Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs

Mike Voets[1], Kajsa Møllersen[2], Lars Ailo Bongo[1]
[1]Department of Computer Science
[2]Department of Community Medicine
mvo010@post.uit.no,{kajsa.mollersen,lars.ailo.bongo}@uit.no
UiT - The Arctic University of Norway

Abstract

We have replicated some experiments in *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs* that was published in JAMA 2016; 316(22)[1]. To our knowledge this study is the first attempt to reproduce their results in this highly cited paper. We re-implemented the methods since the source code is not available.

The original study used fundus images from EyePACS and three hospitals in India for training their detection algorithm. We used a different EyePACS data set that was made available in a Kaggle competition. For evaluating the algorithm's performance the benchmark data set Messidor-2 was used. We used the similar Messidor-Original data set to evaluate our algorithm's performance. In the original study licensed ophthalmologists re-graded all their obtained images for diabetic retinopathy, macular edema, and image gradability. Our challenge was to re-implement the methods with publicly available data sets and one diabetic retinopathy grade per image, find the hyper-parameter settings for training and validation that were not described in the original study, and make an assessment on the impact of training with ungradable images.

For this study we attempted to replicate experiments from the original study using various settings of hyper-parameters, but we were not able to reproduce the performance as reported in the original study. Our algorithm's area under the receiver operating of 0.74 on the Kaggle EyePACS test set and 0.59 on Messidor-Original did not even come close to the reported area under the receiver operating curve of 0.991 on the EyePACS set and of 0.990 on Messidor-2 in the original study. We believe our model did not learn to recognize lesions in fundus images, since we only had a singular grade for diabetic retinopathy per image, instead of multiple grades per images. Furthermore, the original study missed details regarding hyper-parameter settings for training and validation. The orig-

inal study may also have used image quality grades as input for training the network.

We believe that deep learning algorithms should be easily replicated, and that ideally source code should be published so that other researchers can confirm the results of the experiments. Our source code and instructions for running the replication are available at:
https://github.com/mikevoets/jama16-retina-replication.

1 Introduction

In the last decade, deep learning has become one of the most promising machine learning methods due to the increasing affordability of graphics processing units (GPUs) that accelerate the computation for finding patterns in large datasets. This has consequently led to an increase of published articles demonstrating the feasibility of applying deep learning in practice, particularly in fields where images are classified [14][15]. However, to prove that proposed methods are general enough to apply in practice, the methods should be evaluated on many different, preferably public, data sets, with either published source code or a replication of the methods. [19][20] show that replication is challenging in different scientific fields. In this study we make an assessment on the reproducibility of deep learning methods.

We have replicated some methods from *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs*, published in JAMA 2016; 316(22)[1]. In February 2018, this article had been cited 249 times [2][3]. This is to our knowledge the first attempt to replicate their results. The paper describes an algorithm for detection of diabetic retinopathy in retinal fundus photographs. The algorithm is trained using 128 725 fundus images retrieved from EyePACS and from three eye hospitals in India, and the algorithm's performance was evaluated on 2 test sets. The algorithm's area under the receiver operating curve for detecting referable diabetic retinopathy was 0.991 for EyePACS-1 and 0.990 for Messidor-2, and two operating points were selected for high sensitivity and specificity. The operating point for high specificity had 90.3% and 87.0% sensitivity and 98.1% and 98.5% specificity for the EyePACS-1 and Messidor-2 test sets, whereas the operating point for high sensitivity had 97.5% and 96.1% sensitivity and 93.4% and 93.9% specificity, respectively. All fundus images were re-graded by a team of 54 US-licensed ophthalmologists. The source code is not published.

To re-implement their algorithm for detection of diabetic retinopathy, we used similar images from a publicly available EyePACS data set for training, and we used a subset from the EyePACS data set and images from the public Messidor-Original data set for testing. In addition, we attempted to find optimal hyper-parameters for the algorithm's training and validation procedure that the paper did not describe. Our objective is to compare the results of their referable diabetic retinopathy detection algorithm, and our replica algorithm, taking into account potential deviations in the data sets, having fewer grades, and potential

differences in hyper-parameter settings.

