epochs, and the weights during the epoch that had the best performance on the validation data were chosen for inferring the OD location for the rest of the fundus images used in these experiments. If the model had multiple detections for the OD in an image, the detection with the highest confidence was used.

Despite performing adequately well in most cases, the YOLOv5 model failed to detect the OD in 3 images from the Messidor data set and 3 images in e-optha EX. For these 6 images, the location of the OD was manually annotated by the author. For

DIARETDB1, the OD location ground truths annotated by the author were used when extracting the subimages.

3.2. Convolutional Autoencoder

A convolutional autoencoder was used for extracting an representation of lower dimension from the subimages. An autoencoder is a neural network architecture consisting of two parts: an encoder, which compresses the input into a latent representation, and a decoder, which attempts to recover the original input using only the latent features given by the encoder. In the case of a convolutional autoencoder, most of the layers in the model are convolutional layers.

The architecture of the encoder part of the model used consists of 3 convolutional layers with 8, 16 and 32 filters respectively. Each convolution operation uses a kernel of size 3x3 and is followed by a ReLU activation. Each convolution operation has a stride of 2 and is preceded by padding so that after each convolutional layer, the spatial dimension of the input is halved, and the number of channels is doubled. The third convolutional layer produces a volume with a dimension of 12x24x32, which is flattened to a vector of 9216 features. This vector is the input to a fully connected layer which produces the latent representation of 2048 features. As the model extracts a representation of 2048 features from each 192x96 image, it reduces the dimension of the samples to 11.1% of the original 18432.

The decoder part of the model consists of a fully connected layer, which increases the dimension from 2048 back to 9216. This vector is then reshaped to a volume of dimension 12x24x32 which is followed by 3 convolutional transpose layers with 32, 16 and 8 filters, each using 3x3 kernels, a stride of 2, and followed by a ReLU activation. After this, a convolutional transpose layer with one 3x3 kernel followed by a sigmoid activation is used to produce an output of the same shape as the input.

3.2.1. Training the Autoencoder

The 24851 subimages extracted from the Messidor data set were used for training the model, and the 2002 subimages extracted from DIARETDB1 were used as a validation set. The autoencoder model was trained for 50 epochs on the training set using a batch size of 32 and utilizing an Adam optimizer to speed up the training process (Kingma & Ba, 2014). A mean square error (MSE) loss function was used as the learning criterion. After the training process, the model achieved an MSE loss of 7.4422×10^{-4} on the samples on which it was trained and a loss of 7.3907×10^{-4} on the validation set. A figure of the training and validation losses throughout the training process is presented in Figure 2.

In order to evaluate whether the model learns to extract a more useful representation for the purpose of classification the longer it is trained for, a checkpoint of the weights of the model was saved every 10 epochs in addition to also saving the randomly initialized weights of the model before the training procedure.

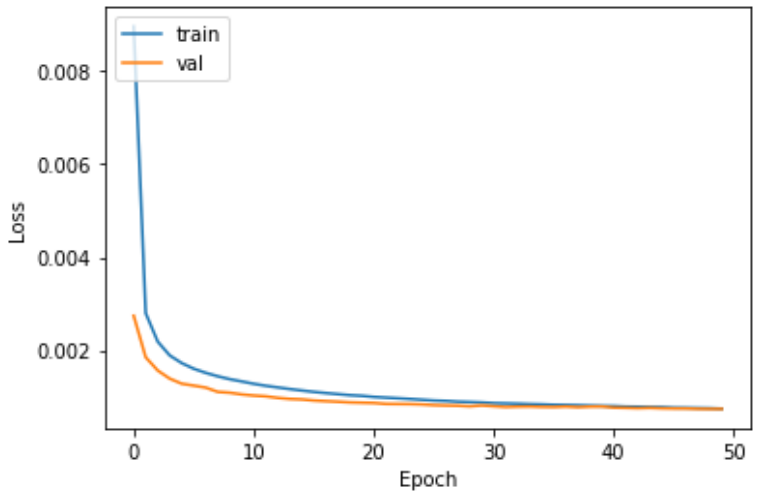

Figure 2. The mean square error loss on the training and validation sets throughout the training process.

3.3. Random Forest Classifier

In order to evaluate the usefulness of the latent features extracted by the autoencoder model in a classification scheme, a random forest (RF) classifier was trained using the feature vectors extracted by the encoder part of the model. A random forest was chosen for the classification experiments because it is robust against overfitting and can achieve good performance even on noisy or imbalanced data, which makes it suitable for the tested experimental setups (Breiman, 2001; Orlando et al., 2018).

