Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks

Daniel E. Worrall(✉), Clare M. Wilson, and Gabriel J. Brostow

Department of Computer Science, University College London, London, UK
d.worrall@cs.ucl.ac.uk

Abstract. Retinopathy of Prematurity (ROP) is an ocular disease observed in premature babies, considered one of the largest preventable causes of childhood blindness. Problematically, the visual indicators of ROP are not well understood and neonatal fundus images are usually of poor quality and resolution. We investigate two ways to aid clinicians in ROP detection using convolutional neural networks (CNN): (1) We fine-tune a pretrained GoogLeNet as a ROP detector and with small modifications also return an approximate Bayesian posterior over disease presence. To the best of our knowledge, this is the first completely automated ROP detection system. (2) To further aid grading, we train a second CNN to return novel feature map visualizations of pathologies, learned directly from the data. These feature maps highlight discriminative information, which we believe may be used by clinicians with our classifier to aid in screening.

1 Introduction and Background

Retinopathy of Prematurity (ROP) has entered a third global epidemic [1]. Higher neonatal survival rates in developing countries and new clinical practices in the West [2] have led to a sharp increase in the number of premature babies at risk of this iatrogenic, sight-threatening disease. The preterm retina can develop abnormally at any time up to 36 weeks gestational age [3] and is treatable, thus screening plays an important role. However, screening is labour-intensive and challenging, due to insufficient understanding of ROP symptomatology, lack of gold-standard ground-truth data and poor quality fundus imaging. We investigate two methods how CNNs can be used to aid in ROP detection. (1) We detail what we believe to be the first fully automated ROP detector, which can classify per image and per examination. It harnesses traditional deep learning and modern variational Bayesian techniques. We provide information on practical tweaks that did and did not work in achieving our goal. (2) We demonstrate how the feature maps of deep CNNs can be used to create visualizations of the pathologies, indicative of disease, learned directly from the data.

ROP is difficult to detect, but conveniently it co-occurs with *plus-disease* [4], which is easier to diagnose. Plus-disease is characterized by increased *dilation* and *tortuosity* of the retinal vasculature about the posterior pole (central zone about optic disc) [5], together called *plusness*. Figure 1 shows a reference

G. Carneiro et al. (Eds.): LABELS 2016/DLMIA 2016, LNCS 10008, pp. 68–76, 2016.
DOI: 10.1007/978-3-319-46976-8_8

image of plus-disease from [4], which very clearly shows vascular dilation and tortuosity, but has been criticized for showing these quantities as more progressed than usually seen in clinic. In practice, these two quantities prove difficult to measure systematically and repeatably. Some common practical issues are: defining the segmentation boundary for vessel extraction, measuring vessel dilation/tortuosity, and discerning retinal from choroidal vessels. Other symptoms [6,7], are known but their use as indicators in screening are limited.

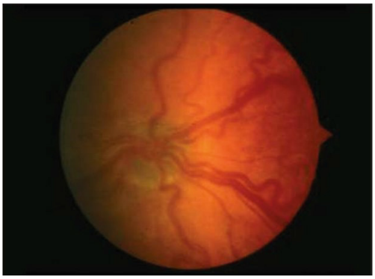
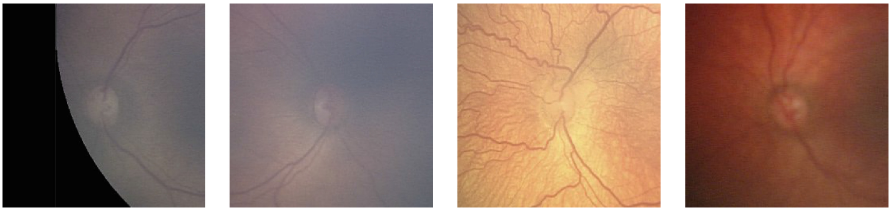


Fig. 1. Standard reference image for identifying plus-disease [4].