In comparison, our algorithm's area under the receiver operating curve for detecting referable diabetic retinopathy for our EyePACS and Messidor-Original test sets were 0.74 and 0.59. The operating point for high specificity (0.2965) had 67.2% and 44.0% sensitivity and 68.2% and 64.8% specificity for our EyePACS and Messidor-Original test sets, and the operating point for high sensitivity (0.2513) had 79.8% and 56.6% sensitivity and 53.7% and 54.3% specificity. The results differ for three reasons. First, we used publicly available grades, whereas in the article the grades have been re-graded multiple times. The study states that re-grading the data set positively contributes to the performance, and that the performance of the algorithm steadily increases as more grades are used. Second, the validation procedure is different due to missing details on the validation procedure. Third, the original study might miss details on the exact used neural network.

We believe our replication results motivate the need for replication studies in deep learning, and that the results give a general insight into the reproducibility of studies that did not use publicly available data. We also believe publishing source code should be a standard procedure, so that other researchers are able to verify the described results. Therefore, we published our source code and give others the opportunity to do further research in addition to this study.

2 Methods

2.1 Data sets

For the development of the algorithm, often called the training procedure, the original study used a training data set (called development set in the paper) with 128 725 macula-centered retinal fundus images obtained from EyePACS in the US, and images from three eye hospitals in India. We used 88 702 images from EyePACS retrieved via a Kaggle competition for diabetic retinopathy detection launched and finished in 2016 [4]. The original study used two additional data sets to evaluate the developed algorithm, and it refers to these data sets by validation data sets. For evaluating an algorithm's performance, the term test data set is more commonly used. The first validation (test) data set was a randomly distributed set of 9963 images retrieved at EyePACS screening sites between May 2015 and October 2015. These images were taken during diabetic retinopathy screening by a variety of cameras, all using 45 degrees fields of view. Approximately 40% of these images were acquired with pupil dilation. This data set did not overlap the EyePACS data used in the training set. The second test set was the publicly available Messidor-2 data set [5][6], consisting of 1748 images. This data set was acquired between January 2005 and December 2010 at three hospitals in France by using a Topcon TRC NW6 non-mydriatic camera and 45 degrees fields of view centered on the fovea. Of these images, approximately 44% were acquired with pupil dilation.

We obtained images for training and testing from two sources: EyePACS

Figure 1: Screenshot of grading tool used to assess gradability for all images.

from Kaggle and the publicly available Messidor-Original set. The Messidor-Original set is a benchmark for algorithms that detect diabetic retinopathy. We separated the Kaggle EyePACS data set consisting of 88 702 images into a training set of 57 146 images and a test set of 8790 images. The leftover images were mostly images graded as having no diabetic retinopathy and were not used for training the algorithm. The reason for the number of images in our training set is to keep the same balance between grades as in the original study's training set. Our EyePACS test set has an identical amount of images and balance between grades as in the original study's EyePACS test set. We used all the available 1200 images from Messidor-Original for testing. We removed duplicate images and made corrections from this set as suggested on the Messidor-Original download page, resulting in a test set of 1187 images. Note that we could not use Messidor-2 since they do not contain grades for diabetic retinopathy. Messidor-Original is a subset of Messidor-2, which means that these data sets are quite similar.

2.2 Grading

The images used for the algorithm training and testing in the original study were all graded by ophthalmologists for image quality (gradability), the presence of diabetic retinopathy, and macular edema. We did not have grades for macular edema for all our images, so we did not train our algorithm to detect macular edema.

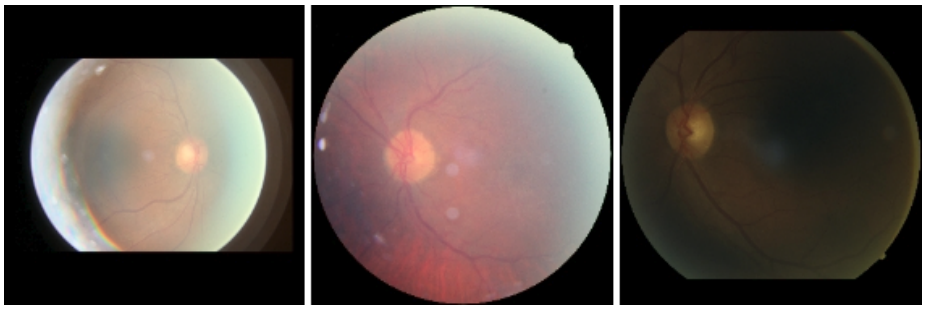

Figure 2: Examples of ungradable images because they are either out of focus, under-, or overexposed.

