OPEN ACCESS

Fully automated grading system for the evaluation of punctate epithelial erosions using deep neural networks

Jing-Hao Qu ,[1,2] Xiao-Ran Qin,[3] Chen-Di Li,[1,2] Rong-Mei Peng,[1,2] Ge-Ge Xiao,[1,2] Jian Cheng,[3] Shao-Feng Gu,[1,2] Hai-Kun Wang,[1,2] Jing Hong [1,2]

[1]Department of Ophthalmology, Peking University Third Hospital, Beijing, China
[2]Beijing Key Laboratory of Restoration of Damaged Ocular Nerve, Peking University Third Hospital, Beijing, China
[3]Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Correspondence to
Dr Jing Hong, Ophthalmology, Peking University Third Hospital, Beijing 100191, China; hongjing196401@163.com

J-HQ, X-RQ and C-DL contributed equally.

Received 26 May 2021
Accepted 8 October 2021
Published Online First 20 October 2021

To cite: Qu J-H, Qin X-R, Li C-D, *et al*. *Br J Ophthalmol* 2023;**107**:453–460.

ABSTRACT

Purpose The goal was to develop a fully automated grading system for the evaluation of punctate epithelial erosions (PEEs) using deep neural networks.

Methods A fully automated system was developed to detect corneal position and grade staining severity given a corneal fluorescein staining image. The fully automated pipeline consists of the following three steps: a corneal segmentation model extracts corneal area; five image patches are cropped from the staining image based on the five subregions of extracted cornea; a staining grading model predicts a score for each image patch from 0 to 3, and automated grading score for the whole cornea is obtained from 0 to 15. Finally, the clinical grading scores annotated by three ophthalmologists were compared with automated grading scores.

Results For corneal segmentation, the segmentation model achieved an intersection over union of 0.937. For punctate staining grading, the grading model achieved a classification accuracy of 76.5% and an area under the receiver operating characteristic curve of 0.940 (95% CI 0.932 to 0.949). For the fully automated pipeline, Pearson's correlation coefficient between the clinical and automated grading scores was 0.908 ($p<0.01$). Bland-Altman analysis revealed 95% limits of agreement between the clinical and automated grading scores of between −4.125 and 3.720 (concordance correlation coefficient=0.904). The average time required for processing a single stained image during pipeline was 0.58 s.

Conclusion A fully automated grading system was developed to evaluate PEEs. The grading results may serve as a reference for ophthalmologists in clinical trials and residency training procedures.

INTRODUCTION

Punctate epithelial erosions (PEEs) are a feature of many ocular surface diseases and present as dots on the corneal epithelium. PEEs can reflect the physiology and function of the epithelium and are easily observed and assessed through corneal staining by slit-lamp microscopy.[1]

Corneal staining uses dyes that are applied on the ocular surface. Sodium fluorescein and lissamine green are two common dyes in clinical practice, the former of which is typically used to highlight corneal defects.[2] A dye-impregnated fluorescein paper strip is instilled into the eye, and punctate dots can be visualised under cobalt blue filter illumination. Currently, the most commonly used corneal fluorescein staining techniques in clinical trials are the Oxford scheme and the Nation Eye Institute/Industry (NEI) workshop grading system. The Oxford scheme was designed to evaluate the severity of dry eye syndrome[3]; however, it produces different features from those in the reference panel, and clinicians, especially at the junior level, may have trouble labelling images.[4] The NEI scale divides the cornea into five zones and summarises the corneal staining in each zone[5]; it combines both the area and intensity of the entire cornea simultaneously. Nevertheless, the NEI scale remains highly subjective and has low accuracy.

Digital image analysis can provide objective and accurate results and more sensitive and reliable assessments than those produced by subjective grading.[6] Computer-aided diagnosis has been applied to PEEs, and many semiautomated grading systems for PEEs have been developed.[7–9] However, these systems still require manual assistance.