Most semi-automated techniques for ROP case detection rely on measuring plusness via a manual registration followed by semi- or fully-automated vessel segmentation, and by various mechanisms to extract width and tortuousity information [8]. *Jomier et al.* [9] measure width and tortuosity in all four quadrants of a vessel segmentation, which is then fed into a neural network, returning a classification of disease presence. *Wallace et al.* [10] do not seek to build a detection system and differentiate between arteriolar and venular diameter, finding that venular diameter is unimportant in classification. Their system requires significant hand preprocessing to make this work. *Swanson et al.* [11] use a custom vessel segmentation software to semi-automatically measure a tortuosity- and dilation-index for user-selected vessels. They identify plus-positive images as having a tortuosity-index above a certain threshold. In contrast to these methods, we use automated registration and feed the entire registered image into a CNN classifier. We are also able to return per-examination classifications; whereas, existing methods only return per-image classifications.

2 Proposed Method

Neonatal fundus images are usually of poor quality (see Fig. 2), captured from the unsedated premature babies, on a low resolution (640 × 480 px RGB) camera. They exhibit high levels of variation with different translations and orientations,



(a) Partial occlusion (b) Strong fades (c) Choroidal vessels (d) Blurring

Fig. 2. Examples factors impeding detection in the neonatal fundus. Only c) is diseased.

high levels of motion blur, illumination artifacts, and strongly visible choroidal vessels. Compared with adult fundus images, like in the Kaggle diabetic retinopathy competition[1], these are much degraded and harder to use for classification. The existing techniques mentioned depend on reliable vessel segmentation, which is extremely difficult in the neonatal fundus and sometimes requires some user-intervention to touch up results. Our images are also few in number ($\sim$ 1500) with high class-imbalance ($\sim$ 10 %). Below we describe our CNN-based classifier and pathology visualization.

2.1 Classifier

The classifier consists of the traditional deep learning pipeline: preprocessing, data augmentation, pretrained CNN, finetuning layers. Presently there are varying gradations of ROP and plus-disease, such as APROP and pre-plus, but we only distinguish 'diseased/healthy', since our dataset was compiled in the late 90s, before these alternatives were used by the mainstream[2].

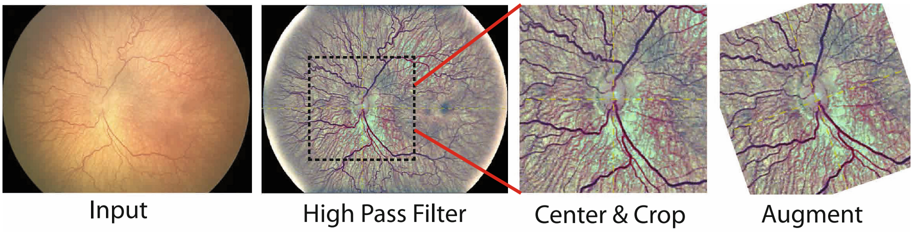


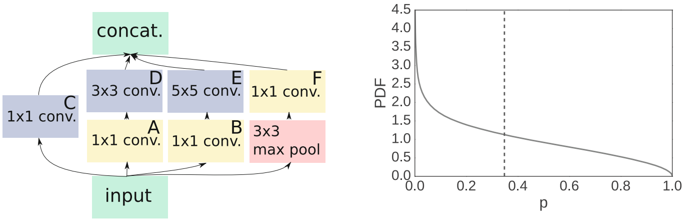
Fig. 3. Fully automated image registration, preprocessing and augmentation pipeline.

Preprocessing and Data Augmentation. Fundus images are translation registered using [12] and cropped to 240 × 240 px about the posterior pole, chosen based by cross-validation. The crop size seems small, but biologically reasonable [5]. Post-registration we high pass filter the RGB channels, removing low frequency illumination changes and global color information. This also removes retinal pigmentation, but we assume ethnicity plays a negligible role in plus screening. For variations in the data, which we cannot 'normalize out', we use data augmentation, such that the particular variation is uniformly sampled. In our case we randomly flip, rotate and take subcrops of 96 % of the original image size. The pipeline is shown in Fig. 3.

The Per-Image Classifier. Our per-image classifier consists of a 2-way softmax classifier with affine layer, stacked on top of an ImageNet pretrained

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection.

[2] Neonatal fundus imaging quality has not improved since, only the labels are different.



(a) Inception module (b) Beta(0.8,1.5)

Fig. 4. (a) An inception module consists of a combination of multiscale convolutions. Lettered blocks contain learnable parameters. The GoogLeNet contains 9 inception modules laid end-to-end. (b) The beta distribution is used in the per-exam classifier. It is biased towards healthy images. Solid line: PDF, dashed line: mean.