Kaggle [17] describes that some of the images or grades in their EyePACS distribution may consist of noise, contain artifacts, be out of focus, or be over- or underexposed. Further [18] states that 75% of the EyePACS images via Kaggle are estimated gradable. For this study one of the authors (MV) graded all Kaggle and Messidor-Original images on their image quality with a simple grading tool (Figure 1). MV is not a licensed ophthalmologist, but we assume fundus image quality can be reliably graded by non-experts. We used the "Grading Instructions" in the Supplement of the original study to assess image quality. We publish the image quality grades with the source code. Images of at least adequate quality were considered gradable.

In the original study, diabetic retinopathy was graded according to the International Clinical Diabetic Retinopathy scale [7], with no, mild, moderate, severe or proliferative severity. 54 US-licensed ophthalmologists or final-year postgraduate ophthalmology trainees were paid to re-grade the training and test sets. Each physician graded between 20 and 62508 images, and 3 to 7 grades were obtained for each image. Intergrader reliability for each individual was measured using pairwise comparison by randomly selecting approximately 10% of the images in the training set to be graded by the same individual two times.

A clinician had graded the images in our Kaggle EyePACS set for the presence of diabetic retinopathy using the same international scaling standard as used in the original study. We have thus only one diagnosis grade for each image. The Messidor-Original test set was graded by medical experts for both the presence of diabetic retinopathy, and for the risk of macular edema. Since we do not have grades for the risk of macular edema in our training set, we did not use these grades in our algorithm. In Messidor-Original, diabetic retinopathy was also graded using a different scale, so we had to convert the grades to the International Clinical Diabetic Retinopathy scale by using the scale's definitions [7]. Fundus images with one to five microaneurysms and no hemorrhages were considered mild; 6 to 14 microaneurysms or up to 5 hemorrhages and no neovascularization were considered moderate; and more than 15 microaneurysms,

Kaggle EyePACS grading (International Clinical Diabetic Retinopathy scale)	**Messidor-Original grading**	**Referable diabetic retinopathy grade**
No diabetic retinopathy	Normal: no microaneurysms and no hemorrhages	0
Mild diabetic retinopathy	1 to 5 microaneurysms and no hemorrhages	0
Moderate diabetic retinopathy	6 to 14 microaneurysms, or up to 5 hemorrhages and no neovascularization	1
Severe diabetic retinopathy	More than 15 microaneurysms, more than 5 hemorrhages, or neovascularization	1
Proliferative diabetic retinopathy	-	1

Table 1: Interpretation of diabetic retinopathy grades from Kaggle EyePACS and Messidor-Original.

more than 5 hemorrhages, or the presence of neovascularization were considered severe or worse diabetic retinopathy. See Table 1 for an overview. As in the original study, we converted the final diabetic retinopathy grade to a binary grade indicating referable diabetic retinopathy, which presents moderate or worse diabetic retinopathy.

2.3 Algorithm development

The objective of this study is to replicate the reference study as accurately as possible. As in the original study, our algorithm is created through deep learning, which involves a procedure of training a neural network to perform the task of grading images. We trained the algorithm with the same neural network architecture as in the original study: the Inception-V3 model proposed by Szegedy et al [8]. This neural network consists of a range of convolutional layers that transforms pixel intensities to local features before converting them into global features.

The fundus images from both training and test sets were preprocessed as described by the original study's protocol for preprocessing. In all images the center and radius of the each fundus were located and resized such that each image gets a height and width of 299 pixels, with the fundus center in the middle of the image. We also scale-normalized the images before passing them to the neural network, as in the original study.