A random forest is an ensemble consisting of T decision trees. Each decision tree is trained using an example randomly drawn with replacement from the training set used in the classification setting. Each node in a tree is determined by a split which uses the best of a maximum of $\sqrt{d}$ features where d is the dimension of the feature vector. The quality of the split is the defined as the decrease in the Gini index it produces (Breiman, 2001; Orlando et al., 2018). In his original article, Breiman suggests the final output of the random forest classifier be determined by the majority class in the predictions given by the trees, but the implementation used in the experiments presented in this thesis defines the output of the classifier as the average of the probabilistic predictions of the trees instead.

3.4. UMAP

A dimensionality reduction algorithm called Uniform Manifold Approximation and Projection (UMAP) was used to create a two dimensional projection of the feature vectors extracted by the convolutional autoencoder model to enable simple visualization of the structure in the data. UMAP was chosen for this purpose due to its ability to capture the information of local neighborhoods in the high dimensional space while also preserving much of the global structure of the original space, which has lead to its widespread adoption in a multitude of research fields. It is also more

computationally efficient comparing to popular alternatives such as t-SNE, which makes it a very powerful dimension reduction tool applicable to a wide range of practical applications (McInnes, Healy, & Melville, 2018).

UMAP is a manifold learning technique with a theoretical foundation based on Riemannian geometry and algebraic topology. The algorithm works by forming approximations of local manifolds in the high dimensional space represented using fuzzy simplicial sets and constructing a topological representation of the high dimensional data by combining these representations. An optimal projection of the data to a low dimensional space is found by minimizing the cross-entropy between the topological representation of the low dimensional projection and the topological representation of the original high dimensional data (McInnes et al., 2018).

4. EXPERIMENTS

In order to evaluate the usefulness of the latent feature representations extracted by the encoder part of the autoencoder model, a series of experiments were performed on the subimages extracted from the DIARETDB1 and e-optha EX data sets. As initial experiments indicated that the representation extracted by the autoencoder did not provide a discriminative enough set of features for the task of detecting MAs or hemorrhages, further experiments focused only on the detection of bright lesions associated with DR: SEs and HEs.

For the purpose of these experiments, a vector of 2048 features was extracted from each subimage using the encoder part of the autoencoder model. All experiments described utilize these features instead of the original subimages or entire fundus images unless stated otherwise.

For the purpose of evaluating the usefulness of the representations extracted by the model when trained for longer, experiments were conducted using three different sets of weights for the autoencoder model: the weights of the model after training for the full 50 epochs, the weights after training for 10 epochs and the randomly initialized weights of the untrained model. The usefulness of the representation was evaluated through visual inspection of the reconstruction performance, visualizing the representations extracted by the model in two dimensions utilizing UMAP, and an RF classifier was trained on the extracted feature vectors to evaluate classification performance.

4.1. Visual Evaluation of Reconstruction Performance

One of the criteria on which the autoencoder was evaluated was its capability to reconstruct the extracted subimages. Although the reconstruction itself is considered to be of less interest than the corresponding latent feature representation, evaluating the reconstruction performance can give insight to what the autoencoder learns to represent during training. As the decoder part of the model attempts to recover the original image using nothing but the latent representation produced by the encoder, the resulting reconstruction can be used as a good estimate of the information contained in the latent feature vector. However, it is important to note that the latent representation does not contain all the information present in the reconstructed image as much of the information in the reconstruction is contained in the weights learned by the decoder part of the model.

Figure 3 contains two results comparing the reconstruction performance of the same autoencoder model after training for 10 and 50 epochs. The figure illustrates the performance of reconstructing an image from the data the model was trained on by displaying the reconstructed image along with a heatmap of the square error between the reconstruction and the original image. When evaluating the results visually, the model that has been trained for longer appears to reconstruct the details of the original image much more accurately. In particular, the error corresponding to the reconstruction of the blood vessels present in the image decreases considerably when trained for longer, which can be clearly seen when comparing the two heatmaps in Figure 3.

However, while reconstruction of the blood vessels improves with training, the reconstruction error outside the veins seems to have increased slightly. This could mean that training the model for longer might not necessarily provide a more useful latent feature representation for the purpose of detecting lesions. Since more of the information present in the latent features seems to correspond to vasculature the longer the model is trained for, training for longer could actually decrease the ability of the model to represent lesions associated with diseases of the fundus.

