applied sciences

MDPI

Article

A Fully Automated Pipeline for a Robust Conjunctival Hyperemia Estimation

Nico Curti [1,†], Enrico Giampieri [1,†], Fabio Guaraldi [2], Federico Bernabei [2], Laura Cercenelli [1,*], Gastone Castellani [3], Piera Versura [2,‡] and Emanuela Marcelli [1,‡]

1 eDIMESLab, Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, 40138 Bologna, Italy; nico.curti2@unibo.it (N.C.); enrico.giampieri@unibo.it (E.G.); emanuela.marcelli@unibo.it (E.M.)
2 Ophthalmic Unit, Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, 40138 Bologna, Italy; fabioguaraldi90@gmail.com (F.G.); federico.bernabei89@gmail.com (F.B.); piera.versura@unibo.it (P.V.)
3 Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, 40138 Bologna, Italy; gastone.castellani@unibo.it
* Correspondence: laura.cercenelli@unibo.it
† Both authors contributed equally to this work.
‡ These authors share seniorship.

Abstract: Purpose: Many semi-automated and fully-automated approaches have been proposed in literature to improve the objectivity of the estimation of conjunctival hyperemia, based on image processing analysis of eyes' photographs. The purpose is to improve its evaluation using faster fully-automated systems and independent by the human subjectivity. Methods: In this work, we introduce a fully-automated analysis of the redness grading scales able to completely automatize the clinical procedure, starting from the acquired image to the redness estimation. In particular, we introduce a neural network model for the conjunctival segmentation followed by an image processing pipeline for the vessels network segmentation. From these steps, we extract some features already known in literature and whose correlation with the conjunctival redness has already been proved. Lastly, we implemented a predictive model for the conjunctival hyperemia using these features. Results: In this work, we used a dataset of images acquired during clinical practice.We trained a neural network model for the conjunctival segmentation, obtaining an average accuracy of 0.94 and a corresponding IoU score of 0.88 on a test set of images. The set of features extracted on these ROIs is able to correctly predict the Efron scale values with a Spearman's correlation coefficient of 0.701 on a set of not previously used samples. Conclusions: The robustness of our pipeline confirms its possible usage in a clinical practice as a viable decision support system for the ophthalmologists.

Keywords: artificial intelligence; computer aided diagnosis; computer vision; conjunctiva; hyperemia; Efron scale

check for updates

Citation: Curti, N.; Giampieri, E.; Guaraldi, F.; Bernabei, F.; Cercenelli, L.; Castellani, G.; Versura, P.; Marcelli, E. A Fully Automated Pipeline for a Robust Conjunctival Hyperemia Estimation. *Appl. Sci.* **2021**, *11*, 2978. http://doi.org/10.3390/app11072978

Academic Editor: Yahan Hu

Received: 27 February 2021
Accepted: 23 March 2021
Published: 26 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The estimation of conjunctival hyperemia is a standard procedure during clinical evaluation in ophthalmology. Conjunctival redness or hyperemia is evaluated on the dilation of blood vessels in the conjunctiva area, and it can be symptomatic of different kinds of inflammations or infections. Its evaluation is commonly investigated for the contact lens effects [1,2] and for the so-called dry-eyes syndrome [3]. The hyperemia estimation is performed on images of the patient's eye acquired using a slit lamp. The slit lamp, also called slit bio-microscope, is an optical instrument used in ophthalmology for the observation of eye tissues. It allows for visualizing the bulb and the ocular annexes, the corneal layers, the vitreous and the anterior chamber, the crystalline lens and the iris. All of this information can be acquired posing no risks for the patient.

Hyperemia is commonly evaluated using qualitative grading scales, where the conjunctival hyperemia is compared to a set of standardized pictures (template images or references). These reference pictures are images artificially generated using software, painted, or real photographs. The clinician establishes the score by matching the patient conjunctiva to the reference one that he considers the most similar. This measure is performed for each eye independently.

There are many grading scales proposed in literature for the hyperemia estimation. Davies (1978), Mandell (1989), and Woods (1989) proposed descriptive grading scales; Koch et al. (1984) and Schnider (1990) proposed art illustrations for the evaluation of single condition grading scales; Annunziato et al. (1992) and Efron (1999) extend the previous evaluation to multiple conditions; Courtney & Lee (1982), Lupelli (1998), McMonnies & Chapman Davies (1987), Price et al. (1982), Begley (1992), and Lofstrom et al. (1998) introduce the use of photographic references. The latest proposed reference images were generated using computer graphic (Jenvis, 2009). Each scale is associated with a different set of reference images and each of them is focused on a different aspect of the hyperemia estimation; for instance, the conjunctival redness in the Efron Grading Scales for Contact Lens Complications [1] is expressed through five images depicting 1–5 grading ranging from normal to severe.

A critical issue of these assessments is their reproducibility: the template matching evaluation is an intrinsic operator-dependent procedure and thus it is affected by subjectivity. The reproducibility of the image acquisition and moreover the hyperemia evaluation are determined by the experience of the operating clinician, and it can be affected by several experimental conditions, leading to imprecisions and biases [4,5]. Generally, the automation of the clinical exam provides a support for the clinical laboratory or clinical practice assessment through faster and more reliable evaluation and reducing the reliance on the operator expertise for the evaluation [6].

Several authors have already proposed semi-automated grading systems for the hyperemia quantification, offering semi-supervised pipelines to process the patient's images [7]. All of these methods facilitate the clinician operations providing a series of automated processing steps, which, however, require the human intervention (e.g., a manual selection of the region of interest and color values) [3,8–14]. At the same time, they allow for introducing new quantitative grading scales that are difficult to relate to the standard clinical assessments. These grading scales can provide a viable medical alternative to standard grading scales ensuring greater accuracy and different information for the clinician, but they can not be used as standalone results prior to a lengthy medical certification trial.Thus, despite the efficiency of the automated methods, their practical usage by clinicians remains limited.

The creation of a fully automated pipeline which starts from the acquired image until the hyperemia score prediction is still an open problem [15]. Some authors have already proposed interesting results on this problem. Sánchez Brea et al. [16] implemented a fully automated pipeline for the segmentation of the conjunctiva, but the predictions of their model were obtained only on a small part of the dataset on which there was a reasonably good agreement between experts' evaluations, discarding any source of issues, that might be instead encountered during clinical practice. Derakhshani et al. [17] tried to predict the assessment of the vascularity of conjunctiva using a neural network approach: the images were rescaled to improve the computational efficiency, losing a great part of the information; moreover, the results reported in their work are not easily interpretable, since the final scores are obtained without a subdivision of the samples in train and test, alongside a best-model selection procedure on the same data.

