Improved interpretability for computer-aided severity assessment of Retinopathy of Prematurity

Mara Graziani[a,c], James M. Brown[b], Vincent Andrearczyk[a], Veysi Yildiz[d], J. Peter Campbell[e], Deniz Erdogmus[d], Stratis Ioannidis[d], Michael F. Chiang[e], Jayashree Kalpathy-Cramer[b], and Henning Müller[a,b,c]

[a]University of Applied Sciences Western Switzerland (HESSO), Sierre, Switzerland;
[b]Martinos Center for Biomedical Imaging, Charlestown, MA, USA;
[c]University of Geneva, Switzerland;
[d]Electrical and Computer Engineering, Northeastern University, Boston, MA, United States;
[e]Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, Portland, OR

ABSTRACT

Computer-aided diagnosis tools for Retinopathy of Prematurity (ROP) base their decisions on handcrafted retinal features that highly correlate with expert diagnoses, such as arterial and venous curvature, tortuosity and dilation. Deep learning leads to performance comparable to those of expert physicians, albeit not ensuring that the same clinical factors are learned in the deep representations. In this paper, we investigate the relationship between the handcrafted and the deep learning features in the context of ROP diagnosis. Average statistics on the handcrafted features for each input image were expressed as retinal concept measures. Three disease severity grades, i.e. *normal, pre-plus and plus*, were classified by a deep convolutional neural network. Regression Concept Vectors (RCV) were computed in the network feature space for each retinal concept measure. Relevant concept measures were identified by bidirectional relevance scores for the *normal* and *plus* classes. Results show that the curvature, diameter and tortuosity of the segmented vessels are indeed relevant to the classification. Among the potential applications of this method, the analysis of borderline cases between the classes and of network faults, in particular, can be used to improve the performance.

Keywords: deep learning, interpretability, machine learning, retinopathy, plus disease

1. INTRODUCTION

Retinopathy of Prematurity (ROP) is a disease that affects premature infants and can lead to blindness if not promptly treated. Particularly important for treatment planning is the correct classification of three grades of ROP, namely *normal, pre-plus and plus*. Fully-automated diagnosis tools have recently shown classification performance comparable to experts and significantly higher than non-experts, especially when Convolutional Neural Networks (CNNs) are used.[1,2] It is still not clear, however, whether a link can be established between the deep CNN features and the handcrafted features of more traditional machine learning approaches.[2] The identification of such a link, if possible, would give clinicians a better understanding of the deep learning black-box. Our approach proposes to link the handcrafted and the deep features by measuring the relevance of the first ones in the classification with a deep CNN. Methods to compute the influence of arbitrary concepts to the network classification were proposed by Testing with Concept Activation Vectors (TCAV)[3] and Regression Concept Vectors (RCVs).[4] In this paper, we apply RCVs to ROP data to compute global and local explanations about the network decisions, mapping the latter to handcrafted features of tortuosity, dilation and curvature of retinal vessels. The results highlight the representations of curvature, dilation and tortuosity learned by the network. By computing bidirectional relevance scores (Br), we show that these concepts are relevant to the network prediction. Moreover, individual relevance scores can be used as support in the diagnosis, since

Further author information: (Send correspondence to M.G.)
M.G.: E-mail: mara.graziani@hevs.ch

they propose complementary information to the prediction. The potential applications of this approach are numerous. The retinal concept measures that are the most responsible for misclassification could, for instance, be detected and penalized. Mislabeled examples at the decision boundaries between the classes can also be rectified, improving the classification accuracy.

2. RELATED WORK

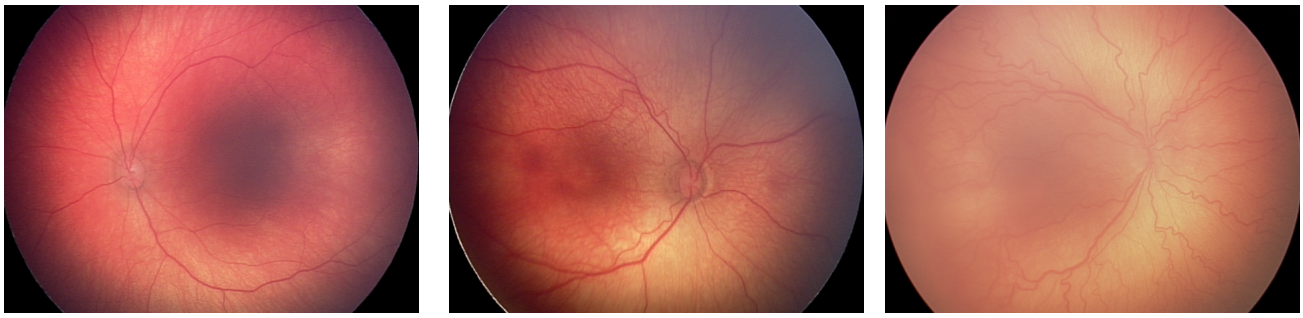
Treatment indication of ROP focuses on the detection of the three grades of ROP. For instance, the *plus* grade is the most important predictor of progression of ROP and usually requires prompt treatment, while *pre-plus* requires very close monitoring. Early work relied on quantitative features of retinal tortuosity, curvature and dilation extracted from manual annotations.[2] A long and intense process of manual vessel tracking is needed to extract the vessel segmentation, which is prone to noisy segmentations. Feature extraction is performed on the segmentations in different modalities, which vary depending on the location at which the feature is computed and on its formulation. Point-based features are extracted from any point on a retinal vessel, while segment based features are extracted from a set of vessel points that are between two vessel intersections. Eleven different types of features were widely evaluated in the literature.[2] Especially, point-based features of curvature performed better than point-based and segment-based features of dilation. Deep learning led to fully automated vessel segmentation and ROP classification at expert-level performances.[5] It raises the question of whether measures of curvature and dilation are still important to the network classification.

