Replication study: Development and validation of a 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
UiT - The Arctic University of Norway
mvo010@post.uit.no,{kajsa.mollersen,lars.ailo.bongo}@uit.no

Abstract

Replication studies are essential for validation of new methods, and are crucial to maintain the high standards of scientific publications, and to use the results in practice.

We have attempted to replicate the main method in *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs* published in JAMA 2016; 316(22)[1]. We re-implemented the method since the source code is not available, and we used publicly available data sets.

The original study used non-public fundus images from EyePACS and three hospitals in India for training. We used a different EyePACS data set from Kaggle. The original study used the benchmark data set Messidor-2 to evaluate the algorithm's performance. We used the similar Messidor-Original data. In the original study, ophthalmologists re-graded all images for diabetic retinopathy, macular edema, and image gradability. There was one diabetic retinopathy grade per image for our data sets, and we assessed image gradability ourselves. Hyper-parameter settings for training and validation were not described in the original study.

We were not able to replicate the original study. Our algorithm's area under the receiver operating curve (AUC) of 0.74 on the Kaggle EyePACS test set and 0.59 on Messidor-Original did not come close to the reported AUC of 0.99 in the original study. This may be caused by the use of a single grade per image, or different hyper-parameter settings. By changing the pre-processing methods, our replica algorithm's AUC increased to 0.94 and 0.82, respectively.

This study shows the challenges of replicating deep learning, and the need for more replication studies to validate deep learning methods, especially for medical image analysis. Our source code and instructions are available at:
https://github.com/mikevoets/jama16-retina-replication

1 Introduction

Being able to replicate a scientific paper by strictly following the described methods is a cornerstone of science. Replicability is essential for the development of medical technologies based on published results. However, there is an emerging concern that many studies are not replicable, raised for bio-medical research [2], computational sciences [3, 4], and recently for machine learning [5].

Deep learning has become a hot topic within machine learning due to its promising performance of finding patterns in large data sets. There are dozens of libraries that make deep learning methods easily available for any developer. This has consequently led to an increase of published articles that demonstrate the feasibility of applying deep learning in practice, particularly for image classification [6, 7]. However, there is an emerging need to show that studies are replicable, and hence be used to develop new medical analysis solutions. Ideally, the used data set and the source code are published, so that other researchers can verify the results by using other data. However, this is not always possible, for example for sensitive data, or for methods with commercial value [4, 5].

In this study, we make an assessment on the replicability of a deep learning method. We have chosen to attempt to replicate the main method 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]. As of April 2018, this article had been cited 350 times [8]. We chose to replicate this study because it is a well-known and high-impact study within the medical field, the source code has not been published, and there are as far as we know not any others who have attempted to replicate this study.

The original study describes an algorithm (hereby referred to as the original algorithm) for detection of referable diabetic retinopathy (rDR) in retinal fundus photographs. The algorithm is trained and validated using 118 419 fundus images retrieved from EyePACS and from three eye hospitals in India. The original algorithm's performance was evaluated on 2 test sets, and achieved an area under the receiver operating curve (AUC) for detecting rDR of 0.99 for both the EyePACS-1 and the Messidor-2 test sets. 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.

To assess replicability of the method used to develop the original algorithm for detection of rDR, we used similar images from a publicly available EyePACS data set for training and validation, and we used a subset from the EyePACS data set and images from the public Messidor-Original data set for performance evaluation. Because many of the details regarding the validation procedure were not described in the original study (for example for hyper-parameter optimization), we had to find optimal hyper-parameters ourselves. Our objective is to compare the performance of the original rDR detection algorithm to our result algorithm after trying to replicate, taking into account potential deviations in the data sets, having fewer grades, and potential differences in hyper-parameter settings.

We were not able to replicate the original study. Our algorithm's AUC for detecting rDR for our EyePACS and Messidor-Original test sets were 0.74 and 0.59. The operating point for high specificity 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 had 79.8% and 56.6% sensitivity and 53.7% and 54.3% specificity. The results can differ for four reasons. First, we used public retinal images with only one grade per image, whereas in the original study the non-public retinal images were re-graded multiple times. Second, the original study lacked details regarding

the training and validation procedure, and the original algorithm may therefore have been tuned better. Third, there might be errors in the original study or methodology. The last possible reason is that we may have done something wrong with replicating the method by having misinterpreted the methodology. We do not know for sure which of the four reasons has led to our considerably worse performance. In further research, apart from this replication study, we improved the algorithm by slightly modifying the pre-processing procedure, and the AUC then increased to 0.94 and 0.82 for the Kaggle EyePACS and the Messidor-Original test sets, respectively.