The original study used distributed stochastic gradient descent proposed by Dean et al [9] as the optimization function for training the parameters (i.e. weights) of the neural network. This implies that their neural network was

trained in parallel, although the paper does not describe it. We did not conduct any distributed training for our replica neural network. Therefore, we used the "normal" stochastic gradient descent as our optimization procedure. Using a different optimization procedure should not affect the final performance of the algorithm. The original study did not describe any learning rate for their training. Therefore we had to experiment with several settings for the learning rate.

As in the original study, we improved performance in the training procedure of our neural network by using batch normalization layers [10] after each convolutional layer. Our weights were also pre-initialized using weights from the neural network trained to predict objects in the ImageNet dataset [11].

The neural network of the original study was trained to output multiple binary predictions: 1) whether the image was graded moderate or worse diabetic retinopathy (i.e. moderate, severe, or proliferative grades); 2) severe or worse diabetic retinopathy; 3) referable diabetic macular edema; or 4) fully gradable. The term referable diabetic retinopathy was defined in the original study as an image having either or both grade 1 and 3. For the training data obtained in this replication study, only grades for diabetic retinopathy were present. That means that our neural network outputs only one binary prediction: moderate or worse diabetic retinopathy (referable diabetic retinopathy).

In this study, the development (training) part of the data set was splitted into two parts as in the original study: 80% of the training set was used for training, and 20% was used for validating the neural network.

It is estimated that the Kaggle EyePACS set contains a substantial amount of ungradable images. We also trained an algorithm with only gradable images. In the original study, the performance of an algorithm trained with only gradable images was also summarized. We do not use the image quality grades for algorithm training.

To compare with other optimization procedures that may improve the speed of training, we re-trained all networks by using Nesterov's accelerated gradient descent.

2.4 Performance evaluation

We measured the performance of the resulting neural network by the area under the receiver operating curve (AUC) on a validation set, as in the original study. We find the area by thresholding the network's output predictions, which are continuous numbers ranging from 0 to 1. By moving the operating threshold on the predictions, we obtain different results for sensitivity and specificity. We then plot sensitivity against 1 – specificity for 200 thresholds. Finally, the AUC of the validation set is calculated, and becomes an indicator for how well the neural network detects referable diabetic retinopathy. The original study did not describe how many thresholds were used for plotting AUC, so we used the de facto standard of 200 thresholds.

The original paper describes that the AUC value of the validation set was used for the early-stopping criterion [12]; training is terminated when a peak

AUC on the validation set is reached. This prevents overfitting the neural network on the training set. In our validation procedure, we also use the AUC calculated from the validation set as an early stopping criterion. To determine if a peak AUC is reached, we have to compare the AUC values between different validation checkpoints. To avoid stopping at a local minimum of the validation AUC function, our network may continue to perform training up to n epochs (i.e. patience of n epochs). The original paper did not describe details regarding the validation procedure, so we assumed validation was performed after each epoch, and we had to experiment with several settings for patience. One epoch of training is equal to running all images through the network once.

For further improving the algorithm's performance, we used ensemble learning [13] by training 10 networks on the same data set, and using the final prediction computed by a linear average of the predictions of the ensemble. This was also done in the original study.

In the original study, additional experiments were conducted to evaluate the performance of the resulting algorithm based on the training set, compared with performance based on subsets of images and grades from the training set. We did not replicate these experiments for two reasons. First, we chose to focus on replicating the main results of the original paper. That is, the results of an algorithm detecting referable diabetic retinopathy. Second, we cannot perform subsampling of grades, as we only have one grade per image.

3 Results

The Kaggle EyePACS set consists of 88 702 images. We had to exclude 15 images from this set because they failed the preprocessing procedure, because the images were either corrupted or their circular mask could not be detected due to bad image quality. We assessed the image quality for all images, and we found that 71 056 images (80.1%) of the Kaggle EyePACS images are gradable. We assigned 57 146 images (16458 [28.8%] with referable diabetic retinopathy, 45615 [79.8%] gradable) for training and validation (8:2 split), and 8790 images (694 [7.89%] referable, 7066 [80.4%] gradable) for testing. Note that the proportion of referable and non-referable images is the same as in the original study. The Messidor-Original test set consists of 1187 images, of which 491 (41.3%) images have referable diabetic retinopathy, and we found 1180 images (99.4%) gradable.