Intense research focused on explaining the internal behavior of neural networks, offering a wide range of visualization techniques and delineating a taxonomy of desiderata, methods and evaluation criteria.[6–8] The notion of global interpretability was defined as the attempt to explain overall behaviors of the model, for example how some filters in the model or combinations of neurons can influence the prediction. Local interpretability, on the other hand, focuses on individual testing inputs, highlighting characteristics that were more salient to the network output. Especially, the relevance or saliency of input factors to the network decision was proposed in several gradient-based methods.[8–11] Outputs of these methods are typically local explanations that are gathered in attribution maps and overlayed to the original input image. Despite their fragility to shifts in the input features,[12,13] saliency maps constitute a standard method for network interpretability. Research on linearity in the activation space has shown how linear classifiers can learn meaningful directions that can be mapped to semantic word embeddings in[14] or human-friendly visual concepts in.[15] Particularly, TCAV and RCV seek a direction (a high-dimensional vector) that represents the presence of a concept in the activations of a layer and that can be used to compute its saliency. The main idea of TCAV is to compute the direction representative of a concept as the normal to the hyperplane that separates a set of concept images from a set of random images. RCV extends this idea to continuous concept measures, identifying their direction of greatest increase by Linear Least Squares regression (LLS). Saliency is computed for each testing input as the directional derivative of the network's output along the concept direction. Bidirectional relevance scores were proposed to obtain global explanations of the model starting from the individual sensitivity scores.[4] More details about RCV and Br are provided in Section 3.

3. METHODS

3.1 Dataset

A database of 4800 de-identified posterior retinal images is used to perform the analysis. The images were captured by a commercially available camera (RetCam; Natus Medical Incorporated, Pleasanton, CA). A set of 3024 images (1084 for *normal*; 1074 for *pre-plus*; 1080 for *plus*) is used to train the network. Figure 1 illustrates one raw images for each class. The images were pre-processed with the pipeline proposed by Brown et al.,[5] which used a deep CNN to segment the retinal vasculature. After segmentation, the images are resized to 224 x 224 pixels and data augmentation is applied to balance the dataset, i.e. 90-degree rotations and horizontal and vertical flipping. A testing set of 985 samples (817 for *normal*; 148 for *pre-plus*; 20 for *plus*) is used to evaluate network performance and to compute relevance scores for the retinal concept measures. Note that ROP is a low prevalence disease (only 3% prevalence), hence data scarcity is the main cause of the class imbalance between *plus* and *normal* cases.



(a) (b) (c)

Figure 1: Sample images for the three classes *normal*, *pre-plus* and *plus*. Best viewed in color.

3.2 Network Architecture

An Inception-v1 network (GoogLeNet)[16] pre-trained on ImageNet is finetuned to classify the preprocessed images. The last softmax activated layer is trained to output a probability for each of the classes *normal*, *pre-plus* and *plus*. The categorical cross-entropy loss is optimized for 100 epochs with stochastic gradient descent and a constant learning rate of 0.0001. The hyperparameters are tuned by 5-fold cross-validation as in Brown et al.[5]

3.3 Model Interpretability

3.3.1 Handcrafted Features

We extract 11 types of handcrafted features from the images, whose importance to the evaluation of the ROP disease was evaluated by Aeter-Cansizoglu et al.[2] The impact of the features choice on the diagnosis was thoroughly investigated in the literature and the selected method constitutes a reference standard with high inter-expert agreement.[2] For each type of feature, 8 traditional statistics (such as minimum, maximum, mean, median and second and third moments) and 5 Gaussian Mixture Model (GMM) statistics are extracted, for a total of 143 handcrafted features (more details in the Appendix 1). Such features are extracted from the automated vessel segmentations and express curvature, tortuosity and dilation of retinal arteries and veins (details reported in Table 1). The features in Table 1 are first computed independently for each vessel in the image. The "vesselness" of the whole retinal sample is then summarized by standard statistics such as mean and median of the per-vessel features. A ranking of the features is computed on the basis of their Gini coefficient for random forest classification of *normal* and *pre-plus* or worse on 100 random train-test splits (with replacement) of the data. The retaining criterion used for this analysis identified a set of six measures that covered a wide set of clinically interpretable features, discarding measures with a frequency of appearance lower than 10% in the ranking. The retained measures were: *curvature mean*, *curvature median*, *avg point diameter mean*, *avg segment diameter mean*, *cti mean* and *cti median*. Notwithstanding, the same analysis can be repeated with a different criterion or with the exhaustive analysis of all the 143 features.

