Automated ROP Diagnostic System based on Comparisons and U-Net Segmentation

Peng Tian
pengtian@ece.neu.edu
Northeastern University
Boston, MA, USA

Jennifer Dy
jdy@ece.neu.edu
Northeastern University
Boston, MA, USA

Deniz Erdoğmuş
erdogmus@ece.neu.edu
Northeastern University
Boston, MA, USA

Susan Ostmo
ostmo@ohsu.edu
Oregon Health & Science University
Portland, OR, USA

J. Peter Campbell
campbelp@ohsu.edu
Oregon Health & Science University
Portland, OR, USA

Michael F. Chiang
michael.chiang@nih.gov
National Eye Institute
Bethesda, MD, USA

Stratis Ioannidis
ioannidis@ece.neu.edu
Northeastern University
Boston, MA, USA

ABSTRACT

Retinopathy of Prematurity (ROP) is a disease affecting premature infants and may lead to childhood blindness. Due to lack of trained ophthalmologists, developing a fully automated ROP diagnostic system can significantly benefit the infants affected by ROP. Based on manually segmented features, previous work produces severity scores for ROP with excellent prediction accuracy. However, when automated segmentation employed, a significant accuracy drop is observed. In this paper, we show that U-Net segmentation, which is automated, comes at no accuracy loss.

KEYWORDS

Retinopathy of Prematurity, learning from comparisons, U-Net

ACM Reference Format:
Peng Tian, Jennifer Dy, Deniz Erdoğmuş, Susan Ostmo, J. Peter Campbell, Michael F. Chiang, and Stratis Ioannidis. 2021. Automated ROP Diagnostic System based on Comparisons and U-Net Segmentation. In *PETRA '21: ACM International Conference on PErvasive Technologies Related to Assistive Environments, June 25 -July 2nd, 2021, Corfu, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3453892.3454002

1 INTRODUCTION

Retinopathy of Prematurity (ROP), a retinal disease affecting prematurely born babies, is a leading cause of childhood blindness worldwide [8, 9, 11]. The cases of ROP have risen in recent years as the survival rate of premature infants is increasing all over the world [2]. Most cases of ROP can be treated if patients receive early and proper diagnosis. Nevertheless, access to the trained ophthalmologists in many regions is limited. Thus, an accurate and automated diagnostic system could significantly improve the chance of patients receiving diagnosis.

ROP has three classes: Plus, Pre-Plus and Normal. *Plus* is defined as tortuorsity of arteries and dilation of veins in a retina greater or equal to that in a standard photograph [5]. In 2005, an intermediate class Pre-Plus was introduced and defined as retinal vessels are more tortuous and dilated than Normal but less than Plus [10]. It is also observed that even though significant inter-expert variabilities exist in ROP diagnosis, experts are more likely to agree when ordering the severties between two ROP images [12]. Tian et al. [16] and Yıldız et al. [17] show that jointly learning from both class (diagnostic) labels and pairwise comparison labels, i.e., relative severities between ROP image pairs, improves the accuracy of ROP diagnosis. In particular, by incorporating pairwise comparison labels, Tian et al. [16] can achieve a 0.97 AUC in ROP classification.

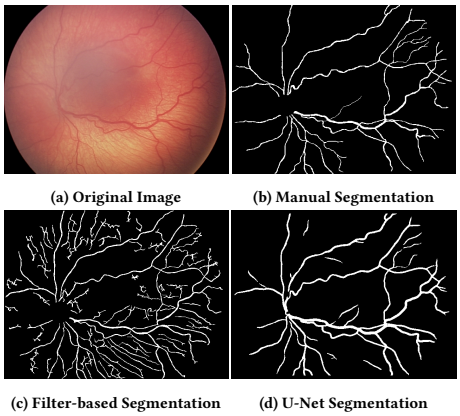


Figure 1: Comparison between different segmentation methods [16]. The original image and its manual segmentation are showed in Fig. 1a, Fig. 1b, respectively. The filter-based segmentation in Fig. 1c is much noiser and U-Net segmentation is similar to the manual segmentation.