All of these methods introduced the usage of modern machine learning and deep learning frameworks for the processing of the features extracted from the original images, after a semi-automated segmentation of the region-of-interests (ROIs). Deep learning convolutional neural network (CNN) image segmentation models have shown promising results in medical applications in the last few years. We can find classical or tailored deep

learning architectures in many medical research fields, including the opthalmologic one [18]. Despite the growth in computational power availability, which allows for enlarging the application of even more complex deep learning models, their usage necessarily requires the manual annotation performed by experts.

In this work, we proposed a fully-automated pipeline for the processing of slit lamp images and hyperemia quantification. The most informative portion of the image for the hyperemia estimation in these images is given by the vessels' network structure [11] and thus it is crucial to extract this component from the whole image. Our pipeline includes the segmentation of the conjunctival area using CNN, the segmentation of the vessels network using image processing algorithms, the extraction of a set of features related to the redness and the prediction of the Efron scale grading value using a Ridge regression model [1]. The full automation of the pipeline allows for including features such as the vessels network structure that would be normally hard to quantify due to the time requirement of the manual segmentation. In this work, we evaluated the performances of this pipeline using a set of features tailored for the Efron scale quantification, as already discussed in literature.

2. Materials and Methods

2.1. Patient Selection

The analyzed images had been obtained during routine ophthalmological examinations in the Ophthalmological Unit at IRCCS S. Orsola University Hospital of Bologna. Images were retrieved from charts of subjects who gave their voluntary consent to research. The study was approved by the Local Ethics Committee, and carried out in accordance with the Declaration of Helsinki.

We collected 70 patients and for each of them the images of both eyes were acquired (Grading Dataset, GD). Using a slit lamp, we obtained 2 images for each eye, the first taken from the nasal part of the eye and the second one from the temporal bulbar conjunctiva. In this way, we obtained a full set of 280 images for our analysis. Individuals between 20 and 50 years of age were selected for the current study. The patients were selected by an heterogeneous population, and thus the dataset includes samples with high and low redness levels. A global description of the dataset is showed in Table 1.

Table 1. Descriptive statistics of the analyzed samples in the dataset. Age and Efron grading values are reported as mean ± standard deviation.

Grading Dataset			
	# of Samples	**Average Age**	**Average Efron**
Male	23	48.5 ± 12.3	2.04 ± 0.81
Female	47	52.3 ± 10.3	1.81 ± 0.79
Total	70	51.4 ± 10.5	1.81 ± 0.86

2.2. Slit Lamp Images

The photos had been taken by two trained clinicians with a digital slit lamp microscope (Topcon SL-D4 slit-lamp biomicroscope). For all the participants, the following protocol was used for photography: 1. We illuminated with a halogen lamp, with a 7.5 KLux illumination input to the slit lamp (both red-free and diffuser filters excluded) and with a wide diffuse illumination of a slit lamp (8 mm circle); 2. We adopted a magnification of 16×; 3. We used a 14 mm diaphragm aperture; 4. We set the sensibility of the digital image to 100 ISO; 5. We set the acquisition time to 1/80 s. All the photographs were acquired under similar room illumination.

For taking images of the nasal and temporal bulbar conjunctiva, each participant was instructed to look horizontally left and right. Gentle pressure was applied to open the lids in order to ensure that they did not obstruct the conjunctiva during the photography. Photographs were taken without flash and quickly to avoid dry eye and irritation.

The images of the GD were collected during clinical practice by the clinicians without standardized parameters of acquisition (such as brightness): each clinician acquired the image according his best judgment, as it is standard clinical procedure.

All of the images were captured in a raw format, i.e., RGB 8-bit, and saved in JPEG format (2576 × 1934, 150 dpi, 24 bit).

2.3. Clinical Scoring of Images

Three trained clinicians performed the evaluation of the full set of 280 images, independently. The clinicians scored each image according to the Efron grading scale: despite the 5 possible values of the Efron scale (from normal, 1, to severe, 5), we allowed the usage of intermediate values in case of doubt. We have chosen the Efron scale as reference since it is a standard reference in ophthalmology and its automation can easily encourage the clinicians community to use our method.

All of the clinicians scored the images in the same physical space, with the same source of illumination and without time limits. Two computer monitors (HP Z27 UHD 4K, 27″, 3840 × 2160 resolution) were used: the first displayed the grading scale images (reference) and in the second one the clinical images were showed. In both of the monitors, the same screen color and brightness were used for all the three clinical evaluations.

We collected the evaluation of the three trained clinicians on the full set of images, and we kept the median value as ground truth score for the analysis.

2.4. Image Processing Pipeline

Our image processing pipeline is composed by a series of independent and fully automated steps (ref. Figure 1):

1. conjunctiva segmentation;
2. vessels network segmentation;
3. vessels network features extraction and color features extraction;
4. Efron scale values prediction.

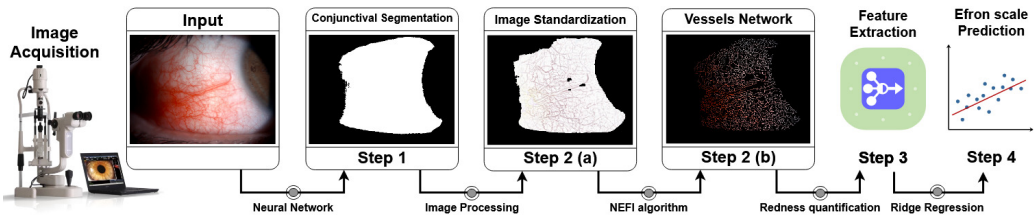

Figure 1. Schematic representation of the pipeline. (**Step 1**) The image acquired by the slit lamp is used as input for the Neural Network (Mobile U-Net) model for the segmentation of the conjunctival area. The model was trained using manually annotated images validated by experts. (**Step 2**(**a**)) Focusing on the conjunctival area the image is standardized using a brighter-correction algorithm given by a combination of median filters and background subtraction. (**Step 2**(**b**)) The image standardization helps us to remove possible artifacts and color issues related to the non-rigid image acquisition procedure. The processed image is used as input for the vessels network segmentation algorithm given by a tuned implementation of the NEFI algorithm. (**Step 3**,**4**) Starting from the vessels' network image, a set of features for the quantification of the conjunctival redness are extracted and used for the development of a predictive model. The full set of steps are performed automatically and thus without any human intervention.