GoogLeNet [13]. The GoogLeNet is formed of a stack of 9 *inception modules*, which are a combination of convolutional layers and max-pooling (see Fig. 4(a)). Please refer to [13] for more details. For training we minimize a binary cross-entropy loss over the model output and target labels using RMSProp [14].

It is common to just retrain the linear classifier on the end of the network, but we found improved performance, if we included some of the convolutional layers within the 9th inception module. Retraining too many layers led to severe overfitting, however, and so we used an iterative procedure of finetuning the final n layers, and if compared to the previous $n-1$ layers validation performance increased, then we proceeded to $n+1$ layers, and so on. With parallel layers, we tried all combinations, for instance A, B and A & B in Fig. 4(a). In the end, we retrained layers ACDEF of inception module 9 with the 2-way softmax classifier.

Bayesian CNNs. CNNs return point-estimate class predictions $\mathbf{y}_* \in \mathbb{R}^D$, where $\sum_{d=1}^{D} y_{*,d} = 1$ given an input image $\mathbf{X}_* \in \mathbb{R}^{N \times M \times C}$. These are overconfident, and a more informative prediction is the posterior predictive distribution $p(\mathbf{y}_*|\mathbf{X}_*, \mathcal{D})$ where $\mathcal{D}$ is the training data. This can be found from the marginal

$$p(\mathbf{y}_*|\mathbf{X}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\mathbf{X}_*, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) \, \mathrm{d}\mathbf{w}, \tag{1}$$

where $p(\mathbf{y}_*|\mathbf{X}_*, \mathbf{w})$ represents the CNN output given image $\mathbf{X}_*$ and weights $\mathbf{w}$, and $p(\mathbf{w}|\mathcal{D})$ is a posterior over the weights, given $\mathcal{D}$. Standard CNN training follows the *maximum likelihood* priniciple, or *maximum a posteriori* when regularlization is involved, so 'traditional' predictions are made with $p(\mathbf{w}|\mathcal{D}) = \delta(\mathbf{w} - \mathbf{w}_{\mathrm{ML}})$ or $p(\mathbf{w}|\mathcal{D}) = \delta(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})$, where $\delta(x)$ is the Dirac delta function.

Recently it has been shown [15] that training CNNs with sampling behaviour, such as dropout [16], is equivalent to fitting an approximation $q(\mathbf{w}; \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ are referred to as the *variational parameters*, to the true Bayesian posterior $p(\mathbf{w}|\mathcal{D})$ over the CNN's weights. Furthermore, these samples are true samples from the approximate posterior. So to approximate Eq. 1, we replace $p(\mathbf{w}|\mathcal{D})$

with $q(\mathbf{w}; \boldsymbol{\lambda})$ and Monte Carlo sample $\mathbf{w}^{(k)} \sim q(\mathbf{w}; \boldsymbol{\lambda})$. For instance, dropout corresponds to $\mathbf{w}_i = z_i \boldsymbol{\lambda}_i$, $z_i \sim \text{Bernoulli}(\mathbf{z}_i; 0.5)$, where $\mathbf{w}_i$ is a set of incoming weights to a neuron. To yield a classification we can then simply threshold the cumulative distribution function of the posterior predictive $\Pr\{y_{*,d} > t\} > s\%$, which in words means, the probability mass of the d^{th} output above threshold t is greater than s%. We can optimize s and t to trade sensitivity–specificity.

Failed Experiments. Here we list some of the techniques, we found to hurt performance. *Vesselness features*: we tried including Frangi vesselness descriptors [17] both as a 4th input channel and as a mask on the input, we presume the network works on a similar representation of the data already. *ADAM solver*: this led to severe overfitting. *Large crops*: increasing the crop size led to underfitting. *More fully-connected layers on output*: this led to overfitting, even with dropout. *Loss function reweighting to remedy class-imbalance*: We found oversampling the smaller class better, because with data augmentation this leads to the network seeing more data per epoch. *Training the softmax classifier from lower layer outputs*: this led to underfitting. Interestingly, one would initially suspect that higher layers are more dataset specific, we found this not to be a problem. *Removing global average pooling (GAP)*: this increased the dimensionality of the output and the number of retrainable parameters, leading to overfitting.