For detecting referable diabetic retinopathy, our algorithm that was trained with all images had an AUC of 0.71 on our Kaggle EyePACS test set, and an AUC of 0.60 on Messidor-Original. With the operating cut point for high specificity (0.3467), the algorithm's sensitivity was 65.7% and specificity was 67.6% on our Kaggle EyePACS test set. In Messidor-Original, the sensitivity was 42.2% and specificity was 68.8%. Using the operating cut point for high sensitivity (0.2915), the algorithm had a sensitivity of 75.4% and specificity of 55.4% on our Kaggle EyePACS test set, and a sensitivity of 57.6% and specificity of 54.6% on Messidor-Original. Note that these results do not compare with the original paper main results (detection of referable diabetic retinopathy), but

Operating threshold	**0.2915**	**0.3467**	**AUC score**
All Kaggle EyePACS test images	75.4% sensitivity 55.4% specificity	65.7% sensitivity 67.6% specificity	0.71
All Messidor-Original images	57.6% sensitivity 54.6% specificity	42.2% sensitivity 68.8% specificity	0.60
Operating threshold	**0.2513**	**0.2965**	**AUC score**
Only gradable Kaggle EyePACS test images	79.8% sensitivity 53.7% specificity	67.2% sensitivity 68.2% specificity	0.74
Only gradable Messidor-Original images	56.6% sensitivity 54.3% specificity	44.0% sensitivity 64.8% specificity	0.59

Table 2: Performance on test sets of ensemble models trained with stochastic gradient descent.

they should be compared with the results for the algorithm that detects moderate or worse diabetic retinopathy (see results for individual subtypes). This is because our algorithm only predicts moderate or worse diabetic retinopathy, and not macular edema, which is also considered referable diabetic retinopathy in the original study. To compare, the original study's results at the operating cut point for high specificity for their EyePACS-1 test set were a sensitivity of 90.1% and specificity of 98.2%, and a sensitivity of 86.6% and specificity of 98.4% on Messidor-2. The original study does not report results at the operating cut point for high sensitivity.

We trained an algorithm with only gradable images to compare potential differences in performance between the two algorithms. We redistributed our Kaggle EyePACS data sets after excluding ungradable images, and trained this new algorithm with 43 688 images (12582 [28.8%] referable) as our Kaggle EyePACS training and validation set, and used 8790 images (694 [7.89%] referable) as our Kaggle EyePACS test set. Our algorithm's performance in making predictions for referable diabetic retinopathy for only gradable images, had an AUC of 0.74 on our Kaggle EyePACS test set, and an AUC on 0.59 on Messidor-Original. The operating cut point for high specificity (0.2965) had a sensitivity of 67.2% and specificity of 68.2% on our Kaggle EyePACS test set, and a sensitivity of 44.0% and specificity of 64.8% on Messidor-Original. The operating cut point for high sensitivity (0.2513) had a sensitivity of 79.8% and specificity of 53.7% on our Kaggle EyePACS test set, and a sensitivity of 56.6% and specificity of 54.3% on Messidor-Original. In comparison, the original paper reported an AUC of 0.991 on their EyePACS-1 set and an AUC of 0.990 on Messidor-2. Their operating cut point for high specificity had a sensitivity of 90.3% and specificity