The first step of processing involves the segmentation of the conjunctiva area from the background, i.e., skin, eyelids, eyelashes, and caruncle. For each image of the datasets, we collected a manual annotation of the conjunctiva area performed and validated by the 3 experts. The manual annotation set includes a binary mask of the original image in which only the conjunctiva area is highlighted. The conjunctiva segmentation allows for

reducing the region of interest of our analysis on the most informative area for the redness evaluation.

Several already published automated grading systems for the hyperemia quantification perform this step using semi-supervised methods, leaving to the final user the manual selection of the region of interest (ROI) [3,8–10]. Our method automatizes this task using a semantic segmentation neural network model trained on a set of images manual annotated and validated by experts.

In the second step, the pipeline performs a second segmentation for the vessels' network identification, starting from the segmented conjunctiva. The vessels network is automatically segmented using a customized version of the NEFI [19] algorithm. The vessels network segmentation provides a mask to apply to the original image.

In step three, the pipeline extracts features starting from the selected areas based on quantities proposed in the literature and based on different color spaces (RGB and HSV, i.e., Hue, Saturation, and Value). In step four, the extracted set of features was used to feed a penalized regression model for the prediction of the final Efron scale value.

2.4.1. Step 1—Conjunctiva Segmentation

In literature, there are several deep learning models proposed to automate the segmentation of eyes' images (e.g., optical coherence tomography [20], sclera segmentation [21–23], retinal vessel [24]); nevertheless, we could not find any model focusing on slit lamp images. These images are theoretically easy to segment since the purpose is to isolate the white-like part of the image (conjunctiva) from the red-like background (skin, eyelids, eyelashes, and caruncle), but in reality the differences between these two regions are not uniform in the different parts of the same image and not as well defined in the color space. An image thresholding [10] or an image quantization [9] could solve this task in the simplest cases, but they can not take care of the extremely variability of the samples provided by a clinical acquisition. There is not a rigid standardized procedure in the image acquisition by a slit lamp during the clinical acquisition and the "elaboration" of the images is left to the experience of the clinicians.

In our work, we divide the available image samples into a training (42 patients, 60% of the full set of images, i.e., 168 images) and a validation set (28 patients, 40% of the full set of images, i.e., 112 images). From the training set, we excluded a subset of 18 images, i.e., 9 patients, as test set for the evaluation of the model performances over the training procedure. For each image, we collected a manual annotation of the conjunctiva area performed and validated by 3 experts. The manual annotation set includes a binary mask of the original image in which only the conjunctiva area is highlighted; this set of images-masks was used as ground truth for our deep learning model.

The amount of the available samples does not justify the usage of a complex deep learning architecture with a large amount of parameters to tune. During the research exploration, we tried several CNN architectures commonly used in segmentation tasks, starting from DenseNet CNN to the lighter U-Net variants [25,26]. The evaluation of model performances has to balance both a good performance on the validation set and a greater ability of extrapolation on new possible samples. We would like to stress that, despite the above requirements being commonly looked for in any deep learning application, they are essential for any clinical application, in which the variability of the samples is extremely high. We remark that even high accuracy scores could still include distortions and artifacts of clinical significance: the most important result is the clinician's evaluation, i.e., the visual segmentation accuracy estimated by experts. All of the predicted images were carefully evaluated by the experts of the Ophthalmology research group of the IRCCS S. Orsola University Hospital Ophthalmic Unit, Laboratory for Ocular Surface Analysis of the University of Bologna and their agreement, jointly with the training numerical performances, lead us to choose a Mobile U-Net model with skip connection as the best model able to balance our needs.

We implemented the Mobile U-Net model using the Tensorflow Python library. The model was trained for 100 epochs with an RMSProp optimizer (learning rate of 10^{-4} and decay of 0.995). For each epoch, we monitored the accuracy score, i.e., the average agreement between the mask produced by the model and the ground truth at pixel level, and the Intersection-Over-Union (IoU) score, i.e., the area obtained by the union of the mask produced by the model and the ground truth divided by their intersection area, on each image of the test set. Since the portion of the image occupied by the conjunctiva area could be smaller compared to the full image size, the IoU score is a more informative metric for the evaluation of the model performances since it is robust to unbalances in the sample, differently from the accuracy score. Our training set includes both nasal and temporal images for both the patient eyes. This means that we have an intrinsic vertical flip of the training images. We however performed a large data augmentation procedure to build a more robust model and to hypothesize the possible variability of the test set. For each image, we performed a vertical and/or a horizontal flip jointly with a random rotation.

2.4.2. Step 2—Vessels Network Segmentation

The conjunctiva segmentation allows for reducing the region of interest of our analysis on the most informative area for the redness evaluation. The first step of processing involves a standardization of the images and the correction of the light dependence due to uneven illumination of the conjunctiva. In the same image, different parts of the conjunctiva could have different brightness based on the angle of the incidental light on the conjunctiva and the aperture of the slit lamp diaphragm. This different brightness produces images with a shadow component or with a flash glare. The vessels network in the conjunctiva can thus appear as a brighter or a darker red component, and this can drastically affect the redness evaluation. The images, in fact, tend to appear with a heterogeneous red component when the acquisition is performed in low-light situations. At the same time, the light standardization allows for removing the possible shadows or hyper-intense areas due to the effect of the flash over the patient's eyes. This standardization is performed independently on each channel of the image (R, G, and B), and it involves a median blurring of the channels followed by a normalization of the difference between the blurred and the original channel. The resulting standardized image is converted into grayscale before the application of the next steps.