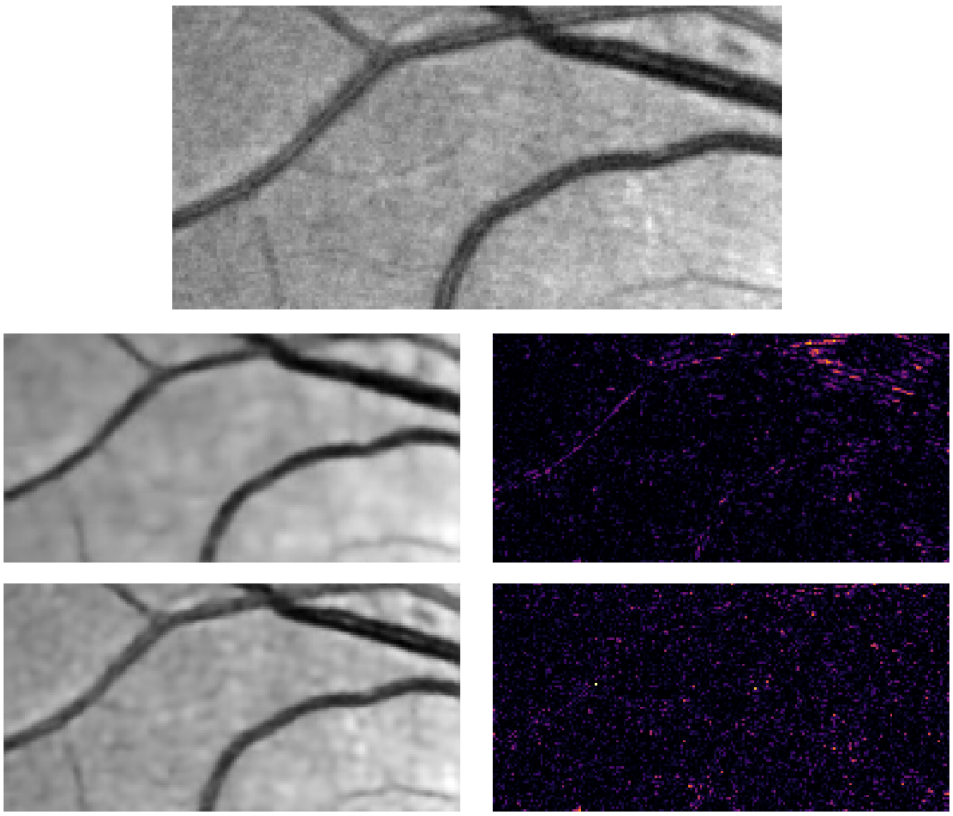

Figure 3. Visualization of the performance of the autoencoder model in reconstructing a subimage from the training set.

In Figure 3 the image on the top is the original, with the reconstructions given by the autoencoder below on the left. The images on the right are heatmaps illustrating the square error between the reconstruction and the original. The result in the middle was produced using the model trained for 10 epochs, while the result on the bottom row was given by the model trained for the full 50 epochs. The MSE of the reconstruction in the upper example is 1.5816×10^{-3} while the MSE in the bottom example is 0.87748×10^{-3}.

4.2. Visualization Using UMAP

The visualization experiments performed using the autoencoder model in conjunction with UMAP were conducted using the subimages extracted from two data sets not used in training the autoencoder model: DIARETDB1 and e-optha EX. UMAP was applied

to the features extracted from these sets to further reduce the dimension of each sample to two dimensions for visualization purposes. The two dimensional projection of each set produced by UMAP was then visualized using a scatter plot with the bright lesion ground truth label information.

If the representation extracted by the autoencoder model provides a discriminative set of features for the purpose of detecting bright lesions, the diseased samples should separate from healthy ones in the visualization; ideally forming their own clusters. Figures 4 and 5 display these scatter plots when using the fully trained autoencoder model to extract the feature vectors. To evaluate whether the autoencoder learns a better means of dimension reduction for such a visualization setting after training, the same experiment was performed on DIARETDB1 using the untrained autoencoder model. The resulting scatter plot is presented in Figure 6.