Per-Exam Classifier. Each exam consists of different images of the same eye from differing views and with different artifacts. We build a per-exam classifier by assuming a Beta distribution $p(\pi|a, b) = \text{Beta}(\pi; a, b)$ prior over the probability π that a given eye is diseased in an examination and a Bernoulli distribution $p(c_i|\pi) = c_i^{\pi}(1 - c_i)^{1-\pi}$ on the probability an image i is classified as diseased c_i given π. The posterior over π is $\text{Beta}(\pi; N_1 + a, N_0 + b)$, where N_1 and N_0 are number of images classified as diseased and healthy, respectively, in that examination. When using the Bayesian predictive distribution, we use classifications from the thresholded cumulative distribution. The posterior predictive distribution is $p(c_* = 1|\{c_i\}_{i=1}^{N_0+N_1}) = \frac{N_1+a}{N_0+N_1+a+b}$, where c_* is the diseased/healthy classification for this exam. We found $a = 0.8, b = 1.5$ through Empirical Bayes on the training data, which places a prior on images being healthy.

2.2 Visualization

We visualize diseased regions of the fundus, by examining the CNN feature maps. GoogLeNet feature maps are too small (7 × 7 px), so we trained a separate 7-layer CNN with 3 × 3 1-padded kernels and 3 × 3 stride 2 max-pooling after every even convolution with 31 × 31 px output feature maps. There is evidence [18] that CNNs trained for the same task learn similar representations at the deepest layers.

For meaningful visualizations, we need to associate activations with a label (diseased/healthy). For this, we manipulate the GAP-layer, found just before

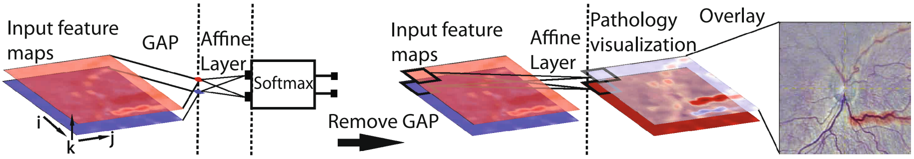


Fig. 5. The linearity of GAP and affine layers means we can swap their order, applying the affine transformation to each stack of pixels in the input.

the softmax classifier. For feature maps $\mathbf{A}_{ijk}$ with spatial indices i, j and channels k, GAP-layers return a spatial mean $a_k = \sum_{i,j} \mathbf{A}_{ijk}$. For GAP-layers feeding directly into a softmax, we need only look at the associated feature maps, but if there is an affine layer between the GAP and the softmax, then we swap the order of the GAP and affine layers,

$$\text{softmax}\left(\mathbf{W}\sum_{i,j}\mathbf{A}_{ij:} + \mathbf{b}\right) = \text{softmax}\left(\sum_{i,j}(\mathbf{W}\mathbf{A}_{ij:} + \mathbf{b})\right), \tag{2}$$

where $\mathbf{A}_{ij:}$ is the vector with entries a_k. The result is a plusness feature map and a health feature map. A schematic of the process is in Fig. 5 and examples of feature maps overlaid on input images are in Fig. 6.

3 Experiments and Results

Here we run experiments on two large and difficult ROP datasets, comparing results against a baseline and competing methods papers.

Datasets *Canada dataset*: there are 1459 usable images from 35 patients, and 347 *exams* of 2–8 images per eye. There is one label per-exam (plus/no-plus) and per-eye, but *not* per-image. We assume all images from an examination share the same label. We used this dataset for training as well as validation. *London dataset*: there are 106 individually labelled images with 4 expert labels per image. For this dataset we cannot group by exam and use this dataset for testing only.

9-fold validation. Table 1 shows results for 9-fold cross-validation on the Canada dataset for our system and a naïve baseline, a 9-layer scratch-trained CNN. Each patient is assigned to a single fold. We contrast the Bayesian model against the 'traditional' maximum likelihood solution CNN. Key statistics are averaged over the folds. Class-normalized accuracy is the mean of sensitivity and specificity and Fleiss' Kappa (FK) [19] is a measure of agreement. FK of 1.0 is full agreement, 0.0 is random agreement and < 0.0 is no agreement.