Starting from the standardized masks, the pipeline performs the vessel segmentation. The vessels' network segmentation is a standard task in ophthalmology image processing, and many studies have already published promising results on this topic [3,9,10,19,27,28]. In our work, we decided to use a customized version of the method proposed in [19], using a watershed adaptive filter to highlight the network vessel component and to segment it from the background. After removing the smaller connected components (up to a predetermined size) to remove noise artifacts, the resulting segmentation is refined using a GuoHall skeletonizer [29]. In this way, only the backbone (only 1 pixel for each branch) of the vessels network is selected. In the next steps, we use the backbone network as a mask for the extraction of the redness features, focusing our analysis on the smallest but most informative region of the image.

2.4.3. Step 3—Redness Features

The redness measurements are standardly performed on the whole area occupied by the conjunctiva, despite the most informative section being given by the only vessels network component. Therefore, for all the features, only the pixels belonging to the vessels network were included.

RGB Redness

Park et al. propose in [9] a redness measure given by a combination of the RGB channels. The authors extract this score on the manually segmented conjunctival area without any preprocessing step, obtaining promising results on the prediction of the

conjunctival redness. The core assumption behind the relation between this feature and the conjunctival redness is the perception of the red color into the conjunctiva area: the mathematical formula tends to emphasize the R intensity using a weighted combination of the three channels. This starting point is generally true in the major part of slit lamp images, but it can suffer from the image light exposition: the RGB channels do not respond in the same way to the image brightness, and this behavior can affect the robustness of the measure.

In our work, we evaluated the same score using the mask provided by the vessels network segmentation: in this way, we can focus our evaluation of the redness only in the most informative area of the conjunctiva, minimizing possible light exposition artifacts. We also improved this measure applying the preprocessing light standardization. Thus, the first feature extracted is the redness score given by

$$score_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2R_i - G_i - B_i}{2 \times (R_i + G_i + B_i)} \tag{1}$$

where we denote with R, G, B the red, green, and blue channels of the standardized image, respectively. The n value represents the amount of pixels in the mask produced by the conjunctival segmentation step.

HSV Redness

A second interesting feature was proposed by Amparo et al. in [8]. The authors suggest to move from the RGB color space to the more informative HSV one, i.e., the hue, saturation, and value. This space is more informative than the RGB one since it is capable of being careful about the different light exposure (saturation). The authors also suggested the usage of a preprocessing step for the slit lamp images, including a white-balance correction to overcome possible light exposition issues. Despite their interesting results, they did not show further results on the dependence of their feature to the same image sampled in different light conditions. Their measure uses a combination of saturation and hue for the redness estimation, thus we can preliminarly expect a more robust behavior to image brightness. In addition, in this case, we applied the same computation for the extraction of a second score on our standardized image after the vessels network segmentation:

$$score_2 = \frac{1}{n}\sum_{i=1}^{n} H_i \times S_i \tag{2}$$

where H and S represent the hue and saturation intensities of the standardized image, respectively.

Fractal Analysis

A third set of measurements were proposed by Schulze et al. in [10], who performed a fractal analysis from the vessels segmented in the conjunctiva. Our technique for the vessels network segmentation is quite different from theirs, but we can, however, apply fractal analysis for the study of the vessels' topology. The vessels network can be compared to a fractal structure and thus analyzed using standard measurements of complex systems analysis [30–32]. Their evaluation can be very informative from a medical point-of-view, and it fills an opened gap in the clinical practice. The estimation of the fractal dimension of the vessels network could be informative of multiple pathologies, since it quantifies the neo-genesis and the ramification of the vessels in the area of interest. Their relation to the hyperemia is straightforward since the greater the area occupied by the vessels, the greater the conjunctival redness. The core assumption in this case is given by the goodness of the segmentation algorithm used for the vessels network extraction: the algorithm should be able to identify the vessels network in any light exposition or the measures can suffer from false positive or missing detections. In particular, we extracted as putative features the fractal dimension of the vessels network using the box-counting and pixel-

counting algorithms. Both algorithms are commonly used methods for the analysis of fractal structures included into images. In our work, we choose a logarithmic set of sizes for the box counting evaluation, and, for each image, the coefficients estimated from the linear interpolation between the box-sizes and the counts were used as features (this line slope is usually referred to as the fractal dimension).

Color Measures

We also computed the averages of the RGB channels $\{\mu_R, \mu_G, \mu_B\}$ and HSV $\{\mu_H, \mu_S, \mu_V\}$ channels of the standardized conjunctival networks, reaching a total of 10 putative features for the redness estimation. There are intrinsic correlations between these features that need to be taken into consideration for the subsequent analyses.

2.4.4. Step 4—Regression Pipeline

The initial step of our regression analysis consists of the standardization of the extracted features. Each feature belongs to a different space/range of values and to combine their values, and we have to rescale all of them into a common range. We rescaled all the features using their median values, normalizing according to the 1st and 3rd quantiles, i.e., a robust scaling algorithm: in this way, we minimize the dependency from possible outliers. Medians and quantiles were estimated on the training set and then applied to the test set to avoid cross contaminations.

The processed set of features is then used in a penalized regression model. We used a penalized ridge regression (or Tikhonov regularization) for the Efron prediction. Ridge regression is a regularized version of the linear regression in which an extra regularization term is added to the cost function, penalizing high values of the coefficients of the regression. In our simulations, we used a penalization coefficient equal to 2.5.

The full set of data was divided into a train/test sets using a shuffled 10-fold cross validation. The model was trained on a subset (90%) of the available samples, and its predictions are compared to the ground truth provided by the corresponding test set (10%).

We want to note that, despite the Efron scale admitting only integer values (from 1 to 5) in our dataset, we have a 8% of floating point values: the experts introduce floating point values when there is no an exact concordance between the patient image and the template image references. We would stress also that the accordance between multiple experts in the Efron scale evaluation is generally low since the measure has an intrinsic subjectivity, and this bounds the best possible predictions of a quantitative model.

3. Results

3.1. Conjunctival Segmentation

The results obtained by our training on the test images are showed in Figure 2a. The masks obtained by the conjunctival segmentation can be applied to the original images to select only the region of interest (ROI) for the following analysis. The results showed in Figure 2b highlight the efficiency of the model in the conjunctiva detection and segmentation. In 100 epochs of training, the model reaches an average accuracy of 0.94 and a corresponding IoU score of 0.88 on the test set.