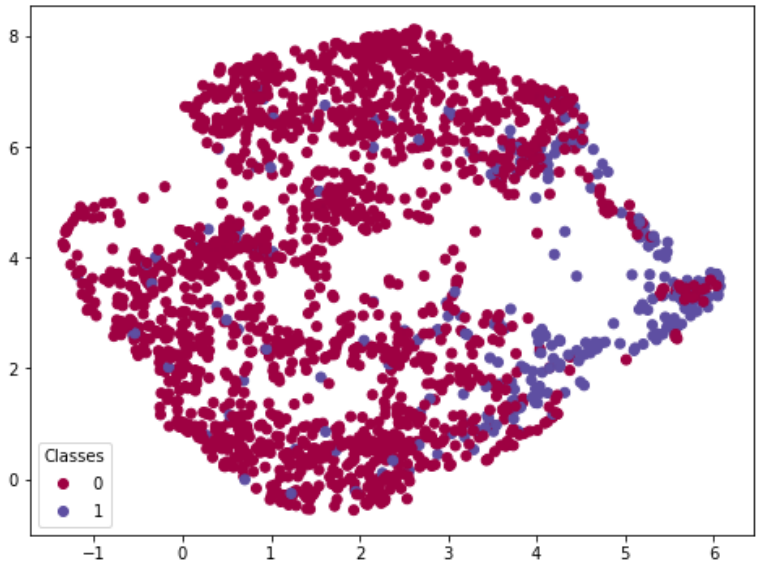

Figure 4. A UMAP projection of the feature vectors extracted from DIARETDB1 using the fully trained model with red dots corresponding to healthy samples and blue dots corresponding to diseased samples.

The scatter plots in Figures 4 and 5 clearly show signs of the diseased samples separating from healthy ones in the lower dimensional feature plane, although the samples belonging to the two classes do overlap in part. The results suggest that utilizing dimension reduction techniques for visualizing data points in a two dimensional plane with the corresponding ground truth information could provide a useful tool for professionals in determining whether an unlabeled new sample contains signs of a disease. However, as the result achieved using an untrained autoencoder model displayed in Figure 6 appears to be very similar to the corresponding result achieved using the fully trained model displayed in Figure 4, it would seem that training the autoencoder model did not improve its performance as a feature extractor for this purpose.

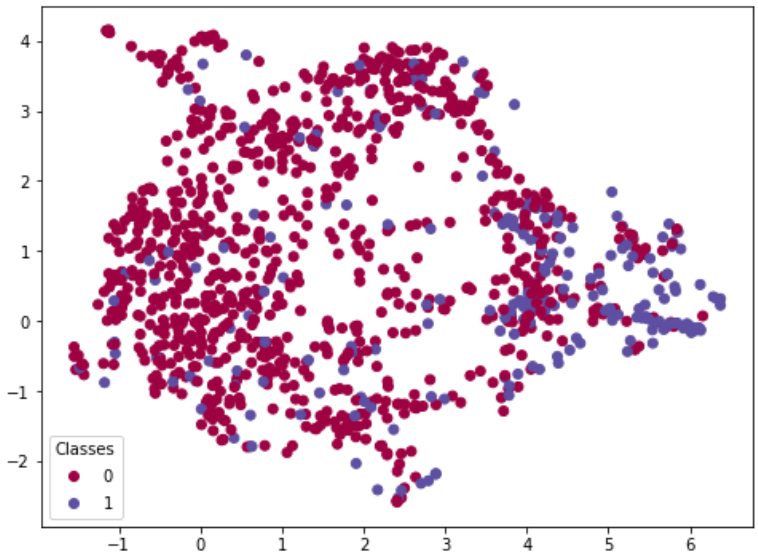

Figure 5. A UMAP projection of the feature vectors extracted from e-optha EX using the fully trained model with red dots corresponding to healthy samples and blue dots corresponding to diseased samples.

4.3. Classification Setup

In order to test the performance of the extracted representations in a classification scheme, the features were used to train a random forest classifier for detecting bright lesions in the extracted subimages. The results achieved when using the model trained for 50, 10 and 0 epochs as a feature extractor were compared in order to evaluate whether the model learns to extract a more useful representation for classification purposes when trained for longer.

For the classification setup, subimages extracted from DIARETDB1 were randomly split into a training set of 1601 samples of which 228 contained bright lesions and a test set of 401 samples of which 57 contained bright lesions. The training set was used for training the random forest classifier, and classification performance was evaluated on the test set of 401 samples as well as the 1118 subimages extracted from e-optha EX.

This approach was also tested by using the designated split into train and test sets provided with the DIARETDB1 data set. 630 subimages were extracted from the 28 fundus images in the designated train set and 1372 from the 61 images in the designated test set. In this setup, 137 of the subimages in the train set and 147 images in the test set contained bright lesions.

Although e-optha EX does not contain ground truth information for SEs like DIARETDB1, the subimages from e-optha EX were used for testing the RF trained for bright lesion detection because the detection of any type of bright lesion was considered more important for these demonstrative results than differentiating between different types of bright lesion.