We believe our failed effort on trying to replicate a highly-cited deep learning paper motivates the need for additional replication studies in deep learning. This result gives a general insight into the challenges of replicating studies that do not use publicly available data and publish source code. We have published our source code with instructions for how to use it with public data. This gives others the opportunity to improve upon the attempted replication.

2 Methods

2.1 Data sets

The data sets consist of images of the retinal fundus acquired for diabetic retinopathy screening. Any other information regarding the patient is not part of the data sets. Each image is graded according to severity of symptoms (see Section 2.2).

The original study obtained 128 175 retinal fundus images from EyePACS in the US and from three eye hospitals in India. 118 419 macula-centered images from this data set were used for algorithm training and validation (referred to as *development set*, divided into *training* and *tuning set* in the original study). To evaluate the performance of the algorithm, the original study used two data sets (referred to as *validation sets* in the original study). For evaluating an algorithm's performance, the term test set is commonly used. The first test set was a randomly sampled set of 9963 images retrieved at EyePACS screening sites between May 2015 and October 2015. The second test set was the publicly available Messidor-2 data set [9, 10], consisting of 1748 images. We provide an overview of the differences in image distribution used in our replication study compared with the original study in Figure 2.

We obtained images for training, validation and testing from two sources: EyePACS from a Kaggle competition [11], and the publicly available Messidor-Original set [12]. The Messidor-Original set is a benchmark for algorithms that detect diabetic retinopathy. We randomly sampled the Kaggle EyePACS data set consisting of 88 702 images into a training and validation 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 and validation set is to keep the same balance for the binary rDR class as in the original study's training and validation set. Our EyePACS test set has an identical amount of images and balance for the binary rDR class 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 official grades for diabetic retinopathy. Messidor-Original is a subset of Messidor-2, which means that these data sets are quite similar.

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

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.