The visualization of the results confirms the goodness of the training, showing a good agreement with the ground truth masks segmented by the experts. Despite the performances, the visualization of the produced masks also allows an empirical interpretation about what the model has learned during the training: in many cases, the ground truth produced by the experts is "rough", and it does not take care about the superposition of the eyelash on the conjunctival surface. The manual segmentation performed by the experts was also not pixel-perfect, but it is a polygon composed of straight segments and thus the model could not reach a perfect learning (and neither it is desirable in terms of model generalization). The developed model is able to better discriminate between the eyelashes and the conjunctiva showing more accurate masks.

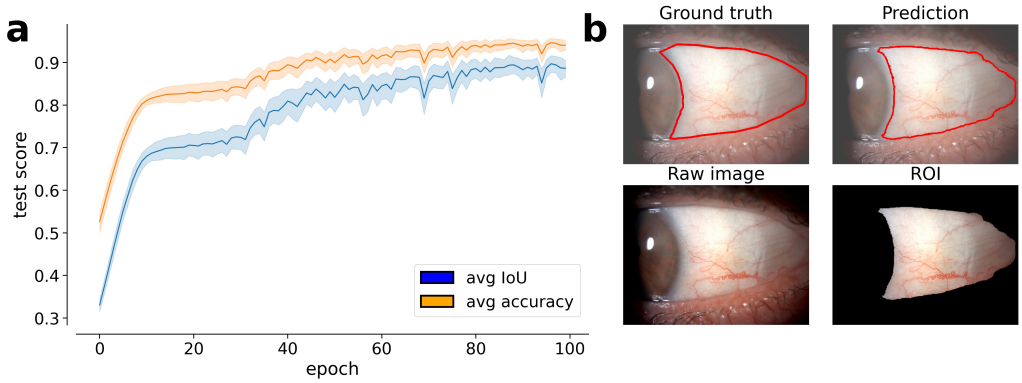

Figure 2. Results obtained by the trained Mobile U-Net model on the 18 test images. (**a**) evolution of the average accuracy and average IoU along the training epochs (100). The IoU score is more informative given that it is robust to unbalances in the sample, differently from the accuracy score; (**b**) example of the results obtained on a test image. On the top-left, the ground truth. On the top-right, the predicted segmentation. On the bottom-left, the raw (input) image. On the bottom-right, the resulting ROI of the conjunctival area.

3.2. Vessels Network Segmentation

An example of this processing is provided in Figure 3. In this case, the red component of the image tends to be uniform along the entire conjunctiva area, making it difficult to visualize the network of vessels. The image standardization algorithm allows for better discriminating the vessels from the background, removing the false positive redness component from the conjunctiva. We use the result of this processing step for both the vessels network extraction and for the next features' evaluation, since the color ranges produced by this step are more robust to the different image expositions. A preliminary analysis of the results obtained by this approach confirms its efficiency: this processing, applied on the standardized image, as discussed above, is powerful enough to extract the most informative vessels from the image. An example of the results obtained by our preprocessing is showed in Figure 3f.

We performed our analyses on a server grade machine (64 GB RAM memory and 1 CPU i9-9900K, with eight cores) and the proposed pipeline, from the conjunctiva segmentation to the vessels network extraction, took less than 2 min per image.

3.3. Efron Scale Prediction

We processed the dataset of 280 samples using our automated pipeline extracting the full set of features. In this dataset, we have two images for each eye (nasal and temporal); thus, we use the average of each feature extracted on the two samples. The full set of 10 features were used for the ridge regression model estimating the correlation between the ground truth Efron values and the predicted ones. We trained the regression model using a 10-fold cross validation: in this way, we can ensure that the model predicts the outcome on a completely novel set of samples. We tuned the pipeline parameters using a grid search algorithm to better fit the available data. The best model found is able to predict the correct Efron scale values with a Spearman's rank correlation coefficient of 0.701 and a corresponding p-value of 10^{-22} (ref. Figure 4a). An example of the prediction obtained on the test set is showed in Figure 5

We apply the same pipeline for 100 different cross validations to test the robustness of our model. In each iteration, a different 10-fold cross validation is provided to the ridge regression: in this way, we test the sensitivity of the model to different training sets. The resulting distribution of Spearman's rank correlation coefficient (ref. Figure 4b) shows the robustness of the developed model: the estimated coefficients are centered around a value of 0.69 with a spread of 0.008.