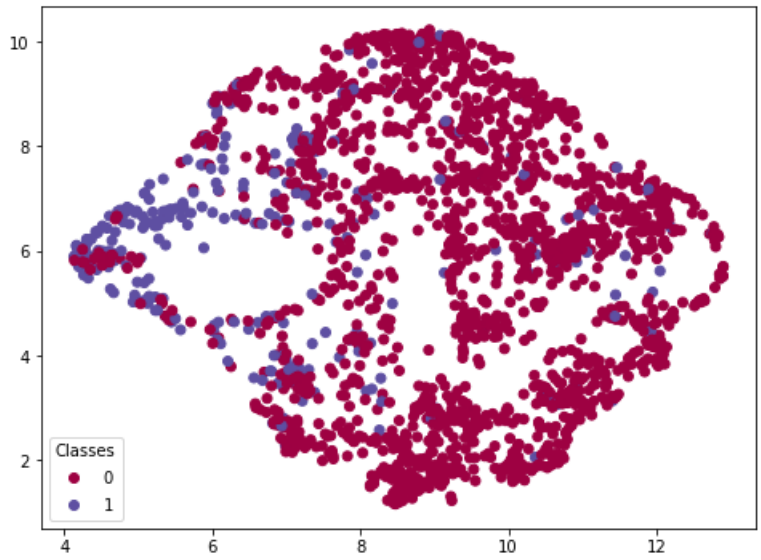

Figure 6. A UMAP projection of the feature vectors extracted from DIARETDB1 using the autoencoder model before training with red dots corresponding to healthy samples and blue dots corresponding to diseased samples.

4.3.1. Evaluation Metrics

The performance of the classification framework was evaluated based on precision, recall and specificity as well as ROC and precision-recall curves. In addition, AUC was calculated in order to summarize the information contained in the ROC curve.

Recall, also called sensitivity or true positive rate, provides the probability that a sample labelled as positive is classified as positive and is defined as follows:

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{1}$$

Precision is the percentage of correctly classified samples among the samples classified as positive and is defined as follows:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{2}$$

Specificity, also called true negative rate, indicates the proportion of correctly classified samples among the samples labelled as negative and is defined as follows:

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \tag{3}$$

When calculating the above metrics, an optimal threshold of 0.5 is used for the confidence of the classifier when determining the classification result. ROC and precision-recall curves can provide more information of the performance of the classifier by calculating the above metrics using multiple different thresholds. In ROC curves, the change in recall at different threshold values is plotted against the

corresponding false positive rate (FPR), where false positive rate is defined as the percentage of negative samples incorrectly classified as positive:

$$\text{FPR} = \frac{\text{false positives}}{\text{true negatives} + \text{false positives}} \tag{4}$$

In precision-recall curves, recall is plotted against precision similarly. Although precision-recall curves are used less frequently than ROC curves in relevant literature, they were chosen as an alternative method of evaluating classifier performance because ROC curves along with the corresponding AUC metrics tend to give misleading estimates of classification performance when experimenting on unbalanced data such as this.

4.3.2. Results

Table 1 lists the classification performances of the RF classifier trained on the randomly selected train set of 1601 subimages from DIARETDB1 when tested on the remaining 401 subimages from DIARETDB1 as well as the 1118 subimages extracted from e-optha EX. Table 2 contains the results of the framework when performing the experiments using the designated split into train and test sets provided with DIARETDB1. Figures 7, 8, 9 and 10 display the ROC and precision-recall curves associated with the experiments of Table 1.

Table 1. Performance of the RF for detecting bright lesions when trained on the train set of 1601 feature vectors extracted from subimages from the DIARETDB1 data set.

Model	Test set	Performance metric			
		Precision	Recall	Specificity	AUC
50 epochs	DIARETDB1	83.33%	26.32%	99.13%	0.866
50 epochs	e-optha EX	73.68%	19.35%	98.33%	0.666
0 epochs	DIARETDB1	78.05%	56.14%	97.38%	0.859
0 epochs	e-optha EX	63.91%	39.17%	94.67%	0.737
10 epochs	DIARETDB1	78.78%	45.61%	97.97%	0.887
10 epochs	e-optha EX	63.06%	32.26%	95.45%	0.717

Table 2. Performance of the RF for detecting bright lesions on the DIARETDB1 test set when using the designated split to train and test samples provided with the DIARETDB1 data set.

Model	Performance metric			
	Precision	Recall	Specificity	AUC
50 epochs	76.31%	19.73%	99.27%	0.832
0 epochs	70.27%	53.06%	97.31%	0.838
10 epochs	71.58%	46.26%	97.80%	0.860