Recently, deep neural networks have become widely used for medical image analysis. According to the literature, deep neural networks have been applied for a variety of retinal diseases and glaucoma.[10–12] Nevertheless, only one article describes an automatic PEEs grading system using a deep convolutional neural network.[13]

In this study, corneal fluorescein-stained samples were observed under cobalt blue filter illumination and photography. Digital photographs were graded by three ophthalmologists (two independent resident ophthalmologists and re-examination by a blinded specialist) according to the NEI scale and then learnt by deep neural networks. Then, a new, fully automated grading system was developed to evaluate PEEs. This system will improve the grading precision of existing methods and help train junior ophthalmologists.

MATERIALS AND METHODS

Subjects

This was a retrospective study. Participants were identified by two experienced ophthalmologists (JH and G-GX) based on the following inclusion criteria: healthy cornea and cornea with PEEs only. Subject who had (1) filamentosa keratitis; (2) a history of corneal transplantation or (3) a condition that the investigator felt may have confounded the study results, may have put the subject at significant risk, or may have interfered significantly with the subject's participation in the study were excluded.

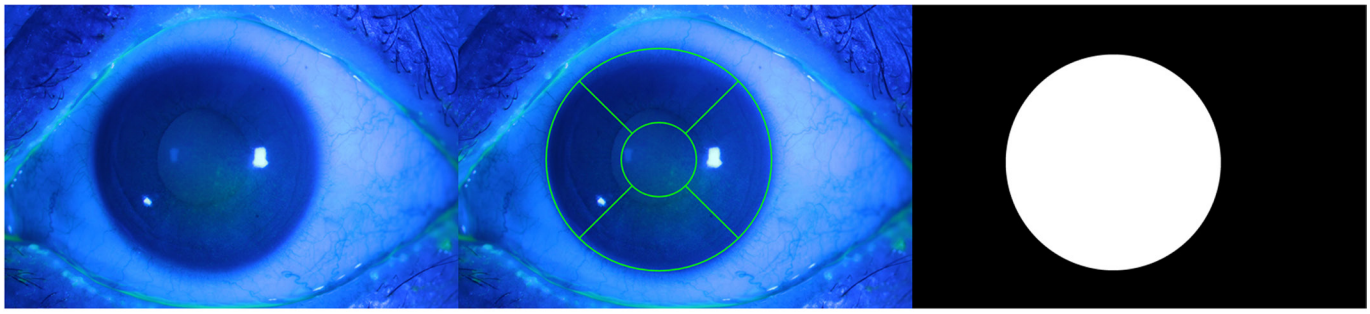

Figure 1 An example of a corneal area annotation. (A) Corneal fluorescein staining image; (B) annotated image and (C) ground truth.

Image capture technique

A sodium fluorescein ophthalmic strip (Meizilin Pharmaceutical Co, Liaoning, China) was made wet with a single drop of ofloxacin ophthalmic solutions; once, the drop had saturated the strip, any excess fluid was shaken off. The lower eyelid was pulled down, and the strip was gently touched onto the lower tarsal conjunctiva (once only). The patients were asked to gently blink to distribute the dye across the ocular surface. A photograph of the entire cornea was taken immediately after gentle blinking.

Images of the entire cornea were acquired with a photo slit-lamp system (BX 900, Haag-Streit, Bern, Switzerland) with a blue filter. The magnification was set to 10× to show all the corneal sections. Tear-layer reflection was minimisedusing a diffused flash system. The ISO, flash intensity, slit beam, illumination angle and other camera parameters were constant across all subjects. All images were captured in RAW format to maximise the acquired information. These pictures were transmitted to a personal computer and saved as JPG files (3456×2304 pixels, 24-bit, RGB).

Image labelling

Two independent resident ophthalmologists (J-HQ and C-DL) selected all corneal sections and graded the photographs with the NEI scale. Each ophthalmologist independently graded the photographs on their own monitor, which was set to a resolution of 1920×1080; the ophthalmologists used a medical annotation website developed by the authors for the labelling within the illumination rooms in their clinics without any time limitations.

The grading system recommended by the NEI divides the cornea into five zones (central, superior, temporal, nasal and inferior), and for each zone, the severity of corneal fluorescein staining is graded on a scale from 0 to 3 based on the reference figures. Therefore, the maximum total score for an entire cornea is 15.