As ROP is defined on tortuosity and dilation on vessels, the first step of several works [1, 4, 16–18], is *segmentation*: one first converts colored ROP images into binary masks, indicating whether a pixel is on the vessel. The quality of segmented vessels is an important factor affecting the diagnosis outcome in an automated system. For example, in the work from Tian et al. [16], automated filter-based segmented masks result an approximately 10% decrease in accuracy when predicting both Plus and Normal classes, compared to using masks manually generated by experts. To that end, we aim to obtain

PETRA '21, June 25 -July 2nd, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3454002

a fully automated model that learns from both class and comparison labels, without sacrificing any prediction accuracy compared to using manually segmented images. In this paper, we follow the same feature extraction method from Ataer-Cansizoglu et al. [1] and extend the work from Tian et al. [16] to use U-Net [14] generated masks. We show that by replacing the filter-based masks with U-Net masks, our model can even outperform the model using manually segmented masks, in predicting class labels, comparisons labels and Reference Standard Diagnostic (RSD) labels, which is produced by consensus decision from multiple experts. This happens despite the fact that the model we employed is shallow and is trained on highly interpretable features. For this reason, establishing a pipeline that involves automatic segmentation while maintaining good prediction is important. As shown in Fig. 1, considering the manual mask as the ground-truth, a filter-based mask is much noisier than a U-Net mask.

2 RELATED WORK

Several works focus on developing automated diagnostic system for ROP. Ataer-Cansizoglu et al. [1] extract features based on manually segmented masks and train Support Vector Machine (SVM) classifiers on RSD labels to predict Plus and Normal classes. We use the same feature extraction pipeline with Ataer-Cansizoglu et al. [1] but we depart by 1) using U-Net segmentations and 2) training on both noisy class and comparison labels instead of RSD labels. Tian et al. [16] propose using severity scores learned from both class and comparison labels and introduce expert-biases to capture the noise in class labels. However, despite manual segmented features reaches an AUC of 0.97, filter-based features only have a 0.90 AUC when predicting Plus disease. This motivates us to replace the filter-based segmentation with U-Net segmentation [14].

Existing works also use neural networks to classify ROP. To predict Plus disease, Brown et al. [4] achieve a 0.98 AUC by training a GoogLeNet architecture [15] with U-Net segmentations as inputs and RSD labels as targets. After comparing several classifiers, Yildiz et al. [18] use a Multi-layer Perceprons to reach an AUC of 0.99 when predicting Plus class. Both works do not involve comparisons. By incorporating comparison labels, Yıldız et al. [17] train a 5 million parameter neural network with only 80 images and such model can achieve an AUC of 0.92. We depart from all of these works using neural networks by 1) directly tackling the noise in class labels instead of training on consensus RSD labels, 2) using a shallow classifier. This has the benefit of interpretability, as features we extract are directly related to tortuosity and dilation [1].

3 PROBLEM FORMULATION

Our goal is to provide an automated model producing a severity score for ROP. Such model regresses automatically generated ROP image features with both class and comparison labels. To obtain ROP images features, our method consists of three steps: segmentation, tracing, and feature extraction. We compare different segmentation methods while following the same process of Tracing and Feature Extraction from Ataer-Cansizoglu et al. [1].

3.1 Notation

We follow the notation of Tian et al. [16]. We consider a dataset of N images, indexed by $i \in \mathcal{N} = \{1, 2, \cdots, N\}$ and E experts, indexed by $e \in \{1, 2, \cdots, E\}$. We denote by $\boldsymbol{p} \in \mathbb{R}^2$ the index of a pixel on an image. Assuming all images have the same size, let Ω be the set of indices of pixels on an image. We use pixels in an image to generate segmented mask, which is further processed for computing interpretable features. The method of segmentation and feature extraction are explained in Section 3.2 and Section 3.3, respectively. We consider two types of labels: class labels, representing the diagnosis of an ROP image, and comparison labels, characterizing the ordering of relative severities between two ROP images. We denote class label set as D_a containing tuples of the form (i, e, y_i^e), where $y_i^e \in \{-1, +1\}$ is the class label generated by expert e. We denote comparison label set as D_c containing tuples of the form $(i, j, e, y_{i,j}^e)$ where $y_{i,j}^e \in \{-1, +1\}$ is the comparison label generated by expert e when ordering the relative severities between images i and j. We assume that $y_{i,j}^e = +1$ if and only if the severity of image i is higher than it of image j and $y_{i,j}^e = -1$ o.w.