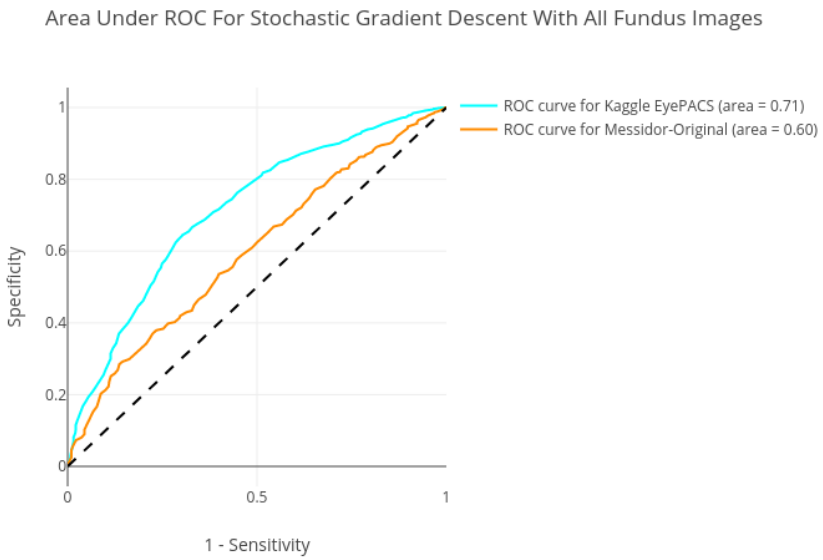

Figure 3: Area under receiver operating curve for training with all fundus images and stochastic gradient descent.

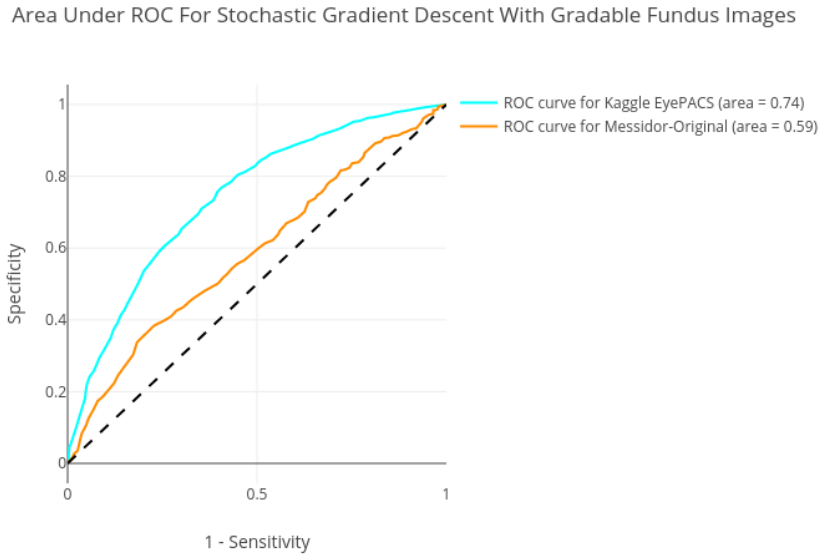

Figure 4: Area under receiver operating curve for training with only gradable fundus images and stochastic gradient descent.

Operating threshold	**0.2864**	**0.3467**	**AUC score**
All Kaggle EyePACS test images	78.4% sensitivity 57.6% specificity	67.3% sensitivity 69.6% specificity	0.75
All Messidor-Original images	56.0% sensitivity 53.0% specificity	38.9% sensitivity 67.2% specificity	0.55
Operating threshold	**0.2563**	**0.3216**	**AUC score**
Only gradable Kaggle EyePACS test images	82.4% sensitivity 51.5% specificity	68.3% sensitivity 70.1% specificity	0.75
Only gradable Messidor-Original images	59.7% sensitivity 57.8% specificity	39.5% sensitivity 74.6% specificity	0.61

Table 3: Performance on test sets of ensemble models trained with Nesterov's accelerated descent.

of 98.1% on EyePACS-1, and a sensitivity of 87.0% and specificity of 98.5% on Messidor-2. Their operating cut for high sensitivity had a sensitivity of 97.5% and specificity of 93.4% on EyePacs-1, and a sensitivity of 96.1% and specificity of 93.9% on Messidor-2. Note that referable diabetic retinopathy only presents moderate or worse diabetic retinopathy in our study, however in the original paper referable macular edema is included in this scope as well.

To compare other optimization procedures during training, we also re-trained algorithms with Nesterov's accelerated gradient descent instead of stochastic gradient descent, and we summarized the performance results of these algorithms in Table 3.

Hyper-parameters settings for the optimization and validation procedure were not specified, so we conducted experiments to find hyper-parameters settings that worked well for training and validating the algorithms. We found that a static learning rate of 0.003 performed well. For Nesterov's accelerated gradient descent we used a momentum value of 0.9. As for our early-stopping criterion at a peak AUC, we introduced a patience of 10 epochs. Our chosen requirement for a new peak AUC was a value of AUC that is larger than the previous peak value, with a minimum difference of 0.01.