Feature	Description	Clinical interpretation
curvature	$\kappa(s)$	rate of direction change
avg segment diameter	$\#pixels/L_c(x)$	global dilation
avg point diameter	$W_n(x)$	absolute dilation
Cumulative Tortuosity Index (CTI)	$\text{cti}(x) = L_c(x)/L_x(x)$	curving, curling, twisting rate

Table 1: Handcrafted feature description and clinical interpretation. $\kappa(s)$ describes the rate of changing velocity between points with respect to the rate of changing the curve length between points. L_c and L_x denote respectively curve and chord length. W_n denotes the width of vessel on the normal direction.

3.3.2 Regression Concept Vectors

RCVs are computed by seeking in the activation space of a layer the direction of greatest increase of a set of measurements for one retinal concept. This direction is computed as the LLS regression of the retinal concept

measures on the training inputs.[4] The regression vector is normalized to obtain a unit vector that identifies the direction of increasing values of the concept measure. The determination coefficient of the linear regression, R^2, is used to evaluate the quality of the RCV and to obtain insights about the layer in the network where the concepts are learned. For each image $\mathbf{x}_i$ and each retinal concept C, the directional derivative of the network output $f(\cdot)$ on the RCV ($\overrightarrow{v}_C$) is computed in the activation space of a layer Φ^l as shown in Eq. 1.

$$S^l_{C,i} = \frac{\partial f(\mathbf{x}_i)}{\partial \Phi^l(\mathbf{x}_i)} \cdot \overrightarrow{v}_C \tag{1}$$

For a given concept, the directional derivative expresses the sensitivity of the network output to changes in the input along the direction of increasing values of this concept.[4] The sensitivity scores are summarized in global explanations by Br scores. Br scores are defined as the ratio between R^2 and the coefficient of variation (standard deviation over the mean) of the individual sensitivities:[4]

$$Br = R^2 \times \left(\frac{\hat{\mu}}{\hat{\sigma}}\right) \tag{2}$$

Br scores were proposed to express the global relevance of a concept on the basis of the magnitude and consistency of the individual sensitivities.[4] However, different scoring systems could be proposed to evaluate the global relevance of a concept. RCVs were computed by extending the Keras Visualization Toolkit*.[17]

3.4 Visualization

We visualize the layer activations with Uniform Manifold Approximation and Projection (UMAP).[18] UMAP is a manifold learning technique for dimensionality reduction, competitive with the more traditional t-SNE for visualization quality and runtime performance. The optimal low-dimensional representation is obtained by minimizing the cross-entropy between two topological representations, namely one for the high dimensional data, and the second for the candidate compression. We visualize the 2D projection of the activations at the filter concatenation layer of the last inception module (for brevity we will refer to this layer as `Mixed_5c_Concatenated` in the rest of the paper). The distribution of the concept measures among the network's representation of each class is visualized by changing the color intensity and the size of each data point. Moreover, the predictions for the testing set are visualized in the same projection space and misclassification errors are shown in different color and size.

4. RESULTS

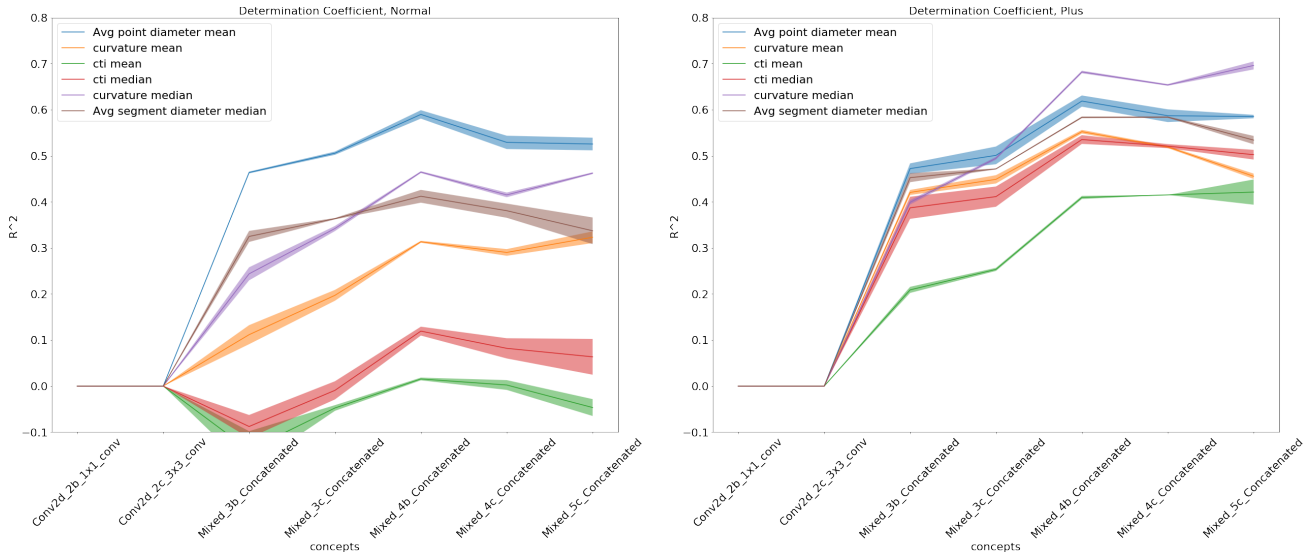
4.1 Model performance and global interpretability

The classification achieved 90% accuracy on the training set and 81.5% accuracy on the test set (0.97 AUC *normal*, 0.93 AUC *plus*).