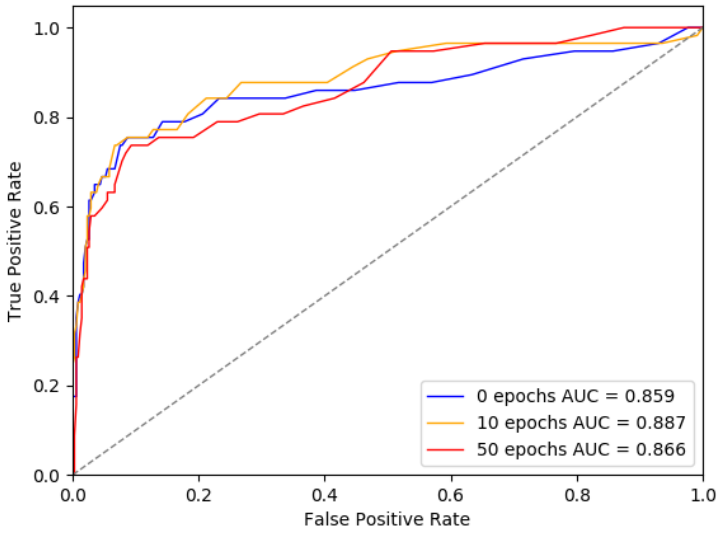

Figure 7. ROC curve of the classifying performance on the test set of 401 subimages from DIARETDB1.

The highest highest recall of 56.14% and highest AUC of 0.887 in Table 1 were achieved on the test set of 401 subimages from DIARETDB1 when using the untrained autoencoder model as a feature extractor. The maximum precision of 83.33% and maximum specificity of 99.13% were achieved on the same test data using the fully trained autoencoder model. The results in Table 2 show a similar pattern with the best results based on recall and AUC achieved when the autoencoder was trained for fewer epochs and the fully trained model performing best in precision and specificity of the used metrics.

An observation worth noting is that precision seems to increase when the autoencoder model used as a feature extractor is trained for longer, whereas recall decreases drastically. The dramatic decrease in recall could be due to the autoencoder primarily learning to represent the veins present in the images when trained for longer, which seems to imply that the lesions associated with the diseases of the fundus become underrepresented in the feature vectors extracted by the model as suggested in Section 4.1.

A reason for the autoencoder not learning to represent the lesions associated with DR could be the observed lack of subimages containing lesions in the set of subimages used for training the autoencoder model. Similarly, most subimages extracted from the DIARETDB1 data set, which were used in the classification setup, contained no signs of lesions associated with DR although the majority of fundus images in DIARETDB1 originally contained signs of at least early DR.

Overall, the suggested classification framework performs poorly when tested using data not seen by either the autoencoder or random forest classifier. The results suggest that an autoencoder based purely on unsupervised learning does not extract a more discriminative set of features for the detection of bright lesions when trained for longer on the subimages extracted from the Messidor data set. A factor contributing to the poor results is most likely the lack of positive examples in the training data used.

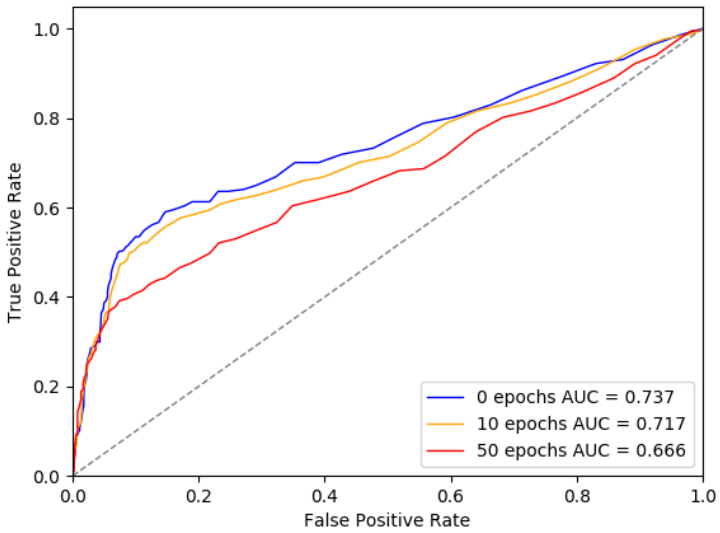

Figure 8. ROC curve of the classifying performance on the 1118 subimages from e-optha EX.

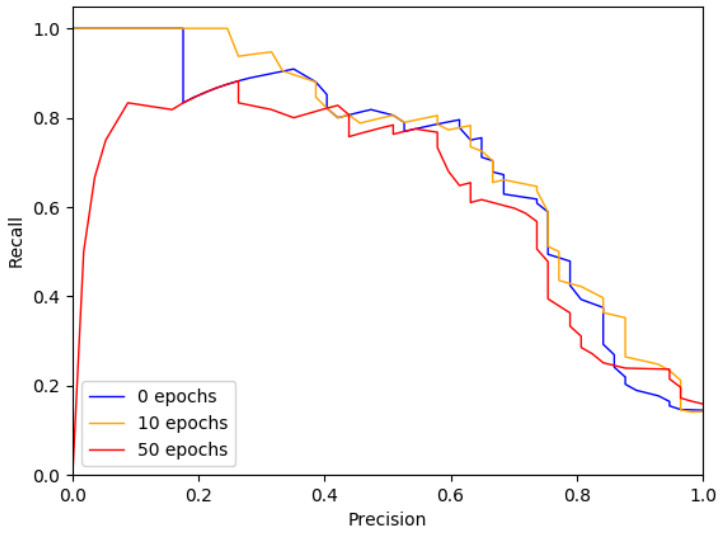

Figure 9. Precision-recall curve illustrating the performance of the classifier on the test set of 401 subimages from DIARETDB1.