3.2 Segmentation

Segmentation converts a colored ROP image to a binary mask. We use three different segmentation methods: manual segmentation, filter-based segmentation, and U-Net segmentation. In manual segmentation, an ROP expert draw the binary mask by manually selecting the pixels on important vessels. Fig. 1b is an example of manual segmentation result. Compared to the original image in Fig. 1a, many tiny and unimportant vessels are removed in the manual segmentation.

U-Net Segmentation [4, 14]. Given the excellent performance of U-Net Segmentation in Brown et al. [4], we employ U-Net Segmentation in our pipeline. For each ROP image, we first generate multiple smaller patches by randomly selecting the center of the patches in the full image. Then we train U-Net with the patches as inputs and their corresponding manually segmented masks as targets. The U-Net is designed to output a value between 0 and 1 for each pixel, which represent the probability of the pixel on the vessel. For the pixels overlapped by multiple patches, we average their output values. Fig. 1d is an example of U-Net mask, which is similar to the manually segmented mask.

Filter-based Segmentation [1]. We first obtain a gray scale image by applying anisotropic diffusion [13] to the green channel of the original image. Then we employ unsharpening filter and enhance the image based on the global mean and variance. After adaptively equalizing the histogram, we filter the image with Frangi filter [6].

Let $f(\boldsymbol{p}) : \mathbb{R}^2 \to \mathbb{R}$ represent the probability that a pixel $\boldsymbol{p} \in \mathbb{R}^2$ is on a vessel. Specifically, the function f is denoted as: $f(\boldsymbol{p}) = \sum_{i \in \Omega} w(\boldsymbol{p}_i) G_{\Sigma_i}(\boldsymbol{p} - \boldsymbol{p}_i)$, where Σ_i is the covariance of the Gaussian kernel $G_{\Sigma_i}(\boldsymbol{p}_i) = C_{\Sigma_i} \exp(-\frac{1}{2}) \boldsymbol{p}_i^T \Sigma_i^{-1} \boldsymbol{p}_i$ and C_{Σ_i} is the normalization parameter. The filterd images provides the weight $w(\boldsymbol{p}_i)$ of each pixel $\boldsymbol{p}_i$. Let $\boldsymbol{H}(\boldsymbol{p})$ be Hessian of f and let $\{(\lambda_i(\boldsymbol{p}), \boldsymbol{q}_i(\boldsymbol{p}))\}_{i=1}^2$ be the eigenvalue-eigenvector pairs of $\boldsymbol{H}(\boldsymbol{p})$. Assuming that $|\lambda_1(\boldsymbol{p})| \geq |\lambda_2(\boldsymbol{p})|$, if $\lambda_1(\boldsymbol{p})$ is negative (non-negative), the pixel $\boldsymbol{p}$ is around (away from) the vessel. We generate a binary mask by selecting the pixels that have negative λ_1. To remove some trivial areas, we further postprocess the binary mask by 1) discarding pixels that is less than $\gamma \times \max_{i \in \Omega} f(\boldsymbol{p}_i)$ where γ is set to 0.0015, and 2) removing all isolated areas less than 100 pixels. Fig. 1c is an example of filter-based segmentation, which is more noisy than manually segmented mask and U-Net mask.

3.3 Tracing and Feature Extraction

After segmentation, we obtain a binary mask where pixels on vessels are selected. Then we use this mask to find the vessel centerlines in tracing [1]. Finally, we use the centerlines to calculate 12 features related to tortuosity and dilation, e.g., vessel width and curvature. For each feature, we compute 13 statistics including maximum, minimum, average, standard deviation etc [1]. As a result, for each image $i \in \mathcal{N}$, we construct a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ where $d = 156$.

3.4 Combining Class and Comparison Labels

We follow the same modeling process from Tian et al. [16], where Bradley-Terry model [3] is used. Bradley-Terry model assumes that each image i is associated with a score $s_i \geq 0$ and all comparison labels $y^e_{i,j} \in D_c$ are independent. Given two images i and j, the probability that image i ranks higher than image j is given by: $P(y^e_{(i,j)} = +1) = s_i/(s_i + s_j), \forall (i, j, e, y^e_{(i,j)}) \in D_c$. We further assume that there exist a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$ and expert-dependent biases $b_e \in \mathbb{R}$ so that $s_i = e^{\boldsymbol{\beta}^T \boldsymbol{x}_i + b_e}$. According to the experiments from Tian et al. [16], introducing expert bias term b_e captures the noise in class labels, and provides the best results in predicting all types of labels. Under such assumptions, given D_c, we learn the parameter $\boldsymbol{\beta}$ via minimizing $L_c(\boldsymbol{\beta}; D_c) = \sum_{(i,j,e,y^e_{(i,j)}) \in D_c} \log\left(1+\exp\{-y^e_{(i,j)}(\boldsymbol{\beta}^T(\boldsymbol{x_i}-\boldsymbol{x_j}))\}\right)$. This is a logistic regression model of the feature difference $\boldsymbol{x}_i - \boldsymbol{x}_j$.