Figure 2 shows the R^2 at different layers for inputs of the *normal* and *plus* classes. *Avg point diameter mean* and *curvature median* are the best performing concepts for both classes. Retinal concept measures about tortuosity (i.e. *cti mean* and *cti median*) are poorly learned on *normal* cases (R^2 is clipped to zero). The 95% confidence intervals of the R^2 are particularly narrow, with RCVs remaining unchanged when computed over different subsets of the data (we computed the angles between computations which were close to zero). Figure 3 shows the global Br scores that were computed at the filter concatenation layer of the last inception module and scaled to be in the range [-1,1].[4] *Curvature median* is the most relevant concept in both classes, but with opposing directions ($Br = -1$ for *normal* and $Br = 1$ for*plus*). *Avg point diameter mean* is the second most important concept for *normal* cases, $Br = -0.99$. For *plus* cases, *Avg point diameter mean* and *cti median* are equally important with both $Br = 0.56$. *Cti median* has almost zero Br in *normal* cases.

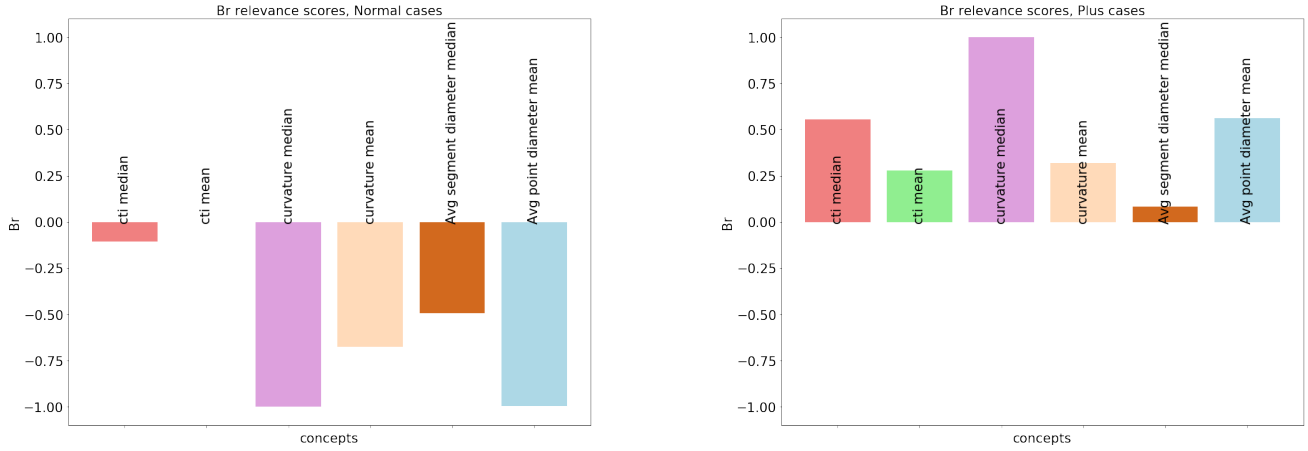
For each of the concept measures analyzed, we visualize the 2D UMAP compression of the representation learned at the `Mixed_5c_Concatenated` layer (in Figure 4). The ROP classes are colored with three different colors, namely green for *normal*, blue for *pre-plus* and red for *plus*. The values of the handcrafted features corresponding to the desired concept measure are represented with different color intensities and sizes.

*Our source code is available at `https://github.com/maragraziani/iMIMIC-RCVs.git`



(a) (b)

Figure 2: R^2 of the regression at different layers of the network, averaged over 3 repetitions, for (a) *normal* and (b) *Plus* input samples. 95% confidence intervals computed over 3 repetitions are illustrated. Best viewed in color.



(a) (b)

Figure 3: Global *Br* relevance scores on the testing set for *normal* and *plus* images. Best viewed in color.

4.2 Misclassification

The *normal* class is the most affected by misclassification. We measure the gravity of the misclassification error as the distance between the predicted class label and the true class label. When the distance is negative, the case is underestimated, namely the prediction assigned a lower class than the real (e.g. *normal* instead of *pre-plus* or *pre-plus* instead of *plus*). Similarly, when the distance is positive, the case is overestimated, with the assignment of *pre-plus* or *plus* to a patient that was less in danger. Figure 5 illustrates the misclassification errors in the 2D UMAP compression of the activations of the `Mixed_5c_Concatenated` layer. Most of the misclassification errors are made at the decision boundary between *pre-plus* and *normal*, at gravity 1 (points in orange color). The emptiness of the left portion of the space shows how misclassification on *plus* cases is kept to a minimum, with a very low number of false negatives. Among them, a false negative sample lies at the decision boundary between the *pre-plus* and *plus*. The individual relevance scores for such data point are presented in Figure 6b.