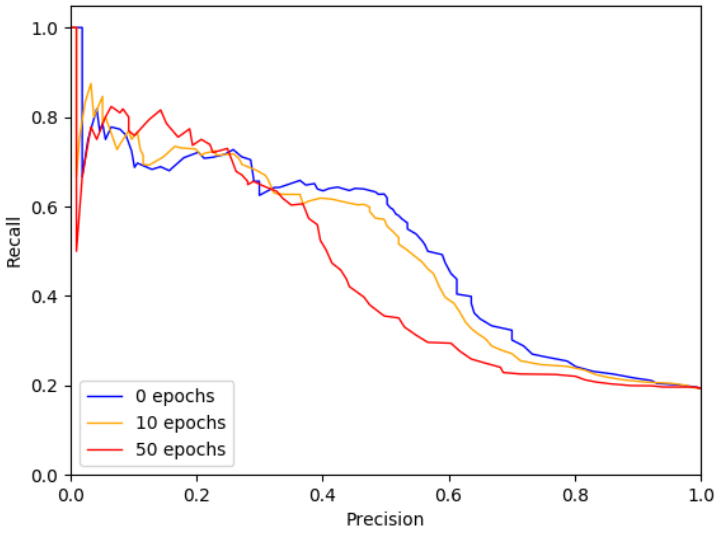

Figure 10. Precision-recall curve illustrating the performance of the RF classifier for detecting exudates on the e-optha EX data set.

5. CONCLUSIONS

The purpose of this thesis was to describe existing solutions for automated fundus image analysis for detecting pathological signs and present a new approach to the problem based on feature extraction using an autoencoder. The results presented in Chapter 4 indicate that the representation that the autoencoder model extracts from the subimages extracted from fundus photographs retains relevant information for the purpose of detecting bright lesions. The visualization results achieved through dimension reduction using UMAP demonstrate the power of unsupervised dimension reduction techniques in data analysis, and could motivate further adoption of visualization techniques in medical image analysis.

However, training the autoencoder for longer did not lead to a more discriminative feature representation for the purpose of detecting bright lesions due to the autoencoder primarily learning to represent the blood vessels present in the training data with continued training. This is partly due to the lack of lesions present in the used training data, which led the autoencoder to learn to represent the blood vessels which appear in both healthy and diseased subimages instead.

Although the autoencoder model in the described framework did not achieve the desired result of learning to extract a more useful representation after training on relevant image data, the proposed approach has much potential for improvement. The training process of the model could be refined to ignore the reconstruction of the blood vessels, leading to a semi-supervised approach to the problem. This would require accurate segmentation of the blood vessels in the eye, which is achievable using specialized neural network models developed for segmentation purposes, such as mask-RCNN models.

One reason for the autoencoder model not learning to represent the lesions associated with diseases of the fundus is the lack of positive examples in the data it was trained on. The subimage extraction procedure could be refined or data augmentation could be applied to increase the number of diseased samples in an attempt to achieve a more balanced set of subimages for training the model.

In addition, further preprocessing is necessary to allow the autoencoder to learn a more meaningful means of representation from its training data. Although multiple preprocessing steps were taken to extract the subimages from the fundus image data, the subimages themselves were only preprocessed using very simple normalization. A number of preprocessing methods have been proposed in literature for the purpose of increasing the contrast of between the lesions associated with diseases of the fundus and the background of the fundus image. Applying such methods in conjunction with detailed annotations of interfering structures, such as the optic disk or blood vessels, could certainly allow for the development of a more useful feature extractor for this purpose.