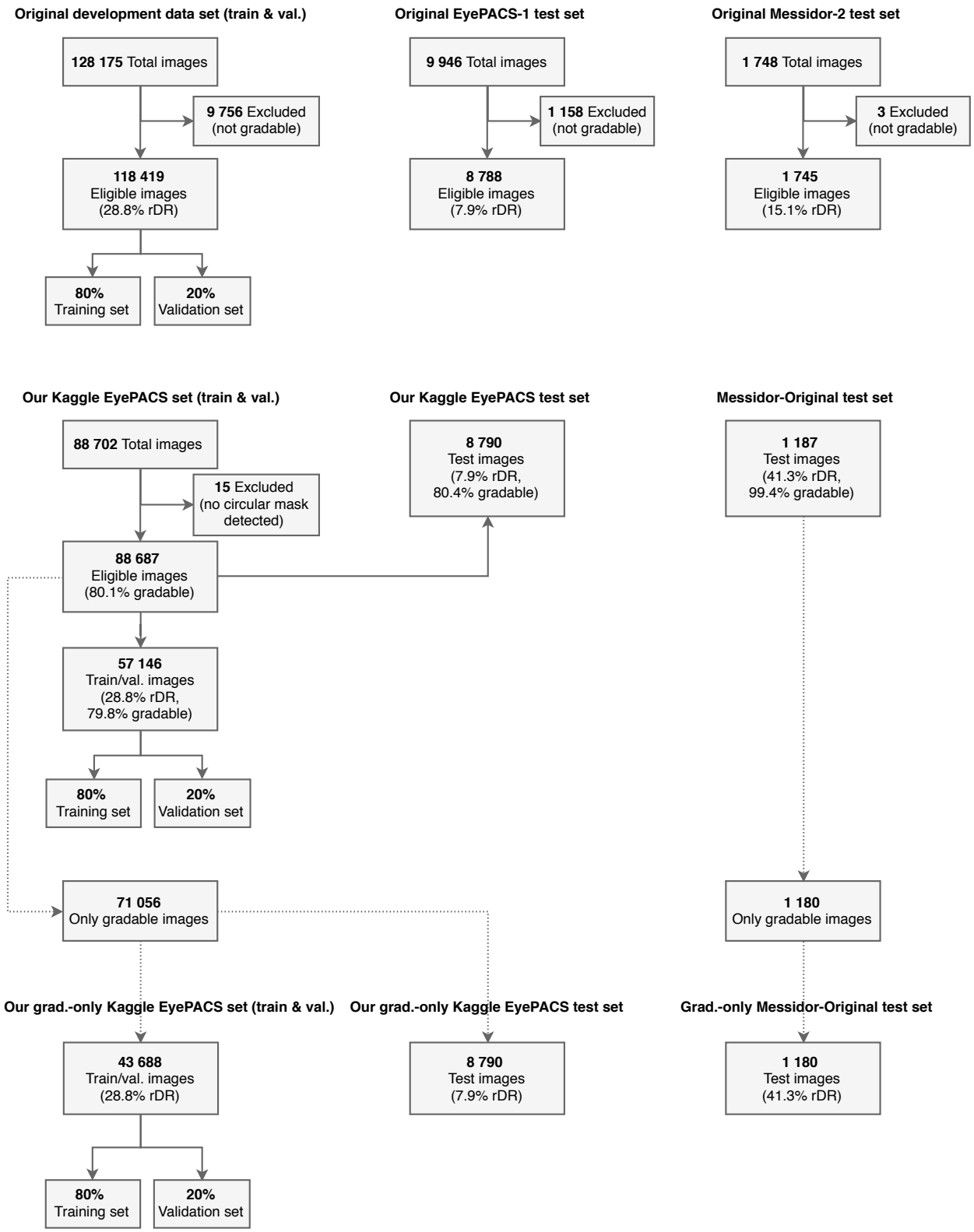

Figure 2: Data set distribution in original study vs. this replication study.

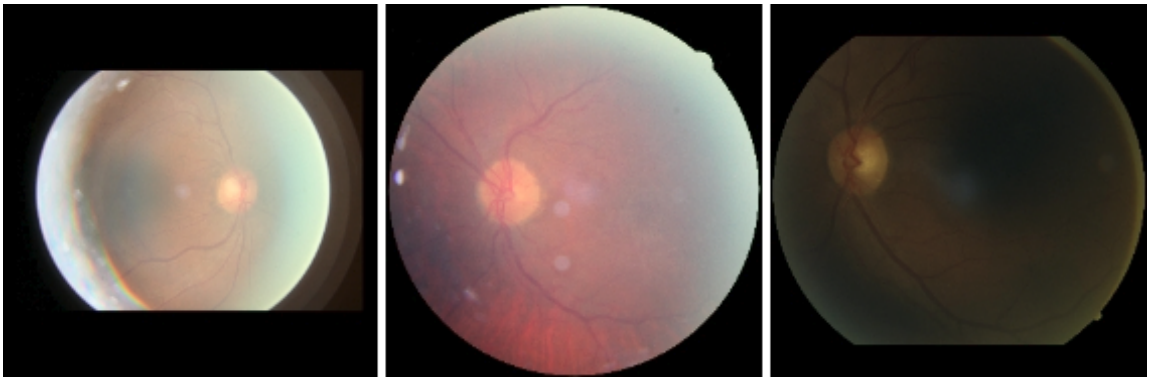

Figure 3: Examples of ungradable images because they are either out of focus, under-, or overexposed.

Kaggle [13] describes that some of the images in their EyePACS distribution may consist of noise, contain artifacts, be out of focus, or be over- or underexposed. [14] states further 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 [15], with no, mild, moderate, severe or proliferative severity.

The Kaggle EyePACS set had been graded by one clinician 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. Kaggle does not give more information about where the data is from. 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 converted the grades to the International Clinical Diabetic Retinopathy scale by using the scale's definitions [15]. 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, 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 training

The objective of this study is to assess replicability of the original study. We try to replicate the method by following the original study's methodology as accurately as possible. As in the original study, our algorithm is created through deep learning, which involves a procedure of training a

Kaggle EyePACS grading ***(International Clinical Diabetic Retinopathy scale)***	**Messidor-Original grading**	**rDR 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 referable diabetic retinopathy (rDR) grades from the grading used in Kaggle EyePACS and Messidor-Original.

neural network to perform the task of classifying images. We trained the algorithm with the same neural network architecture as in the original study: the InceptionV3 model proposed by Szegedy et al [16]. 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 pre-processed as described by the original study's protocol for pre-processing. 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 [17] 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 non-distributed stochastic gradient descent [18] as our optimization procedure. Using a different optimization procedure affects the time consumption, but not 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 used batch normalization layers [19] after each convolutional layer. Our weights were also pre-initialized using weights from the neural network trained to predict objects in the ImageNet dataset [20].