4 Discussion

The results show substantial performance differences between the original study's algorithm and our replica algorithm. Even though we followed the methodology of the original study as closely as possible, our algorithm did not seem to "learn" how to recognize lesions in fundus images as local features. This is probably

because our algorithms were trained under different hyper-parameters, and because in the original study ophthalmologic experts re-graded all their images. According to the original study the validation and test sets should have multiple grades per image, because it will provide a more reliable measure of a model's final predictive ability. Their results on experimenting with only one grade per image show that their algorithm's performance declines with 36%.

The methods in the original study do not specify some details. First, the details on hyper-parameter settings for the validation procedure, or for the optimization function are missing. The original study also briefly mentions that image preprocessing is performed in the validation procedure, but it does not further elaborate on this. Second, it is unclear how the algorithm's predictions for diabetic retinopathy or macular edema are interpreted in case of ungradable images. The image quality grades might have been used as an input for the network, or the network might be concatenated with another network that takes the image quality as an input. Third, apart from the "main" algorithm that detects referable diabetic retinopathy and outputs 4 binary classifications, other algorithms seem to have been trained as well. An example is the described algorithm that only detects referable diabetic retinopathy for gradable images, and an algorithm that detects "all-cause" referable diabetic retinopathy, which presents moderate or worse diabetic retinopathy, referable macular edema, and ungradable images. Details on how these other algorithms are built are however not reported. It is unclear whether the main network has been used or if the original study trained new networks. Lastly, the original paper did not state how many iterations it took for their proposed model to converge during training, or describe how to find a converging model.

The main challenge in this replication study was to find hyper-parameters, which were not specified in the original paper, such that the algorithm does not converge on a local maximum of the validation AUC function. To understand how we should adjust the hyper-parameters, we measured the Brier score on the training set and the AUC value on the validation set after each epoch of training. We observed the following. First, during the first 15 epochs, the AUC value on the validation set increases and stabilizes at approximately 0.65. From then, the validation AUC does not increase, but stays around the same value. The Brier score measured on the training set gradually decreases, indicating that the algorithm is learning features from the images in the training set. This scenario continues for many epochs: the validation AUC stays around 0.65, with the Brier score of the training set gradually decreasing for every epoch. After around 50 epochs, the validation AUC decreases again, and the algorithm clearly overfits on the training data. One possible reason for the algorithm to not converge may be the dimensions of the fundus images. As the original study suggests, the original fundus images were preprocessed and scaled down to a width and height of 299 pixels. To be able to initialize the InceptionV3 network with ImageNet pre-trained weights, which have been trained with images of 299 by 299 pixels. We believe it is difficult for ophthalmologists to find lesions in fundus images of this size, so we assume the algorithm has difficulties with detecting lesions as well. [18] also points out this fact, and suggests re-training

an entire network with larger fundus images and randomly initialized weights instead. And as mentioned before, it seems like the original study extended the InceptionV3 model architecture for their algorithm to use image gradability as an input parameter.

A potential drawback with the Kaggle data is that it contains grades for diabetic retinopathy for all images. We found that 80.1% of these images are gradable, and it is thus possible that the algorithm will "learn" features for ungradable images, and make predictions based on anomalies. This is likely to negatively contribute to the algorithm's predictive performance, but we were not able to show a significant difference of performance between an algorithm trained on all images and an algorithm trained on only gradable images. Further research is needed to determine the negative effects of training with ungradable images.

We also trained the algorithms with accelerated gradient descent, instead of the stochastic gradient descent proposed by the original paper. Our observation is that the algorithm's performance was not substantially improved, but that less epochs were needed to find the peak validation AUC value.

Compared to biomedical experiments, we assumed that deep learning experiments are easy to replicate. The original study is a good candidate for replication because it reports results on the publicly available Messidor-2 data, and we could use the Kaggle data to train our model with the same data. However, we could not replicate the results. We believe, ideally source code should be published, so that other researchers can evaluate the experiments with their own data. However, reasons for holding back data, methods or smaller details, are for example that the data or methods contain sensitive information, or that experiments may be commercialized [21][22].