Per-exam results are mostly higher than per-image, as expected, since averaging over exams smooths over erroneous per-image labels. For both per-image and per-exam classification, the Bayesian model adds about 5 % class-normalized

Table 1. 9-fold cross-validation results on the Canada dataset. **Bold** denotes the best result for each row within per-image or per-exam.

Experiment	Per-image					Per-exam		
	Bayes.	Trad.	Base.	Jomier[9]	Wallace[10]	Bayes.	Trad.	Base.
Raw Acc	**0.918**	0.892	0.833	-	-	**0.936**	0.919	0.852
Sensitivity	0.825	0.809	0.598	0.800	**0.950**	**0.954**	0.852	0.625
Specificity	**0.983**	0.909	0.846	0.920	0.780	**0.947**	0.929	0.860
Precision	**0.607**	0.547	0.295	-	-	**0.713**	0.665	0.322
Norm. Acc	**0.904**	0.859	0.722	0.860	0.865	**0.951**	0.890	0.742
Fleiss' Kappa	**0.590**	0.547	0.246	-	-	**0.714**	0.657	0.278

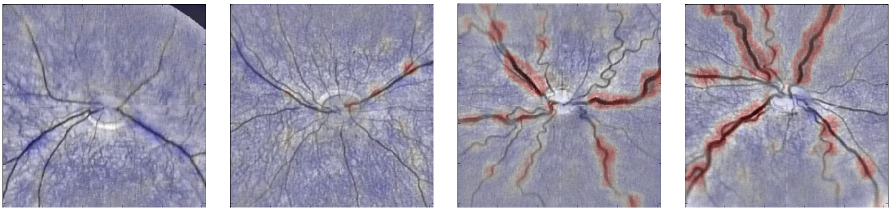
accuracy, with significant gains in sensitivity per-exam. Comparing to other methods, we are competitive, although losing on per-image sensitivity to *Wallace et al.*. We note though that the comparison of results is not straight-forward, since they use smaller test sets (20 images) and *Jomier et al.* use different methodology, testing only non-borderline images. Looking at FK, we see agreement is $0.54 - 0.72$ for our model, considered "moderate" to "substantial".

Multigrader Agreement. With the London dataset there is no groundtruth, so we report the FK score only. For a single prediction, we ensemble the outputs of the 9 cross-validation trained CNNs, taking a mean and thresholding at 50 %, results are in Table 2. Among the experts there is an FK of 0.427, but with our system this drops to 0.366/0.372. It turns out that the system agrees very strongly with one expert and disagrees strongly with another (see Table 2), and that the agreement with the closest expert is stronger than amongst the closest and furthest experts (0.194). For comparison, [20] report an FK of 0.32 for inter-clinician agreement, albeit on a separate dataset.

Table 2. Multigrader agreement is similar to levels found in [20].

Experiment	Experts alone	All experts		Closest expert		Furthest expert	
		Bayes.	Trad.	Bayes.	Trad.	Bayes.	Trad.
Fleiss' Kappa	0.427	0.366	0.372	0.551	0.546	-0.118	-0.084

Pre-GAP Visualization. Figure 6 shows the pre-GAP visualization, where red indicates diseased and blue healthy. The blue channel has been intensified for easier visualization. There is a clear indication that the CNN focuses on the vasculature in its decision-making, and that this is by far the most important indicator for plus-disease. This agrees with the current guidance for clinicians as per [4], which focuses on qualitative measurements of the width and tortuosity of retinal blood vessels.



(a) Healthy retina (b) Healthy retina (c) Diseased retina (d) Diseased retina

Fig. 6. Visualizations of learned retinal pathologies with the projected pre-GAP activation layer superimposed. Blue is healthy tissue and red is diseased tissue. The CNN has learned that wide and tortuous vessels correlate with plus-disease, as we expect. (Color figure online)

4 Conclusion, Limitations and Future Work

We have presented the first fully automated ROP detection system. We have listed techniques to finetune a GoogLeNet to small datasets, which did and did not work for us. We have also demonstrated a simple Bayesian framework to increase the accuracy of the output of a dropout trained CNN. The system copes with single images or multiple images from a single examination. For understanding we have also demonstrated how to return augmented pathology visualizations from CNNs with large enough feature maps. The code and dataset are available to download upon request.