The neural network in 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 associated with either or both category 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 training and validation sets were split like in the original study: 80% was used for training and 20% was used for validating the neural network. It is estimated that 25% of the Kaggle EyePACS set consists of ungradable images [14]. So we also assessed image gradability for all Kaggle EyePACS images, and we 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 as an input for algorithm training.

Hyper-parameter settings for the optimization and validation procedure were not specified, so we conducted experiments to find hyper-parameter settings that worked well for training and validating the algorithms.

2.4 Algorithm validation

We validate the algorithm by measuring 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 [21]; 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 compared the AUC values between different validation checkpoints. To avoid stopping at a local maximum of the validation AUC function, our network may continue to perform training up to n epochs (i.e. patience of n epochs). Since the original paper did not describe details regarding the validation procedure, we had to experiment with several settings for patience. One epoch of training is equal to running all images through the network once.

We used ensemble learning [22] 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

We found that a static learning rate of 0.003 performed well under training the algorithm. 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.

The replica algorithm's performance was evaluated on two independent test sets. We provide an overview of the differences in image distribution used in our replication study compared with the original study in Figure 2 in Section 2.2. Our replicated results are an area under the receiver operating curve of 0.74 and 0.59 on our Kaggle EyePACS test dataset and Messidor-Original (Figure 4 and Table 2).

We observe mainly three things from Table 2. First, there is a large discrepancy between the AUC of our replication study and the original study. Second, the AUC did not change substantially when excluding non-gradable images. Third, the AUC increased substantially when altering the pre-processing method (see Section 4.3), but it is still low compared to the original study.

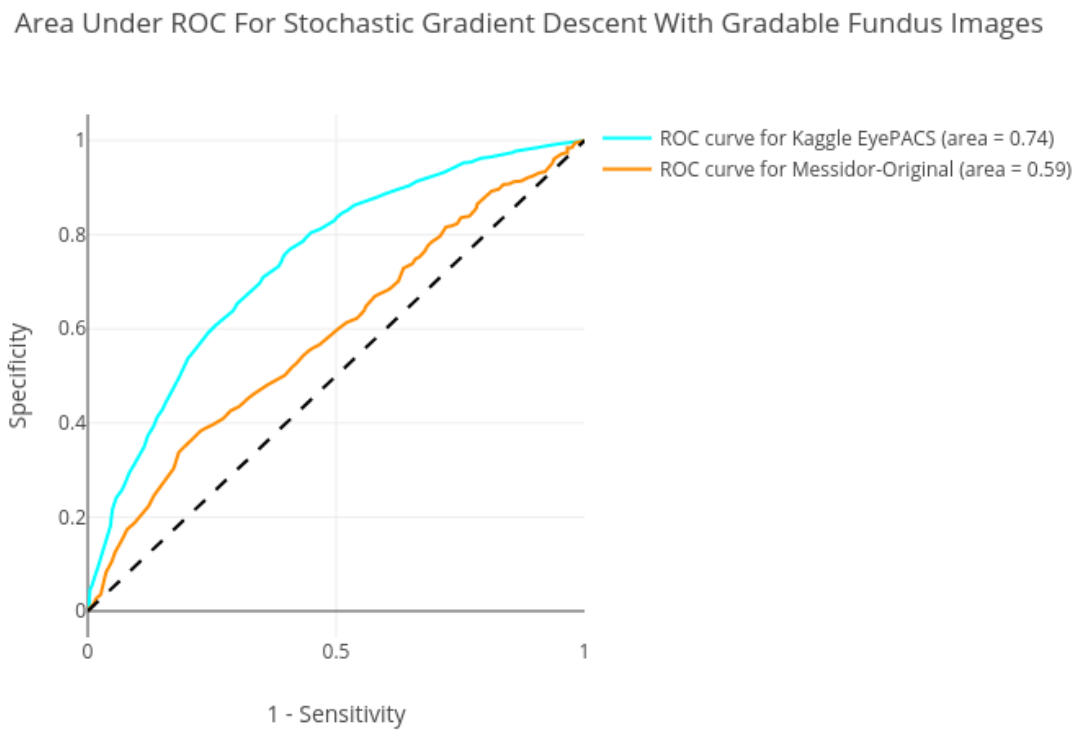

Figure 4: Area under receiver operating curve for training with only gradable fundus images and stochastic gradient descent.