5 Conclusion

We reimplemented some methods from JAMA 2016; 316(22), but we were not able to get the same performance as reported in that study. The findings of this study confirm the need for deep learning replication studies.

The source code of this replication study and instructions for running the replication are available at https://github.com/mikevoets/jama16-retina-replication.

References

[1] Gulshan V, Peng L, et al. *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.* JAMA. 2016;316(22):2402–2410. doi:10.1001/jama.2016.17216.

[2] Google. *Google Scholar.* `https://scholar.google.no/scholar?cites=16083985573643781536`. Accessed February 22, 2018.

[3] Abràmoff MD (2017). *Image Processing.* Ryan's Retina E-Book: Edition 6 (8). Retrieved from https://books.google.co.jp/books?lr=&id=YWWwDgAAQBAJ&oi=fnd&pg=PA196&ots=Uy_u4eneJ3&sig=CJXNSrkLd-Vn7T2_9Frfmt4oR9w&redir_esc=y#v=onepage&q&f=false.

[4] Kaggle. *Diabetic Retinopathy Detection (2015).* https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed March 11, 2018.

[5] Decencière E, Etienne D, Xiwei Z, et al. *Feedback on a publicly distributed image database: the Messidor database.* Image Anal Stereol. 2014;33(3):231-234. doi:10.5566/ias.1155, Google Scholar

[6] Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. *Optimal wavelet transform for the detection of microaneurysms in retina photographs.* IEEE Trans Med Imaging. 2008;27(9):1230-1241.

[7] American Academy of Ophthalmology. *International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table.* http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf. Accessed October 14, 2016.

[8] Szegedy C, Vanhouke V, Ioffe S, Shlens J, Wojna Z. *Rethinking the Inception Architecture for Computer Vision.* December 2015. http://arxiv.org/pdf/1512.00567v3.pdf. Accessed November 14, 2016.

[9] Dean J, Corrado G, Monga R, et al. *Large scale distributed deep networks.* Adv Neural Inf Processing Syst. 2012:1223-1231.

[10] Ioffe S, Szegedy C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* March 2015. http://arxiv.org/pdf/1502.03167v3.pdf. Accessed November 14, 2016.

[11] Russakovsky O, Deng J, Su H, et al. *ImageNet large scale visual recognition challenge.* Int J Comput Vis. 2015;115(3):211-252.

[12] Caruana R, Lawrence S, Giles L. *Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping.* Adv Neural Inf Processing Syst. 2001:402-408. Google Scholar.

[13] Krizhevsky A, Sutskever I, Hinton GE. *ImageNet classification with deep convolutional neural networks.* Adv Neural Inf Processing Syst. 2012:1097-1105.

[14] Litjens G, Kooi T, et al. *A survey on deep learning in medical image analysis.* Medical Image Analysis vol. 42:60-88.

[15] Travers C, Himmelstein DS, et al. *Opportunities And Obstacles For Deep Learning In Biology And Medicine.* bioRxiv 142760; doi: https://doi.org/10.1101/142760

[16] ADCIS. *Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology (MESSIDOR).* `http://www.adcis.net/en/Download-Third-Party/Messidor.html`. Accessed March 11, 2018.

[17] Kaggle. *Database for Kaggle EyePACS fundus images and labels.* `https://www.kaggle.com/c/diabetic-retinopathy-detection/data`. Accessed March 11, 2018.

[18] Rakhlin A (2017). *Diabetic Retinopathy detection through integration of Deep Learning classification framework.* bioRxiv 225508; doi: doi.org/10.1101/225508

[19] *Reproducibility in Cancer Biology: The challenges of replication.* Editorial. 19 January 2017. eLife 6:e23693.

[20] *Announcement: Reducing our irreproducibility.* Editorial. 25 April 2013. Nature 496:398.

[21] *Reality check on reproducibility.* Editorial. May 25, 2016. Nature 533:437.

[22] Hutson M. *Missing data hinder replication of of artificial intelligence studies.* Science. February 15, 2018. `http://www.sciencemag.org/news/2018/02/missing-data-hinder-replication-artificial-intelligence-studies`. Accessed March 11, 2018.