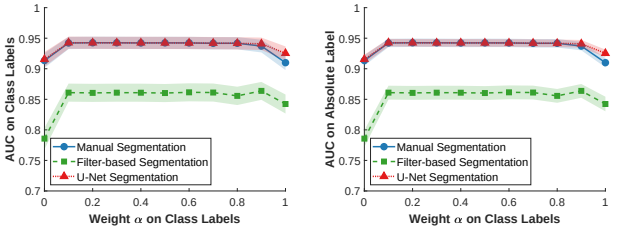
We also assume that the class labels are generated by a logistic model with the same parameter $\boldsymbol{\beta}$ from the comparison labels. As a result, we write the joint loss function combining class labels and comparison labels as follows: $L(\boldsymbol{\beta}; D_a, D_c) = \alpha L_a(\boldsymbol{\beta}; D_a) + (1 - \alpha)L_c(\boldsymbol{\beta}; D_c) + \lambda\|\boldsymbol{\beta}\|_1$, where L_a is a logistic loss for class labels, $\alpha \in [0, 1]$ is a balancing parameter between losses of class labels and comparison labels, and $\lambda \in \mathbb{R}^+$ is the regularization parameter. When $\alpha = 0$, we only use comparison labels and when $\alpha = 1$, we only use class labels.

4 EXPERIMENTS

4.1 Dataset

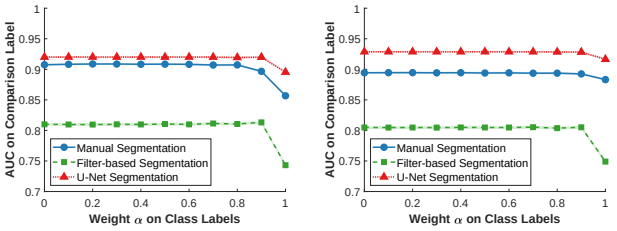
The ROP dataset contains 100 images (size of 640×480) and each image is independently labeled by 13 experts as Plus, Pre-Plus or Normal. Thus, our class labels dataset contains $1,300$ class labels. We binarize these class labels as Plus vs not Plus and Normal vs not Normal. There are 5 experts among the 13 experts additionally providing comparison labels. All 5 experts labeled all possible pairs of comparisons and some pairs are labeled more than once by all 5 experts. As a result, our comparison dataset contains $29,705$ comparison labels. We also obtain Reference Standard Diagnostic (RSD) labels for all 100 images via the following process: all 100 images receive a clinical diagnosis provided by an ophthalmologist who fully evaluate the retina of the patient. Then three experts independently label the 100 ROP images. If the majority label from three experts is consistent with the clinical diagnosis, the RSD label is assigned to the clinical diagnosis; if they differs, three experts discuss and decide the RSD label. According to RSD labels, this dataset contains 15 Plus, 31 Pre-Plus and 54 Normal images[1].

[1]Our code is available at https://github.com/neu-spiral/ROP_UNetSeg_Comparison



(a) Plus vs Not Plus (b) Normal vs Not Normal

Figure 2: Different segmentations on predicting class labels. The shaded area shows the standard deviation.



(a) Plus vs Not Plus (b) Normal vs Not Normal

Figure 3: Different segmentations on predicting comparison labels. The shaded area shows the standard deviation.

4.2 Experimental Setup

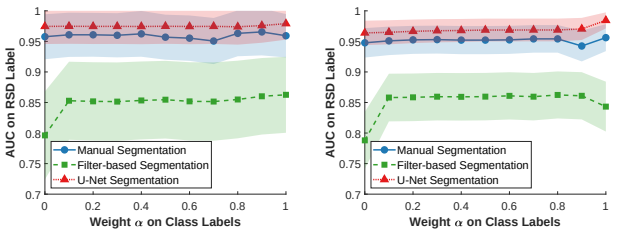
We evaluate our model performances over 5-fold cross-validaition. We guarantee that there is no overlapping images between training and testing set, even considering comparison pairs. We use the area under the receiver operating characteristic curve (AUC) as an evaluation metric. We repeat the experiments 10 times with different cross-validation partitions, and we report the average AUC. We also report the standard deviation of AUC [7].