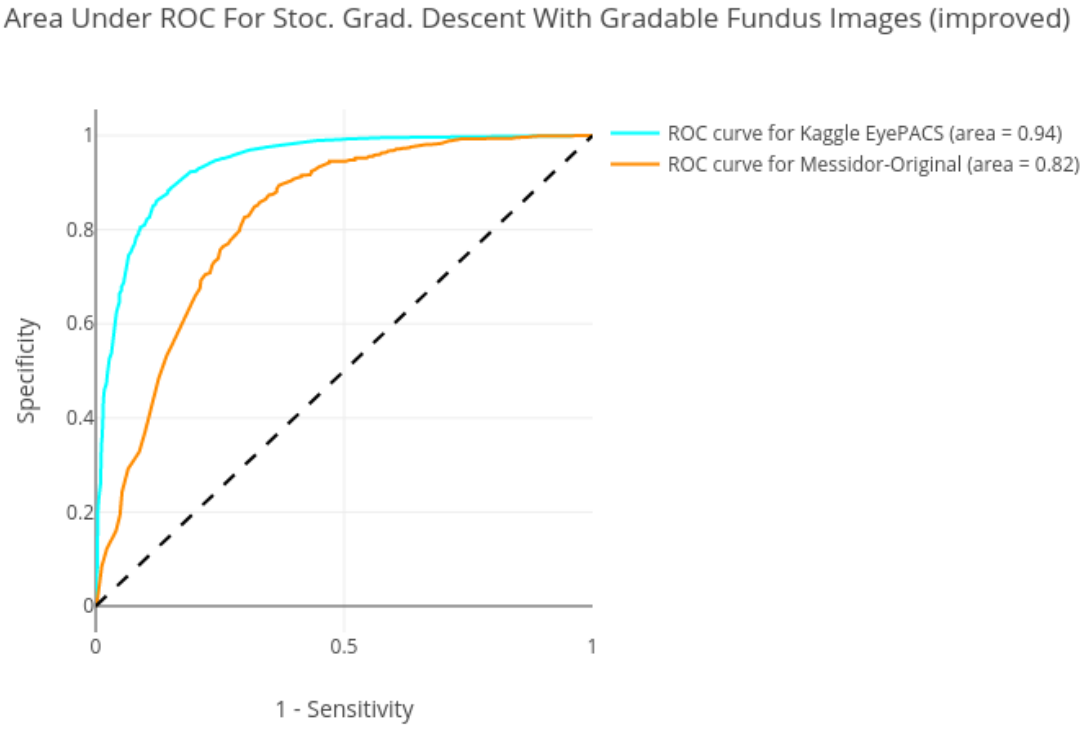

Figure 5: Area under receiver operating curve for improved training with only gradable fundus images and stochastic gradient descent.

Replication results			
Operating threshold	***High sens.***	***High spec.***	***AUC score***
Kaggle EyePACS (orig. EyePACS-1)	75.4% sens. 55.4% spec.	65.7 (90.1)% sens. 67.6 (98.2)% spec.	0.71
Messidor-Original (orig. Messidor-2)	57.6% sens. 54.6% spec.	42.2 (86.6)% sens. 68.8 (98.4)% spec.	0.60
Operating threshold	***High sens.***	***High spec.***	***AUC score (orig.)***
Only grad. Kaggle EyePACS test (orig. EyePACS-1)	79.8 (97.5)% sens. 53.7 (93.4)% spec.	67.2 (90.3)% sens. 68.2 (98.1)% spec.	0.74 (0.99)
Only grad. Messidor-Original (orig. Messidor-2)	56.6 (96.1)% sens. 54.3 (93.9)% spec.	44.0 (87.0)% sens. 64.8 (98.5)% spec.	0.59 (0.99)
Improved results			
Operating threshold	***High sens.***	***High spec.***	***AUC score***
Kaggle EyePACS test (orig. EyePACS-1)	87.0% sens. 81.9% spec.	80.6 (90.1)% sens. 88.1 (98.2)% spec.	0.93
Messidor-Original (orig. Messidor-2)	76.0% sens. 70.7% spec.	70.1 (86.6)% sens. 78.3 (98.4)% spec.	0.81
Operating threshold	***High sens.***	***High spec.***	***AUC score (orig.)***
Only grad. Kaggle EyePACS test (orig. EyePACS-1)	90.0 (97.5)% sens. 81.4 (93.4)% spec.	83.3 (90.3)% sens. 90.5 (98.1)% spec.	0.94 (0.99)
Only grad. Messidor-Original (orig. Messidor-2)	77.0 (96.1)% sens. 70.8 (93.9)% spec.	70.1 (87.0)% sens. 82.6 (98.5)% spec.	0.82 (0.99)

Table 2: Performance on test sets of replication and improved ensemble models trained with stochastic gradient descent, compared to results from the original study. In the first two rows, we summarize results for training on all images, and in the last two rows we summarize results for training on only gradable images. The results of the original study are depicted in parenthesizes. The original study did not report results for the algorithm and/or operating point in columns with missing results in parenthesizes.