For the corneal area annotations, an initial circle consisting of five regions was generated, and the ophthalmologists could adjust the circle to fit the position and size of the corneal area. A corneal area annotation example and its corresponding binary ground truth are shown in figure 1. For grading annotation, ophthalmologists could assign a score from 0 to 3 for each subregion. All scores were checked again by a blinded specialist (R-MP). During this checking, the blinded specialist picked out the images assigned inconsistent scores by the two resident ophthalmologists, reviewed those images, and gave final scores based on the specialist's own judgement; for images assigned consistent scores, the specialist did not change their scores. The final score for each subregion was then regarded as the ground truth score.

Datasets

A total of 1046 images were collected. Among these, 283 images displaying corneal ulcers, filaments, ambiguous corneal limbus or blurring were excluded. The remaining 763 images were used as the corneal fluorescein staining dataset. The dataset was randomly divided into three parts: a training set (534 images) to train the deep neural networks, a validation set (76 images) to tune the hyperparameters of the training process (such as early stopping conditions and the learning rate) and a testing set (153 images) to evaluate the performance of the trained models. The process of dataset creation is shown in figure 2.

To improve the performance and generalisability of the developed cornea segmentation model, an external public dataset, the SUSTech-SYSU dataset, was used as an additional training set (only for the cornea segmentation task).[14] The SUSTech-SYSU dataset contains 712 fluorescein staining images and the corresponding segmentation labels of the corneal areas.

Fully automated grading system

In this study, a fully automated grading system for the assessment of corneal punctate staining was developed based on deep learning. The pipeline of the automated grading system is shown in figure 3. First, given a corneal fluorescein staining image, the corneal segmentation model extracts an elliptical corneal boundary. Then, based on this extracted elliptical boundary, the corneal area is separated into five subregions. Each image patch is determined by the minimum bounding rectangle of each subregion. The associated five image patches are cropped from the original staining image for further staining grading. Finally, the staining grading model extracts deep features and predicts a score for each input image patch. The total score of the original staining image is calculated after the grading process is completed for all five image patches.

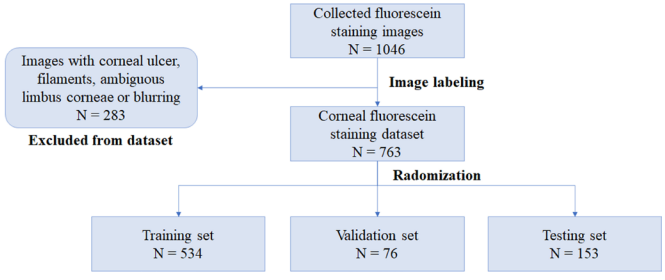

Figure 2 The process of dataset creation.