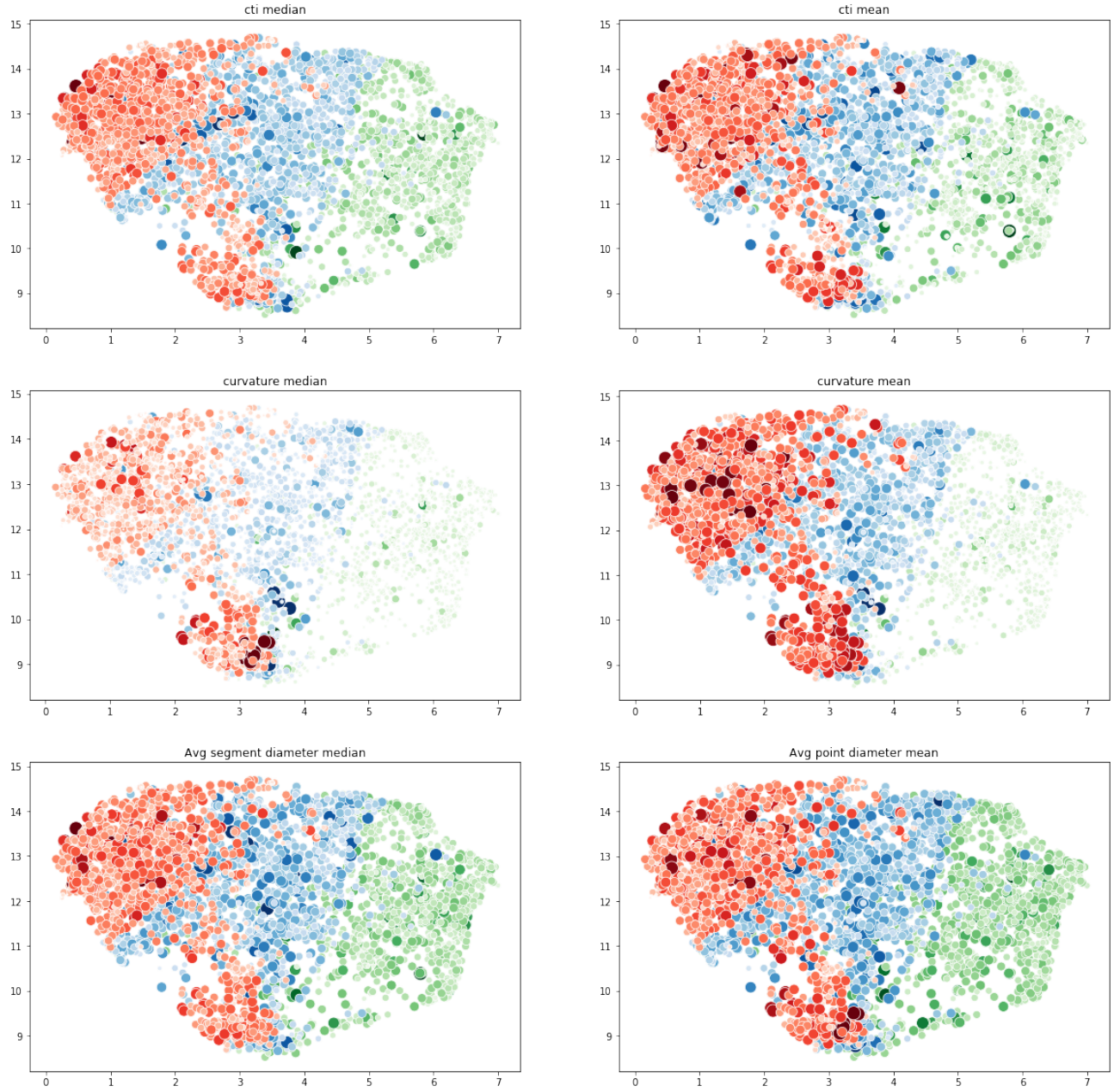


Figure 4: UMAP representation at the `Mixed_5c_Concatenated` layer of the Inception model. The three colors respresent the three classes, respectively *normal* as green, *pre-plus* as blue and *plus* as red. The color intensity and the size of the data point have been changed to show the value of the concept measure. Smaller values of the concept measure correspond to less bright, smaller data points. Best viewed in color.

4.3 Individual relevance scores

Figure 6 presents the individual relevance scores for a correctly classified sample (6a) and for the misclassification of *plus* as *pre-plus* (6b). The original values of the handcrafted features (which were used as concept measures) are reported on the left of the image. The network probability of each class is shown on top of the segmentation, as p_n, p_{pre} and p_{plus}. For the misclassified data point, the scores highlight that higher values of curvature and tortuosity would increase the probability of the *plus* class.