4.3 U-Net Segmentation Experimental Setup

The dataset used for U-Net segmentation includes the ROP dataset, and additional 100 ROP images as well as their manually segmented masks. We generate all $500,000$ patches (size of 48×48) from these 200 images. As the dataset is imbalanced, the training set is augmented 8 times by 90 degree rotations and both vertical and horizontal flips. Then the augmented set is randomly sampled so that the number of patches from each class equals.

4.4 Evaluation

We compare the prediction performance under three different segmentation methods: manual segmentation, filter-based segmentation [16], and U-Net segmentation. Fig. 2 plots AUC on class labels as a function of the balancing parameter α under different segmentation methods. In Fig. 2a, for all α values, filter-based segmentation performs worse than manual segmentation and in most cases, U-Net segmentation performs as good as manual segmentation. This is because the filter-based segmentation is much noisy compared to manual segmentation and U-Net segmentation (see Fig. 1). When $\alpha = 1$, U-Net segmentation outperforms manual segmentation. For all segmentation methods, the best AUC is obtained when $\alpha \in (0, 1)$, which indicates that combining both class and comparison labels



(a) Plus vs Not Plus (b) Normal vs Not Normal

Figure 4: Different segmentations on predicting comparison labels. The shaded area shows the standard deviation.

Table 1: Best AUC between ROP Systems. The Deep, Comparison and RSD columns show whether deep neural networks, comparisons and RSD labels, respectively, are used.

ROP System	AUC	Deep	Comparison	RSD
Tian et al. [16] Manual Seg.	0.97	No	Yes	No
Tian et al. [16] FB Seg.	0.90	No	Yes	No
Yildiz et al. [18]	**0.99**	Yes	No	Yes
Yıldız et al. [17]	0.93	Yes	Yes	Yes
Brown et al. [4]	0.98	Yes	No	Yes
This work	0.98	No	Yes	No

improves the testing AUC. These trends are also observed in Fig. 2b when predicting Normal vs Not Normal category.

Fig. 3 shows AUC on comparison labels as a function of parameter α. Similar to Fig. 2, for all α, filter-based segmentation performs worst. Surprisingly, in all cases, U-Net segmentation outperforms manual segmentation. Even though manually segmented masks are used as targets when training U-Net segmentation, the manually segmented masks contain noises. The CNN model is capable to capture the noise in manual segmentation and reduces the noise in testing. Thus the prediction results from U-Net segmentation outperforms those from manual segmentation. Compared to Fig. 3a, the gap between U-Net segmentation and manual segmentation is even larger when predicting Normal class in Fig. 3b. For all segmentations, compared to the AUC at $\alpha = 1$, AUC improves when $\alpha \in [0, 1)$. This demonstrates that incorporating comparison labels in training improves the prediction on comparison labels.

Fig. 4 shows AUC on RSD labels as a function of the balancing parameter α. Compared to manual segmentation, using filter-based segmentation can cause an AUC decreasing as much as 15%. However, in all cases, U-Net segmentation can still outperforms manual segmentation even the latter reaches an AUC of 0.95.

In Table 1, we report the best AUC of predicting RSD labels, achieved by several works. Although Tian et al. [16] achieve an 0.97 AUC by manual segmentation, the prediction AUC drops to 0.90 when filter-based (FB) segmentation is used. Yıldız et al. [17], Yildiz et al. [18], and Brown et al. [4] use U-Net segmentations; however, all of them train neural networks as classifiers, thus it is difficult to interpret the importance of features. We use a shallow model in this work to obtain the interpretabilities while maintaining the prediction AUC almost the same with the best deep model.

5 CONCLUSIONS

We extend Tian et al. [16] by replacing manual segmentations with U-Net segmentations. Our experimental results show that in many cases, features using U-net segmentation outperforms manually segmented features. Moreover, compared to other neural networks models, our model achieves almost same AUC while maintaining the interpretability of ROP features.