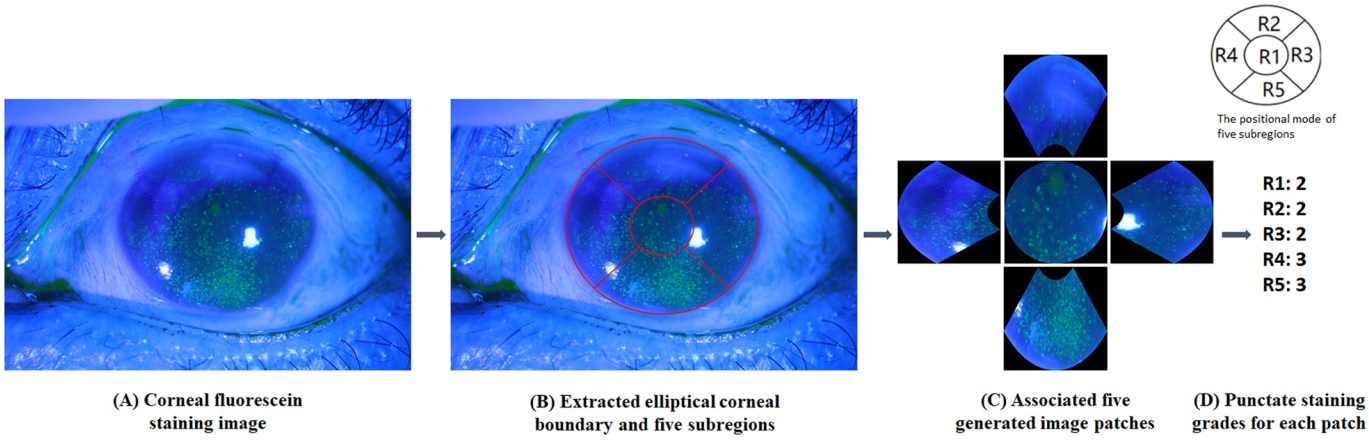

Figure 3 Pipeline of the fully automated grading system. (A) Corneal fluorescein staining image; (B) extracted elliptical corneal boundary and five subregions; (C) associated five generated image patches and (D) punctate staining grades for each patch.

Corneal segmentation

The cornea segmentation task was formulated as a binary segmentation problem, where the corneal area belongs to the foreground and other areas belong to the background. A corneal segmentation model was established to segment the corneal area from the corneal fluorescein staining image. Recently, fully convolutional networks (FCNs) have shown great progress in semantic segmentation and medical image segmentation tasks. FCNs typically use multiple convolution layers to extract features and downsize the resolution of feature maps, and then use transposed convolution layers to recover the resolution of output feature maps. Specifically, the convolution layer is composed of several convolution units, each of which can be seen as a filter and performs a convolution operation with the input image or feature maps. The first few convolution layers extract shallow features, such as edges and corners, and the deeper convolution layers extract high-level semantic features. U-Net and D-LinkNet, which follow the FCN architecture, were investigated to determine their applicability to corneal segmentation.[15 16] As shown in figure 4, the corneal segmentation model has an encoder–decoder architecture.

The encoder uses a ResNet34 backbone (excluding the last basic block) and a dilation block to extract image features.[17] The adopted ResNet34 backbone, which contains three residual blocks, is relatively light and easily converges. The dilation block is composed of four stacked dilated convolution layers with dilation rates of 1, 2, 4 and 8, each of which is followed by a rectified linear unit.[18] The dilated convolution layers are based on the original convolution layer and are implemented by adding zeros between each number in the original convolution kernel; the number of zeros minus 1 is called the dilation rate. The original convolution layer is a special dilated convolution layer with a dilation rate of 1. The intermediate feature maps produced by each dilated convolution layer are summed to generate the final output of the dilation block. Using the dilation block can increase the size of the receptive fields of the resulting feature maps and aggregate multiscale context information without reducing the