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%.

Some of the details regarding the methods in the original study were not specified. 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 pre-processing 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.

4.1 Hyper-parameters

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 pre-processed 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. [14] 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.

4.2 Kaggle images

A potential drawback with the images from Kaggle 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.

4.3 Improvements

We made minor changes to the replicated method. First, we modified the pre-processing procedure. In the original study the images were scale-normalized, which we assumed meant normalizing the image values by scaling them down to being in a range from 0 to 1 [23]. We have seen from many entries in the Kaggle-competition that as a pre-processing procedure image standardization was performed, so we standardized the images instead of scale-normalizing them, and we re-trained all algorithms. This resulted in a substantial increase in performance (Figure 5 and Table 2). Why this small difference in pre-processing yielded such a large increase in performance is unclear.

Second, we re-trained the algorithms with Nesterov's accelerated gradient descent instead of stochastic gradient descent. This did however not affect the performance, but less epochs were needed to find the peak validation AUC value.

5 Conclusion

We re-implemented the main method 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 additional 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] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA - Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, 2016. doi: 10.1001/jama.2016.17216

[2] "The challenges of replication," *eLife*, vol. 6, no. e23693, 2017. doi: 10.7554/eLife.23693

[3] "Announcement: Reducing our irreproducibility," *Nature*, vol. 496, no. 7446, pp. 398–398, 2013. doi: 10.1038/496398a

[4] "Reality check on reproducibility," *Nature*, vol. 533, no. 7604, p. 437, 2016. doi: 10.1038/533437a

[5] M. Hutson, "Missing data hinder replication of artificial intelligence studies," 2018. [Online]. Available: http://www.sciencemag.org/news/2018/02/missing-data-hinder-replication-artificial-intelligence-studies

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," pp. 60–88, 2017.

[7] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, "Opportunities And Obstacles For Deep Learning In Biology And Medicine," 2018.

[8] Google, "Citations for Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," 2018. [Online]. Available: https://scholar.google.no/scholar?cites=16083985573643781536

[9] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publicly distributed image database: the Messidor database," *Image Analysis & Stereology*, vol. 33, no. 3, p. 231, 2014. doi: 10.5566/ias.1155

[10] G. Quellec, M. Lamard, P. M. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Optimal Wavelet Transform for the Detection of Microaneurysms in Retina Photographs," *IEEE Transactions on Medical Imaging*, vol. 27, no. 9, pp. 1230–1241, 2008. doi: 10.1109/TMI.2008.920619

[11] Kaggle, "Diabetic Retinopathy Detection (Data)," 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/data

[12] Messidor, "Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology," 2004. [Online]. Available: http://www.adcis.net/en/Download-Third-Party/Messidor.html

[13] Kaggle, "Diabetic Retinopathy Detection," 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection

[14] A. Rakhlin, "Diabetic Retinopathy detection through integration of Deep Learning classification framework," *bioRxiv*, 2017. doi: 10.1101/225508

[15] C. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer, and Global Diabetic Retinopathy Project Group, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003. doi: 10.1016/S0161-6420(03)00475-5

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *arXiv*, 2015. doi: 10.1109/CVPR.2016.308. [Online]. Available: http://arxiv.org/abs/1512.00567

[17] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. A. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng, "Large Scale Distributed Deep Networks," in *NIPS 2012: Neural Information Processing Systems*. Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012. doi: 10.1109/ICDAR.2011.95. ISBN 9781627480031. ISSN 10495258 pp. 1–11.

[18] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*. Heidelberg: Physica-Verlag HD, 2010, pp. 177–186. ISBN 978-3-7908-2603-6

[19] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv*, 2015. doi: 10.1007/s13398-014-0173-7.2

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. doi: 10.1007/s11263-015-0816-y

[21] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *the 13th International Conference on Neural Information Processing Systems*. Advances in Neural Information Processing Systems 13 (NIPS 2000), 2000. doi: 10.1109/IJCNN.2000.857823. ISBN 1049-5258. ISSN 10495258 pp. 402–408.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc., 2012, pp. 1097–1105.

[23] S. Theodorides and K. Koutroumbas, *Pattern Recognition*. Academic Press, 2009. ISBN 978-1-59749-272-0