ACKNOWLEDGMENTS

Our work is supported by NIH (R01EY019474), NSF (SCH-1622542 at MGH, SCH-1622536 at Northeastern, SCH-1622679 at OHSU), a Facebook Research Award, and by unrestricted departmental funding from Research to Prevent Blindness (OHSU).

REFERENCES

[1] Esra Ataer-Cansizoglu, Veronica Bolon-Canedo, J Peter Campbell, Alican Bozkurt, Deniz Erdogmus, Jayashree Kalpathy-Cramer, Samir Patel, Karyn Jonas, RV Paul Chan, Susan Ostmo, et al. 2015. 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. *TVST* 4, 6 (2015), 5–5.

[2] Hannah Blencowe, Joy E Lawn, Thomas Vazquez, Alistair Fielder, and Clare Gilbert. 2013. Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010. *Pediatric* 74 (2013), 35–49.

[3] Ralph Allan Bradley and Milton E Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952).

[4] James M Brown, J Peter Campbell, Andrew Beers, Ken Chang, Susan Ostmo, RV Paul Chan, Jennifer Dy, Deniz Erdogmus, Stratis Ioannidis, Jayashree Kalpathy-Cramer, et al. 2018. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA ophthalmology* 136, 7 (2018), 803–810.

[5] Cryotherapy for Retinopathy of Prematurity Cooperative Group and others. 1988. Multicenter Trial of Cryotherapy for Retinopathy of Prematurity: Preliminary Results. *Archives of Ophthalmology* 106, 4 (1988), 471–479.

[6] Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever. 1998. Multiscale vessel enhancement filtering. In *MICCAI*. Springer, 130–137.

[7] James A Hanley and Barbara J McNeil. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 143, 1 (1982).

[8] M Elizabeth Hartnett and John S Penn. 2012. Mechanisms and Management of Retinopathy of Prematurity. *New England Journal of Medicine* 367, 26 (2012).

[9] Ann Hellström, Lois EH Smith, and Olaf Dammann. 2013. Retinopathy of Prematurity. *The lancet* 382, 9902 (2013), 1445–1457.

[10] International Committee for the Classification of Retinopathy of Prematurity. 2005. The international classification of retinopathy of prematurity revisited. *JAMA Ophthalmology* 123, 7 (2005), 991–999.

[11] Jonathan Javitt, Ronald Dei Cas, and Yen-pin Chiang. 1993. Cost–Effectiveness of Screening and Cryotherapy for Threshold Retinopathy of Prematurity. *Pediatrics* 91, 5 (1993), 859–866.

[12] Jayashree Kalpathy-Cramer, J Peter Campbell, Deniz Erdoğmuş, Peng Tian, Dharanish Kedarisetti, Chace Moleta, James D Reynolds, Kelly Hutcheson, Michael J Shapiro, Michael X Repka, et al. 2016. Plus Disease in Retinopathy of Prematurity: Improving Diagnosis by Ranking Disease Severity and Using Quantitative Image Analysis. *Ophthalmology* 123, 11 (2016), 2345–2351.

[13] Pietro Perona and Jitendra Malik. 1990. Scale-space and edge detection using anisotropic diffusion. *TPAMI* 12, 7 (1990), 629–639.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 234–241.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. 1–9.

[16] Peng Tian, Yuan Guo, Jayashree Kalpathy-Cramer, Susan Ostmo, John Peter Campbell, Michael F Chiang, Jennifer Dy, Deniz Erdogmus, and Stratis Ioannidis. 2019. A severity score for retinopathy of prematurity. In *KDD*. 1809–1819.

[17] İlkay Yıldız, Peng Tian, Jennifer Dy, Deniz Erdoğmuş, James Brown, Jayashree Kalpathy-Cramer, Susan Ostmo, J Peter Campbell, Michael F Chiang, and Stratis Ioannidis. 2019. Classification and comparison via neural networks. *Neural Networks* 118 (2019), 65–80.

[18] Veysi M Yildiz, Peng Tian, Ilkay Yildiz, James M Brown, Jayashree Kalpathy-Cramer, Jennifer Dy, Stratis Ioannidis, Deniz Erdogmus, Susan Ostmo, Sang Jin Kim, et al. 2020. Plus Disease in Retinopathy of Prematurity: Convolutional Neural Network Performance Using a Combined Neural Network and Feature Extraction Approach. *TVST* 9, 2 (2020), 10–10.