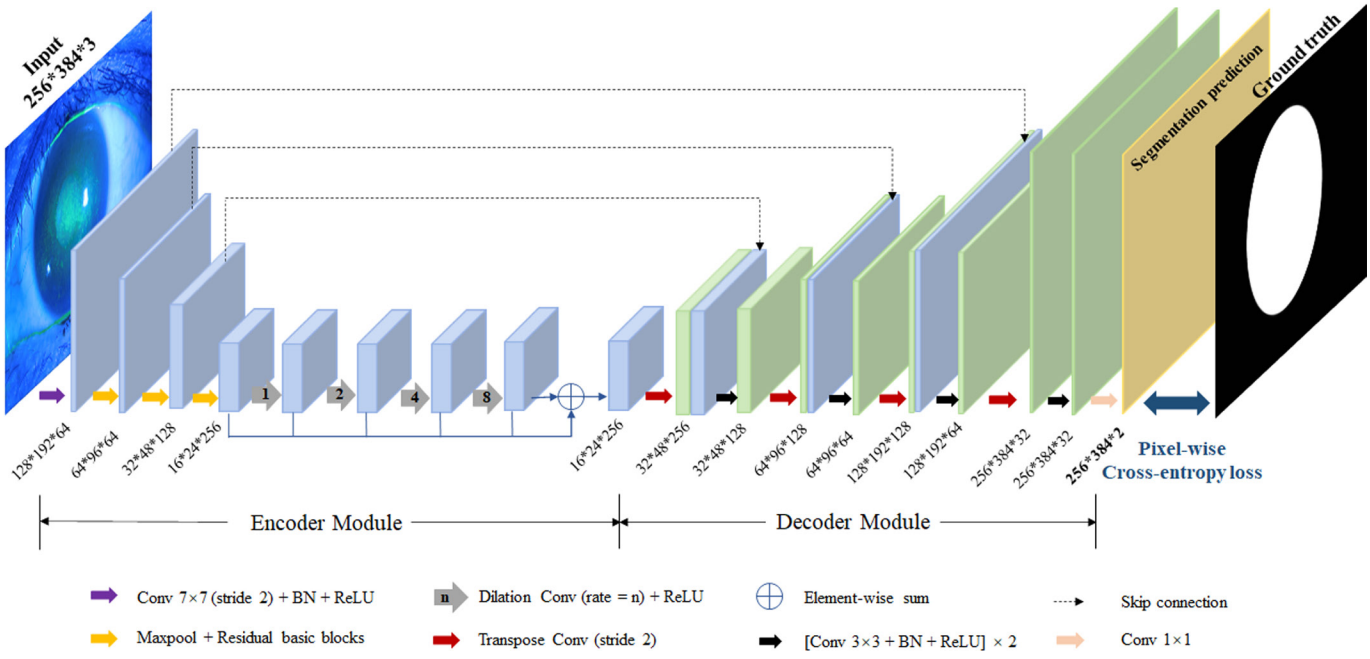

Figure 4 Architecture of the corneal segmentation model. ReLU, rectified linear unit.

resolution of the output feature maps. The decoder adopts the design of the U-Net decoder, which uses transposed convolution layers to upscale the resolutions of the feature maps and skip connections to concatenate the correspondingly encoded feature maps. This helps to restore precise segmentation features with detailed information. Finally, a convolution layer is used to generate the final segmentation prediction with two output channels.

During the training process, the corneal segmentation model is learnt under supervision of the ground truth segmentation using a pixel-wise cross-entropy loss. In the inference process, a softmax layer is applied to the segmentation prediction along the channel axis, to classify each pixel of the prediction map as belonging to either the corneal area or the background, and a binary mask is obtained for the original staining image. As the shape of the cornea is usually formulated as an ellipse and the predicted corneal area has an irregular shape in the binary mask, contour extraction and ellipse fitting are proposed to produce an elliptical corneal boundary. The centre coordinates, long axis and short axis of the detected elliptical boundary are recorded to determine the position of the cornea and its five subregions.

Generation of the image patches to be graded

Referring to the NEI scale, the proposed grading system divides the corneal area into five parts and assigns a score to each part. Based on the detected elliptical boundary of the corneal area, the five subregions can be calculated by a prebuilt positional mode. The first subregion is determined by an inner ellipse located at the centre of the detected cornea, whose axis is one-third of the axis of the detected elliptical boundary. The remaining four subregions are determined by four line segments that partition the ring into four sectors of 90° each.

Because each subregion should be graded separately, five image patches are cropped from the original staining image, where each image patch is determined by the minimum bounding rectangle of each subregion. More specifically, for each image patch, the intensity values of the pixels that do not fall into the corresponding subregion are set to zero, which guarantees the exact representations of the five subregions.

Punctate staining grading

The punctate staining grading task for each generated image patch was regarded as a four-class classification problem, where the four-class labels correspond to four severity scores (0–3). A staining grading model was employed for classification, taking cropped and resized image patches with sizes of 512×512×3 as inputs.

In recent years, deep convolutional neural networks have demonstrated amazing performance on many image classification tasks and have even surpassed human experts with some large datasets. Models pretrained on such datasets learn general features from various images and thus can be transferred to specific tasks in which only a relatively small dataset is available. Inspired by this, in this study, ResNet34 was adopted as the architecture of the proposed staining grading model, which was trained by fine-tuning the pretrained parameters on the ImageNet dataset to adapt them to the proposed grading task.[17 19] Because the input size for the staining grading model (512×512×3) differs from the original input size for ResNet34 (224×224×3), the intermediate feature maps produced by the last basic block have a relatively large resolution of 16×16×512, where the dimensions indicate height, width and channel. To reduce the complexity of the model, an adaptive pooling layer is used to downsize the obtained feature maps and produce a fixed-size feature vector. Finally, the feature vector is fed into a fully connected layer with four output nodes to represent the four classes.