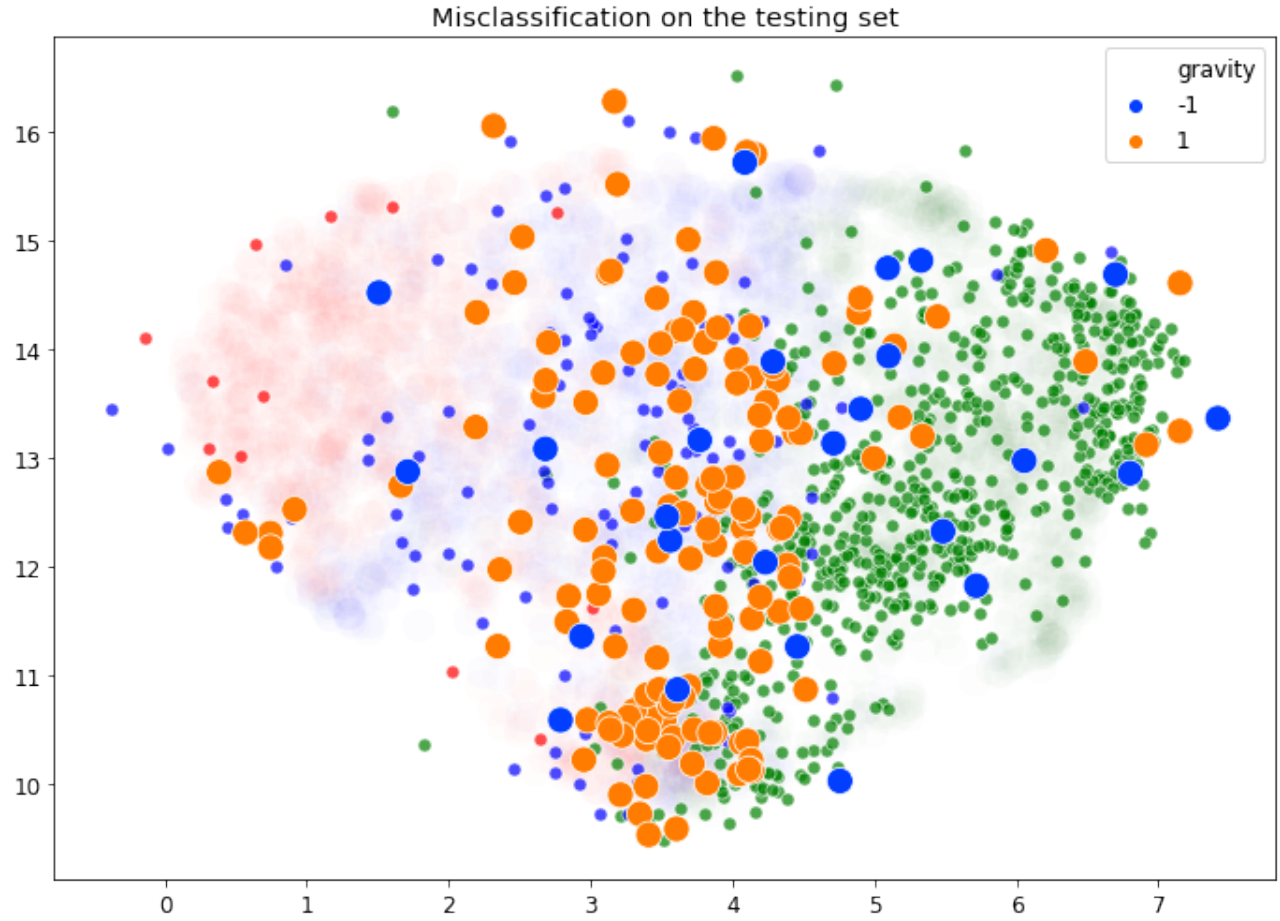
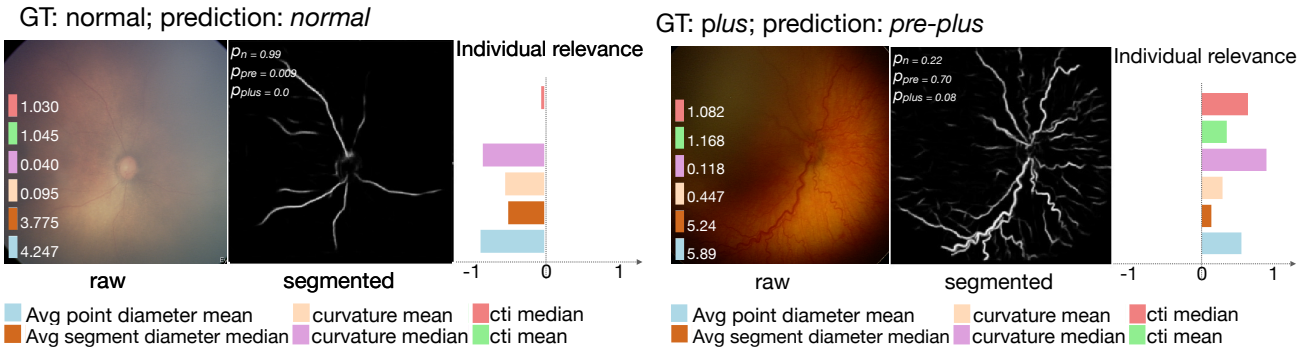


Figure 5: 2D UMAP of the `Mixed_5c_Concatenation` layer activations for the training data points (in transparency in the background), the testing data points (in brighter color) and the miclassified data points. The three classes are colored in green for *normal*, blue for *pre-plus* and red for *plus*. Misclassification data points are in larger size and colored according to misclassification gravity (false negatives in dark blue, false positives in orange). Best viewed in color.