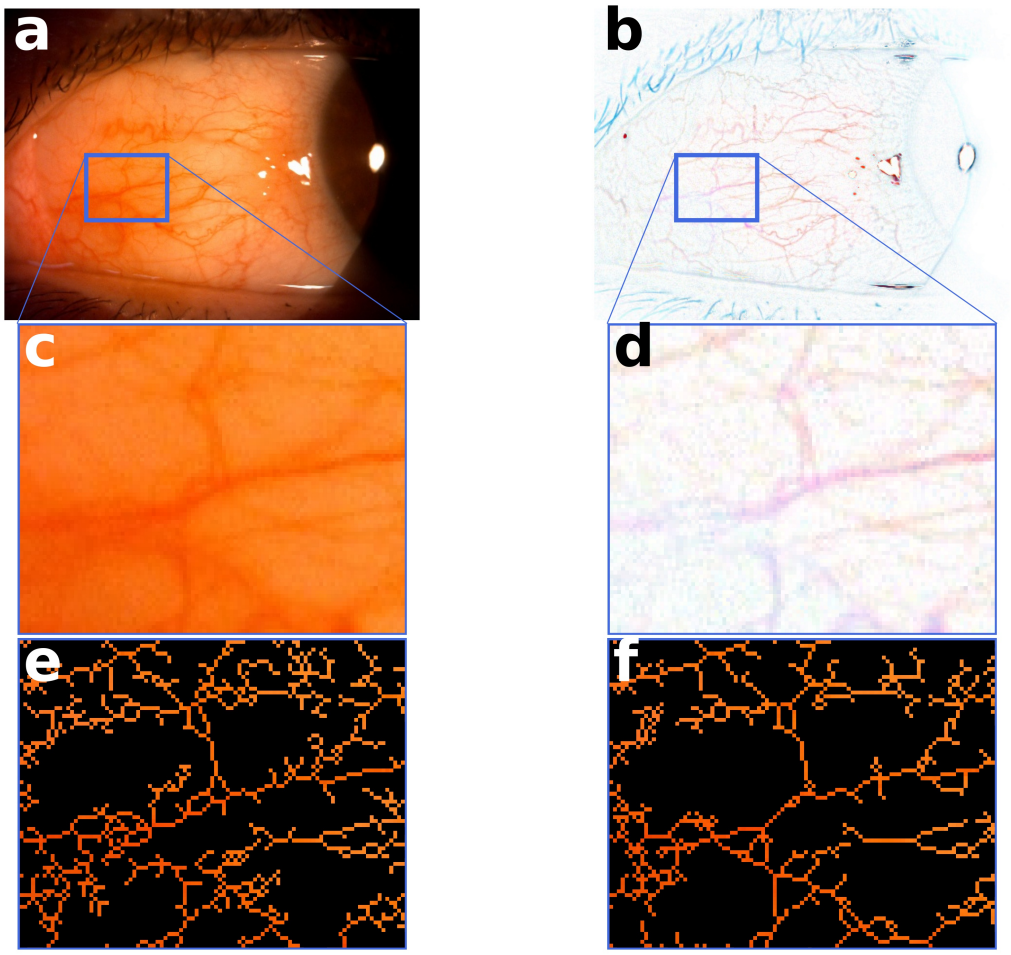

Figure 3. Comparison between the raw image and the result of our standardization algorithm. (**a**,**c**) the original image and corresponding enlarged area. In this case, the image acquisition was performed with an incorrect light exposition, which leads to an unbalanced redness component along the entire conjunctiva area. It is evident in the enlarged section how the vessel color tends to be very close to the background, making it their evaluation difficult and creating a false positive redness component in the conjunctiva. (**b**,**d**) the standardized image and corresponding enlarged area. The standardization algorithm allows for better discriminating the vessels network component from the red-like background. In the enlarged portion, we can see how the algorithm is able to split the vessels with the uniform background, despite the latter having a non-negligible redness component; (**e**) the vessels network segmentation estimated on the raw image; (**f**) the vessels network segmentation estimated on the standardized image. The standardization processing allows for removing the artifacts (false positive branches) related to the red-like background component.

The developed model uses a combination of the extracted features to find the best parameters for the regression model. Despite the resulting performances, we evaluate the informative power of each feature independently. We apply the same cross validation procedure using each feature singularly one by one: in this way, we can assign to each of them its informative power given by the associated Spearman's correlation coefficient. As expected, not all the features are equally informative, but only three of them are associated with a Spearman's correlation coefficient greater than 0.55 (in absolute values). We found as the most informative feature the $score_1$ (ref. Table 2) with a correlation coefficient of 0.64, immediately followed by the *average saturation* and *average blue channel* with 0.61 and −0.56, respectively. Thus, the results obtained by our predicted model are related to the informative power of these three features.

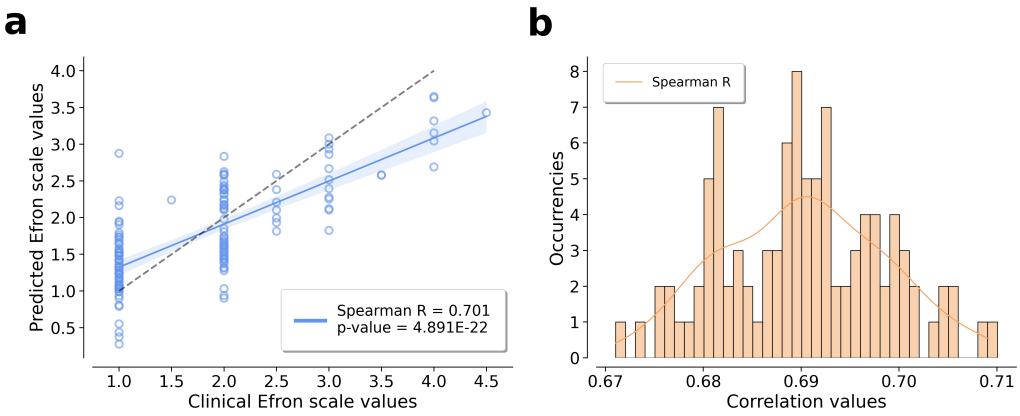

Figure 4. Results of the regression model for the prediction of the Efron scale values, developed starting from the set of features extracted. The correlation between the ground truth and the predicted values is estimated using the Spearman's rank correlation coefficient (ref. plot legend); (**a**) results of a single cross validation of the model. We highlight with the dashed line the axes bisector that corresponds to a perfect prediction. Despite the significance of the correlation found, the model finds many difficulties with low Efron scale values. We note that the predictions are performed on a set of data completely independent to the training set; (**b**) results obtained by the same pipeline on 100 different cross validations. In each iteration, a different 10-fold cross validation was applied for the estimation of the Spearman's rank correlation coefficients. The average Spearman's rank coefficient found is 0.69 with a standard deviation of 0.008.

Figure 5. Example of the predictions obtained by the regression model on three samples. For each image, we report the assigned Efron score and our method's predicted score. We would like to stress that the predicted scores are floating point numbers and therefore they are more descriptive for the redness evaluation, ensuring a finer redness grading scale of values.

Table 2. Spearman's correlation coefficient scores of single features in relation to the Efron grading scale values, ranked from the highest to the lowest. The analysis of the single feature correlation highlights an unbalanced informative power in the prediction of the Efron scale values. Only three features have a correlation coefficient greater than 0.55 (in absolute values).

Feature	Spearman's Correlation
$score_1$	0.642
μ_S	0.612
μ_B	−0.564
μ_G	−0.505
μ_H	−0.480
$score_2$	0.408
pixel-counting	−0.395
μ_R	0.314
μ_V	0.267
box-counting	−0.113

4. Discussion

The results obtained on the Efron scale predictions highlight a statistical agreement between only a small set of extracted features and the grading scores. In particular, we show the efficiency of some RGB measurements which have statistical significant correlations with the Efron grading scale values. Their efficiency is justified by their robustness to the most common issues in the image acquisition given by the light exposition and aperture of the slit lamp diaphragm.