BIBLIOGRAPHY

Antal, B., & Hajdu, A. (2012, June). An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. *IEEE Transactions on Biomedical Engineering*, *59*(6), 1720-1726. doi: 10.1109/TBME.2012.2193126

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cho, N., Shaw, J., Karuranga, S., Huang, Y., da Rocha Fernandes, J., Ohlrogge, A., & Malanda, B. (2018). Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, *138*, 271 - 281. Lainattu saatavilla http://www.sciencedirect.com/science/article/pii/S0168822718302031 doi: https://doi.org/10.1016/j.diabres.2018.02.023

Ciulla, T. A., Amador, A. G., & Zinman, B. (2003). Diabetic retinopathy and diabetic macular edema. *Diabetes Care*, *26*(9), 2653–2664. Lainattu saatavilla https://care.diabetesjournals.org/content/26/9/2653 doi: 10.2337/diacare.26.9.2653

Decencière, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.-C., Meyer, F., ... Chabouis, A. (2013). Teleophta: Machine learning and image processing methods for teleophthalmology. *IRBM*, *34*(2), 196 - 203. Lainattu saatavilla http://www.sciencedirect.com/science/article/pii/S1959031813000237 (Special issue: ANR TECSAN: Technologies for Health and Autonomy) doi: https://doi.org/10.1016/j.irbm.2013.01.010

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., ... Klein, J.-C. (2014, elokuuta). Feedback on a publicly distributed database: the messidor database. *Image Analysis & Stereology*, *33*(3), 231–234. Lainattu saatavilla http://www.ias-iss.org/ojs/IAS/article/view/1155 doi: 10.5566/ias.1155

Gargeya, R., & Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, *124*(7), 962-969. Lainattu saatavilla https://www2.scopus.com/inward/record.uri?eid=2-s2.0-85016221341&doi=10.1016%2fj.ophtha.2017.02.008&partnerID=40&md5=07287985cdbfc42c76a8b6ba19d0594e doi: 10.1016/j.ophtha.2017.02.008

Giancardo, L., Meriaudeau, F., Karnowski, T. P., Li, Y., Garg, S., Tobin, K. W., & Chaum, E. (2012). Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Medical image analysis*, *16*(1), 216-226.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - Journal of the American Medical Association*, *316*(22), 2402-2410.

Kauppi, T., Kalesnykiene, V., Kamarainen, J. ., Lensu, L., Sorri, I., Raninen, A., ... Uusitalo, H. (2007). The diaretdb1 diabetic retinopathy database and evaluation protocol. Teoksessa *Bmvc 2007 - proceedings of the british machine vision conference 2007*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60 - 88. Lainattu saatavilla http://www.sciencedirect.com/science/article/pii/S1361841517301135 doi: https://doi.org/10.1016/j.media.2017.07.005

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Niemeijer, M., van Ginneken, B., Staal, J., Suttorp-Schulten, M. S. A., & Abramoff, M. D. (2005, May). Automatic detection of red lesions in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, *24*(5), 584-592. doi: 10.1109/TMI.2005.843738

Niemeijer, M., Van Ginneken, B., Cree, M. J., Mizutani, A., Quellec, G., Sánchez, C. I., … others (2009). Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE transactions on medical imaging*, *29*(1), 185–195.

Niemeijer, M., van Ginneken, B., Russell, S. R., Suttorp-Schulten, M. S., & Abramoff, M. D. (2007). Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Investigative ophthalmology & visual science*, *48*(5), 2260–2267.

Orlando, J. I., Prokofyeva, E., del Fresno, M., & Blaschko, M. B. (2018). An ensemble deep learning based approach for red lesion detection in fundus images. *Computer Methods and Programs in Biomedicine*, *153*, 115 - 127. Lainattu saatavilla http://www.sciencedirect.com/science/article/pii/S0169260717307897 doi: https://doi.org/10.1016/j.cmpb.2017.10.017

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, *3*(3), 25.

Quellec, G., Lamard, M., Josselin, P. M., Cazuguel, G., Cochener, B., & Roux, C. (2008, Sep.). Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Transactions on Medical Imaging*, *27*(9), 1230-1241. doi: 10.1109/TMI.2008.920619

Schmidt, D. (2008). The mystery of cotton-wool spots-a review of recent and historical descriptions. *European journal of medical research*, *13*(6), 231.

Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., & van Ginneken, B. (2004). Ridge based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, *23*(4), 501-509.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Teoksessa *Proceedings of the ieee conference on computer vision and pattern recognition* (s. 2818–2826).

ultralytics/yolov5: v3.0. (s.a.). Lainattu 2020-09-25, saatavilla https://zenodo.org/record/3983579 doi: 10.5281/zenodo.3983579

Wilkinson, C., Ferris, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., … Verdaguer, J. T. (2003). Proposed international clinical diabetic retinopathy

and diabetic macular edema disease severity scales. *Ophthalmology*, *110*(9), 1677 - 1682. Lainattu saatavilla http://www.sciencedirect.com/science/article/pii/S0161642003004755 doi: https://doi.org/10.1016/S0161-6420(03)00475-5