(a) (b)

Figure 6: Individual relevance scores for (a) correct classification (b) misclassification of *plus* as *pre-plus*. The original values of the concept measures are overlayed on top of the raw image on the left. The network probabilities for *normal*, *pre-plus* and *plus* are respectively p_n, p_{pre} and p_{plus}. Best viewed on the screen.

5. DISCUSSION

The analysis of the R^2 (in Figure 2) highlights that concepts are learned better at deeper layers of the network. Interestingly, *avg point diameter mean* has a better regression than *curvature median* in Figure 2a. However, this situation is reversed for *plus* cases. After a crossing point in the middle layers, the R^2 of *curvature median* overcomes the one of *avg point diameter mean* (Figure 2b). Moreover, the narrow confidence intervals of the R^2 show that the solution of the regression is robust to multiple repetitions, leading to consistent directions. The opposing signs of the Br scores for the two classes can be interpreted as a decrease (if scores are negative) or an increase (if scores are positive) in the probability for this class when the values of the concept measure increase. For instance, the negative Br scores of the *normal* class can be interpreted as a shift in the prediction towards *pre-plus* or *plus*. Since measures of curvature and dilation emerged as the most relevant, the shift in the prediction towards *pre-plus* or worse is more marked when the inputs present larger measures of such features. Tortuosity has little relevance in the classification of *normal* cases, while it becomes as important as dilation measures in the classification of *plus*. Finally, visualizing the misclassified data points on the 2D compression of the activation space helped in finding regions where the network faults were more frequent. For instance, most of the errors were false positives (on inputs of class *normal*), and the false negatives were very few. The visualization of the individual relevance scores of the misclassified data points shows that larger values of curvature or tortuosity may have increased the probability of the class *plus*.

Overall, the application of the RCV framework allowed to identify relevant concepts in a fully-automated system for ROP diagnostics. The interpretations of the network can be related to the handcrafted features used in more traditional machine learning approaches. Our analysis could be extended to a variety of alternative handcrafted features. One limitation of this approach is the linearity assumption of the LLS. Especially, a non-linear regression could be found for concept measures that were not possible to regress linearly. Moreover, regularized regression would lead to even more stable solutions. Notwithstanding, the results show that the proposed approach can generate explanations of the network decisions that are clinically interpretable and consistent with previous findings in ROP.

6. CONCLUSIONS

This work presented a method to compare deep learning features to more traditional features of vessel curvature, tortuosity and dilation for the diagnosis of ROP. We linked deep features to handcrafted features by using the latter as concepts measures in the RCV framework. RCVs were computed at different layers of the network and the relevance of each concept to the classification of ROP grades was computed on testing inputs of the classes *normal* and *plus*. The Br scores highlighted the network focus on vessel measures to perform the classification. Curvature emerged as a particularly discriminating factor between the classes *plus* and *normal*. Tortuosity is only relevant for *plus* cases, suggesting that the neuron activating for the class *plus* is the only one attending features of vessel tortuosity. However, more experiments are needed to estimate the attention given to a concept by a single neuron. Individual scores were used to highlight the factors most responsible for misclassification. The penalization of these in the network optimization function could improve the classification accuracy. Moreover, the approach versatility allows the extension to similar pathologies which may have larger datasets available. Concepts that could be relevant in more than one clinical application (e.g. vessel features in diabetic rethinopathy) could be identified and eventually used during network training. In conclusion, this paper is a step towards a better understanding of the connection between deep features and retinal vascular features. Further research is still needed to fully uncover the deep learning mechanisms, thus broadening the perspectives on the application of deep learning to computer-aided ROP diagnosis.

ACKNOWLEDGMENTS

This work was possible thanks to the project PROCESS, part of the European Unions Horizon 2020 research and innovation program (grant agreement No 777533). This work was also supported by the National Institutes of Health (R01EY019474, P30EY10572, P41EB015896), by the National Science Foundation (SCH-1622542 at MGH; SCH-1622536 at Northeastern; SCH-1622679 at OHSU), by unrestricted departmental funding from Research to Prevent Blindness (OHSU), and by a training grant from the NIH Blueprint for Neuroscience

Research (T90DA022759/R90DA023427). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Conflict of Interest Disclosures Dr Chiang is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA), a Consultant for Novartis (Basel, Switzerland), and an initial member of Inteleretina (Honolulu, HI). Dr Chan is a Consultant for Visunex Medical Systems (Fremont, CA). Henning Müller is in the advisory board of Zebra Medical Vision and of ContextVision.

REFERENCES

[1] J. M. Brown, J. P. Campbell, A. Beers, K. Chang, K. Donohue, S. Ostmo, R. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, *et al.*, "Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning," in *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, **10579**, p. 105790Q, International Society for Optics and Photonics, 2018.