The best correlation performances are achieved by the measure proposed by Park et al. ($score_1$): this measure is already known in literature as a feature related to the redness as much as the blue channel (the blue channel conserves the most informative component for the discrimination between the white-like conjunctiva area and the red-like vessels network). It is, however, interesting to notice how the saturation of the segmented vessels network area shows a not negligible correlation with the redness measure of the Efron scale. Moreover, the correlation between the image saturation and Efron values is a positive correlation, implying that the higher is the image exposition to the light with brighter red color and the higher tends to be the evaluation of the clinicians. The quality of the image certainly affects the evaluation of the clinicians, and it is quite obvious that, in a low contrast image, the evaluation of the vessels is harder for human eyes.

The analysis of the single features correlation also explains the apparently low efficiency of the $score_2$: despite the HSV color space being theoretically more robust than the RGB one and the high correlation between saturation and Efron values, the hue channel (μ_H) tends to anti-correlate with the Efron scale. The combination of the HSV channels proposed by Amparo et al. does not take care about them and thus their score is penalized in relation to the others.

The lowest predictive powers are achieved by the fractal features: the fractal analysis of the vessels network seems to be unrelated to the Efron values despite their theoretically information power in the description of the neo-genesis and ramification of the vessels. The fractal measures that are independent from the extension of the vessels network by definition can be seen as informative of the dynamic of the process of neo-genesis. Therefore, this lack of correlation seems to imply that the vessels growth is not different from the physiological condition.

The methods to estimate conjunctival redness are usually assessed by evaluating the agreement of the estimates of a panel of experts. The same set of samples is shown to a pool (at least two) of experts, and they perform the grading evaluation under the same experimental conditions. In this way, one can double check the obtained results, comparing the efficiency of the model against the expert evaluation using the agreement between the experts as the *maximal obtainable correlation*: the rationale is that any automated method can not agree with the opinion of any single expert more than how much two experts can agree between themselves. In our work, each image was independently evaluated by three experts. It is important to notice that, despite our automated method not reaching a perfect prediction, we observed an internal agreement between the experts' evaluations of at most 0.84 (evaluated as Spearman's correlation coefficient). In this case, any automated pipeline trained on this dataset could not achieve a predictive power greater than this value. Any result higher than this should be regarded as over fitting.

Our results also highlight a non-negligible correlation between the saturation of the image and the Efron scale values. It is not surprising that a brighter image is more easily analyzed by the human eye, but the correlation between the brightness of the image color and the Efron scale values leads us to hypothesize a possible criticality in the clinical evaluation that can lead to an overestimation of the hyperemia severity by the clinicians. The proposed pipeline tries to overcome the problems related to image exposure using the standardization algorithm described above, but further analysis will be required to completely remove the light dependence from the feature extraction procedure.

The robustness of our pipeline on a set of images sampled with a no-rigid acquisition protocol confirms its possible usage in a clinical practice as a viable decision support

system for the ophthalmologists. Our pipeline can help the clinicians in the evaluation and quantification of the hyperemia scores, providing a reference score given by a fully automated processing but also intermediate images which better highlight the conjunctival characteristics. We would like to stress that the entire pipeline is able to perform the prediction of the Efron values using floating point numbers, allowing a finer scale of values for the redness estimation. Moreover, the entire pipeline performs its evaluation in real time, since the most time-consuming task is given by the training procedure which must be performed once and for all. The proposed pipeline is currently used in the Ophthalmological Unit of IRCCS S. Orsola University Hospital of Bologna in Italy, and it is still being perfected to overcome the current limitations of the method. These improvements will be the subject of future works.

Author Contributions: N.C.: Formal analysis, Methodology, Software, Writing—Original Draft. E.G.: Formal analysis, Methodology, Software, Writing—Original Draft. F.G.: Resources, Data curation, Writing—Reviewing and Editing. F.B.: Resources, Data curation, Writing—Reviewing and Editing. L.C.: Methodology, Validation, Visualization, Data curation, Writing—Reviewing and Editing. G.C.: Methodology, Writing—Reviewing and Editing. P.V.: Project administration, Supervision, Conceptualization, Writing—Reviewing and Editing. E.M.: Project administration, Supervision, Conceptualization, Writing—Reviewing and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of AOU of Bologna, Policlinico S.Orsola-Malpighi (protocol number 100/2016/O/Sper).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patients to publish this paper.

Data Availability Statement: Data available on request due to privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Efron, N. Grading scales for contact lens complications. *Ophthalmic Physiol. Opt.* **1998**, *18*, 182–186. [CrossRef]
2. Efron, N.; Morgan, P.B.; Katsara, S.S. Validation of grading scales for contact lens complications. *Ophthalmic Physiol. Opt.* **2001**, *21*, 17–29. [CrossRef] [PubMed]
3. Rodriguez, J.; Johnston, P.; Ousler, G.R.; Smith, L.; Abelson, M.B. Automated grading system for evaluation of ocular redness associated with dry eye. *Clin. Ophthalmol.* **2013**, *7*, 1197–1204. [CrossRef] [PubMed]
4. Schulze, M.M.; Hutchings, N.; Simpson, T.L. Grading Bulbar Redness Using Cross-Calibrated Clinical Grading Scales. *Investig. Ophthalmol. Vis. Sci.* **2011**, *52*, 5812–5817. [CrossRef] [PubMed]
5. Peterson, R.C.; Wolffsohn, J.S. Sensitivity and reliability of objective image analysis compared to subjective grading of bulbar hyperaemia. *Br. J. Ophthalmol.* **2007**, *91*, 1464–1466. [CrossRef]
6. Alexovic, M.; Urban, P.L.; Tabani, H.; Sabo, J. Recent advances in robotic protein sample preparation for clinical analysis and other biomedical applications. *Clin. Chim. Acta* **2020**, *507*, 104–116. [CrossRef]
7. Bunya, V.Y.; Chen, M.; Zheng, Y.; Massaro-Giordano, M.; Gee, J.; Daniel, E.; O'Sullivan, R.; Smith, E.; Stone, R.A.; Maguire, M.G. Development and Evaluation of Semiautomated Quantification of Lissamine Green Staining of the Bulbar Conjunctiva From Digital Images. *JAMA Ophthalmol.* **2017**, *135*, 1078–1085. [CrossRef]
8. Amparo, F.; Wang, H.; Emami-Naeini, P.; Karimian, P.; Dana, R. The Ocular Redness Index: A novel automated method for measuring ocular injection. *Investig. Ophthalmol. Vis. Sci.* **2013**, *54*, 4821–4826. [CrossRef] [PubMed]
9. Park, I.K.; Chun, Y.S.; Kim, K.G.; Yang, H.K.; Hwang, J.M. New Clinical Grading Scales and Objective Measurement for Conjunctival Injection. *Investig. Ophthalmol. Vis. Sci.* **2013**, *54*, 5249–5257. [CrossRef] [PubMed]
10. Schulze, M.M.; Hutchings, N.; Simpson, T.L. The Use of Fractal Analysis and Photometry to Estimate the Accuracy of Bulbar Redness Grading Scales. *Investig. Ophthalmol. Vis. Sci.* **2008**, *49*, 1398–1406. [CrossRef] [PubMed]
11. Papas, E.B. Key Factors in the Subjective and Objective Assessment of Conjunctival Erythema. *Investig. Ophthalmol. Vis. Sci.* **2000**, *41*, 687–691.
12. Fieguth, P.; Simpson, T. Automated Measurement of Bulbar Redness. *Investig. Ophthalmol. Vis. Sci.* **2002**, *43*, 340–347.
13. Huntjens, B.; Basi, M.; Nagra, M. Evaluating a new objective grading software for conjunctival hyperaemia. *Contact Lens Anterior Eye J. Br. Contact Lens Assoc.* **2020**, *43*, 137–143. [CrossRef]