Our multigrader experiments show that it is possible to train classifiers on subjective labels. These classifiers exhibit good agreement with some of the expert labelers. From a supervised learning perspective, a classifier can only ever be as good as its training data, as such we need to look to less human-dependent training data if we are to surpass human performance. This may involve harnessing unsupervised and semi-supervised learning. It would also be sensible to explore building spatio-temporal models of ROP progression, to see if sequences of images form better predictors of disease than single instances in time.

Acknowledgements. We would like to thank Alistair Fielder and Philip ‘Eddie’ Edwards for insightful conversations. Daniel Worrall is supported by Fight for Sight, registered charity number 1111438.

References

1. Zin, A., Gole, G.A.: Retinopathy of prematurity-incidence today. Clin. Perinatol. **40**(2), 185–200 (2013)
2. Fleck, B.W., Stenson, B.J.: Retinopathy of prematurity and the oxygen conundrum: lessons learned from recent randomized trials. Clin. Perinatol. **40**(2), 229–240 (2013)
3. Wilkinson, A., Haines, L., Head, K., Fielder, A., et al.: UK retinopathy of prematurity guideline. Eye (London, England) **23**(11), 2137 (2009)

4. Gole, G.A., Ells, A.L., Katz, X., Holmstrom, G., Fielder, A.R., Capone Jr., A., Flynn, J.T., Good, W.G., Holmes, J.M., McNamara, J., et al.: The international classification of retinopathy of prematurity revisited. JAMA Ophthalmol. **123**(7), 991–999 (2005)
5. Saunders, R.A., Bluestein, E.C., Sinatra, R.B., Wilson, M.E., Rust, P.F.: The predictive value of posterior pole vessels in retinopathy of prematurity. J. Pediatr. Ophthalmol. Strabismus **32**(2), 82–85 (1995)
6. Binenbaum, G.: Algorithms for the prediction of retinopathy of prematurity based on postnatal weight gain. Clin. Perinatol. **40**(2), 261–270 (2013)
7. Oloumi, F., Rangayyan, R.M., Ells, A.L.: Quantification of the changes in the openness of the major temporal arcade in retinal fundus images of preterm infants with plus diseaseanalysis of fundus images of infants with plus disease. Invest. Ophthalmol. Vis. Sci. **55**(10), 6728–6735 (2014)
8. Aslam, T., Fleck, B., Patton, N., Trucco, M., Azegrouz, H.: Digital image analysis of plus disease in retinopathy of prematurity. Acta Ophthalmol. **87**(4), 368–377 (2009)
9. Jomier, J., Wallace, D.K., Aylward, S.R.: Quantification of retinopathy of prematurity via vessel segmentation. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2879, pp. 620–626. Springer, Heidelberg (2003)
10. Wallace, D.K., Zhao, Z., Freedman, S.F.: A pilot study using roptool to quantify plus disease in retinopathy of prematurity. J. Am. Assoc. Pediatr. Ophthalmol. Strabismus **11**(4), 381–387 (2007)
11. Swanson, C., Cocker, K., Parker, K., Moseley, M., Fielder, A.: Semiautomated computer analysis of vessel growth in preterm infants without and with ROP. Br. J. Ophthalmol. **87**(12), 1474–1477 (2003)
12. Worrall, D.E., Brostow, G.J., Wilson, C.M.: Automated optic disc (OD) localization in the neonatal fundus image. In: ARVO (2016)
13. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
14. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop. COURSERA: Neural Netw. Mach. Learn. **4**, 2 (2012)
15. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. arXiv preprint arXiv:1506.02142 (2015)
16. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
17. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, p. 130. Springer, Heidelberg (1998)
18. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 991–999 (2015)
19. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. Wiley, Hoboken (2013)
20. Gschließer, A., Stifter, E., Neumayer, T., Moser, E., Papp, A., Pircher, N., Dorner, G., Egger, S., Vukojevic, N., Oberacher-Velten, I., et al.: Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. Am. J. Ophthalmol. **160**(3), 553–560 (2015)