[2] E. Ataer-Cansizoglu, V. Bolon-Canedo, J. P. Campbell, A. Bozkurt, D. Erdogmus, J. Kalpathy-Cramer, S. Patel, K. Jonas, R. P. Chan, S. Ostmo, *et al.*, "Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the i-rop system and image features associated with expert diagnosis," *Translational vision science & technology* **4**(6), pp. 5–5, 2015.

[3] B. Kim, J. Gilmer, F. Viegas, U. Erlingsson, and M. Wattenberg, "Tcav: Relative concept importance testing with linear concept activation vectors," *arXiv preprint arXiv:1711.11279* , 2017.

[4] H. M. M. Graziani, V. Andrearczyk, "Regression concept vectors for bidirectional explanations in histopathology," *to be presented at Interpretability of Machine Intelligence in Medical Image Computing at MICCAI* , 2018.

[5] J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, J. Kalpathy-Cramer, *et al.*, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," *JAMA ophthalmology* .

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arxiv:1702.08608* , 2017.

[7] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490* , 2016.

[8] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing* , 2017.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization.," in *ICCV*, pp. 618–626, 2017.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034* , 2013.

[11] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR* **abs/1311.2**, 2013.

[12] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," *arXiv preprint arXiv:1711.00867* , 2017.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, ACM, 2016.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[15] B. Kim, J. Gilmer, F. Viegas, U. Erlingsson, and M. Wattenberg, "TCAV: Relative concept importance testing with linear concept activation vectors," *arXiv preprint arXiv:1711.11279* , 2017.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[17] R. Kotikalapudi and contributors, "keras-vis." https://github.com/raghakot/keras-vis, 2017.

[18] L. McInnes and J. Healy, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426* , 2018.

7. APPENDIX

7.1 ROP features

Feature types Eleven types of features that quantify either tortuosity or dilation were used for the experiments. These features are divided into three groups depending on where and how the feature is computed: point-based, segment-based or tree-based. *Average point diameter* is an example of point-based feature, which describes the width of the vessel (on the direction normal to the blood flow) at each point in the image. Segmented-based features are computed on segments of the vessel trace. For example, *average segment diameter* is obtained by dividing the number of pixels on the vessel by the vessel curve length. Finally, an example of tree-based feature is the *distance to the center of the optic disc*, namely for each vessel segment the distance between the ending point of the vessel and the disc center. An exhaustive comparison of the feature types can be found at.[2]

Feature statistics Statistics are computed on the features. Traditional statistics include the minimum and maximum first and second values, the mean, median and higher central moments (up to the third). Moreover, the pool of feature values is separated into two clusters to separate out the signal from normal and abnormal vessels. The normal and abnormal clusters are then fit into a GMM and the mean, variances and component coefficients are respectively used as GMM statistics. For each feature type 8 traditional statistics and 5 GMM statistics were computed, generating a total of 143 measures.

Feature ranking Feature ranking was used to identify the most important features for the classification of the ROP disease. We trained 100 random forest classifiers on random data shuffles, ranking the features for importance according to their Gini coefficient and selecting the top 5 in all the 100 repetitions. Interestingly the median of the Cumulative tortuosity index appeared in the top 5 for all 100 models. Table N reports how often each feature appeared in the top 5, while Figure 7 shows the kernel density estimation of the feature values for each of the three ROP classes.

Feature Name	type	frequency in the top5
cti median	segment-based	100%
cti mean	segment-based	87%
curvature median	point-based	85%
curvature mean	point-based	42%
average point diameter mean	point-based	21%
average segment diameter median	segment-based	10%

Table 2: Top 5 frequency in the classification of ROP disease over 100 random forests repetitions

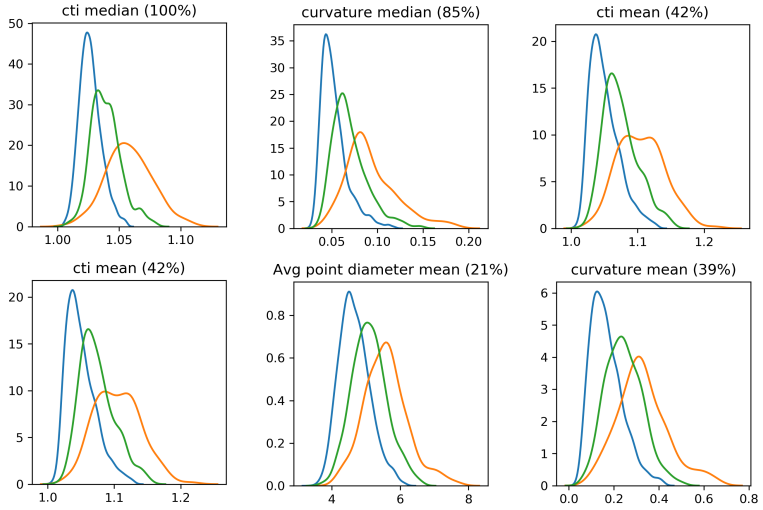


Figure 7: Kernel Density Estimation of the handcrafted feature values for the three classes (*normal* in orange, *pre-plus* in green, *plus* in blue).