14. Koprowski, R.; Wilczynski, S.; Olczyk, P.; Nowinska, A.; Weglarz, B.; Wylegala, E. A quantitative method for assessing the quality of meibomian glands. *Comput. Biol. Med.* **2016**, *75*, 130–138. [CrossRef] [PubMed]
15. Wu, X.; Liu, L.; Zhao, L.; Guo, C.; Li, R.; Wang, T.; Yang, X.; Xie, P.; Liu, Y.; Lin, H. Application of artificial intelligence in anterior segment ophthalmic diseases: Diversity and standardization. *Ann. Transl. Med.* **2020**, *8*. [CrossRef]
16. Brea, M.L.S.; Rodríguez, N.B.; González, A.M.; Evans, K.; Pena-Verdeal, H. Defining the Optimal Region of Interest for Hyperemia Grading in the Bulbar Conjunctiva. *Comput. Math. Methods Med.* **2016**, *2016*. [CrossRef]
17. Derakhshani, R.; Saripalle, S.K.; Doynov, P. Computational methods for objective assessment of conjunctival vascularity. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 1490–1493. [CrossRef]
18. Shin, S.Y.; Lee, S.; Yun, I.D.; Lee, K.M. Deep vessel segmentation by learning graphical connectivity. *Med. Image Anal.* **2019**, *58*, 101556. [CrossRef] [PubMed]
19. Dirnberger, M.; Kehl, T.; Neumann, A. NEFI: Network Extraction From Images. *Sci. Rep.* **2015**, *5*. [CrossRef]
20. Liu, X.; Bi, L.; Xu, Y.; Feng, D.D.F.; Kim, J.; Xu, X. Robust deep learning method for choroidal vessel segmentation on swept source optical coherence tomography images. *Biomed. Opt. Express* **2019**, *10*, 1601. [CrossRef] [PubMed]
21. Radu, P.; Ferryman, J.; Wild, P. A robust sclera segmentation algorithm. In Proceedings of the 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), Arlington, VA, USA, 8–11 September 2015; pp. 1–6. [CrossRef]
22. Naqvi, R.; Loh, W.K. Sclera-Net: Accurate Sclera Segmentation in Various Sensor Images Based on Residual Encoder and Decoder Network. *IEEE Access* **2019**, *7*, 98208–98227. [CrossRef]
23. Rot, P.; Emersic, Ž.; Struc, V.; Peer, P. Deep Multi-class Eye Segmentation for Ocular Biometrics. In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, Costa Ricam, 18–20 July 2018; pp. 1–8. [CrossRef]
24. Hassan, G.; El-Bendary, N.; Hassanien, A.E.; Fahmy, A.; Snasel, V. Retinal Blood Vessel Segmentation Approach Based on Mathematical Morphology. *Procedia Comput. Sci.* **2015**, *65*, 612–622. [CrossRef]
25. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; [CrossRef]
26. Falk, T.; Mai, D.; Bensch, R.; Cicek, O.; Abdulkadir, A.; Marrakchi, Y.; Bohm, A.; Deubner, J.; Jackel, Z.; Seiwald, K.; et al. U-Net: Deep learning for cell counting, detection, and morphometry. *Nat. Methods* **2019**, *16*. [CrossRef] [PubMed]
27. Zhiping, L.; Hua, W.; Hong, J.; Rosa, G.G.; Jianhua, W. Quantitative analysis of conjunctival microvasculature imaged using optical coherence tomography angiography. *Eye Vis (Lond.)* **2019**, *6*. [CrossRef]
28. Tamuli, J.; Jain, A.; Dhan, A.V.; Bhan, A.; Dutta, M.K. An image processing based method to identify and grade conjunctivitis infected eye according to its types and intensity. In Proceedings of the 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, India, 20–22 August 2015; pp. 88–92. [CrossRef]
29. Guo, Z.; Hall, R.W. Parallel Thinning with Two-Subiteration Algorithms. *Commun. ACM* **1989**, *32*, 359–373. [CrossRef]
30. Xu, Z.; Jiang, H.; Tao, A.; Wu, S.; Yan, W.; Yuan, J.; Liu, C.; DeBuc, D.C.; Wang, J. Measurement variability of the bulbar conjunctival microvasculature in healthy subjects using functional slit lamp biomicroscopy (FSLB). *Microvasc. Res.* **2015**, *101*, 15–19. [CrossRef] [PubMed]
31. Hang, S.; Mingzhou, Z.; Xiaodan, H.; Kaixiu, L.; Xiaodan, J.; Yan, L.; Huibin, L.; Xuemin, L. Correlation Analysis of Ocular Symptoms and Signs in Patients with Dry Eye. *J. Ophthalmol.* **2017**, *2017*. [CrossRef]
32. Chen, W.; Deng, Y.; Jiang, H.; Wang, J.; Zhong, J.; Li, S.; Peng, L.; Wang, B.; Yang, R.; Zhang, H.; et al. Microvascular abnormalities in dry eye patients. *Microvasc. Res.* **2018**, *118*, 155–161. [CrossRef] [PubMed]