During training, given the imbalance in the samples among the different classes (shown in table 1), a weighted cross-entropy loss is employed with predefined weights of 0.1, 0.2, 0.5 and 0.2. During inference, a softmax layer is used after the final fully connected layer to produce probability estimates for the four class labels, and the class label with the maximum probability value determines the predicted score. When preparing input image patches during the training process, the positions of the five subregions were based on the ground truth corneal boundary, while for the fully automated grading pipeline during the inference process, they were based on the detected elliptical boundary of the corneal area.

Data augmentation and experimental details

For the corneal segmentation model experiment, a combined training dataset consisting of 1246 corneal fluorescein staining images (534 images from the training set and 712 images from the SUSTech-SYSU dataset) was used to train the model, while the validation set was used to assess the convergence of the model every few training epochs. As the original high-resolution images took up a large amount of graphics processing unit (GPU) memory and the corneal area covers nearly one quarter of the image area, all images were resized to 384×256 pixels. The pixel intensities were normalised to values of 0-1 for better convergence. Data augmentation methods, including random horizontal flipping, vertical flipping and jittering in brightness, contrast and sharpness, were used to enhance the training dataset. The ResNet34 backbone in the encoder was initialised by the parameters pretrained on ImageNet, while other layers adopted the Kaiming initialisation algorithm.[20] During training, the Adam optimiser was used with a batch size of 8 and a weight decay of 0.0005. The learning rate started at 0.001 and decreased gradually with a step size of 5000 iterations and a factor of 0.25. After 100 epochs, the model that achieved the best performance on the validation set was selected for further evaluation.

For the staining grading model experiment, a set of image patches were cropped from the corneal fluorescein staining images at original resolutions (see details in the Generation of the image patches to be graded section) in the original training set, validation set, and testing set and then collected to train, validate and test the model, respectively. For the collected data,

Table 1 Numbers of image patches before and after offline augmentation in the training data for different classes

Class	Numbers of image patches without offline augmentation	Numbers of image patches with offline augmentation
Score 0	1350	8100
Score 1	616	3696
Score 2	262	1572
Score 3	442	2652

each image patch had a ground truth score of 0–3 (referring to the ground truth score of its corresponding subregion). All image patches were resized to 512×512 pixels, and the image pixels are normalised to the values of 0–1. The training data were relatively small and demonstrated an imbalance in the representation of the different classes: 1350 score 0 image, 616 score 1 image, 262 score 2 images and 442 score 3 images. To reduce the risk of overfitting, data augmentation methods were employed as follows. First, when cropping the image patches from the original staining images, the ground truth corneal boundary was shifted and resized randomly for better generalisation while randomly rescaling its centre coordinates, long axis and short axis in the range [−20, 20]. After these offline augmentation methods were executed, the number of image patches was augmented to six times the amount in the original training data, as shown in table 1. Subsequently, the input image patches were randomly flipped, rotated by 0°/90°/180°/270° and jittered in brightness, contrast, sharpness and colour. During training, the Adam optimiser was used with a batch size of 50 and a weight decay of 0.0005. The learning rate started from 0.0001 and decreased gradually with a step size of 1200 iterations and a factor of 0.25. The grading model was trained with 15 epochs.

Both the corneal segmentation model and the staining grading model were implemented with the PyTorch framework, and trained on a RedHat operating system with a 12 GB NVIDIA Tesla K80 GPU.

Evaluation metrics for the deep models

To evaluate the performance of the proposed model on the corneal segmentation task, the intersection over union (IoU) was calculated as follows

$$\mathrm{IoU} = \frac{area\left(EC_{pred} \cap EC_{gt}\right)}{area\left(EC_{pred} \cup EC_{gt}\right)}$$

where EC_{pred} is the prediction result of the detected elliptical cornea for the original staining images and EC_{gt} is the ground truth elliptical cornea. The IoU is defined as the ratio of the intersection of two regions to the union of the two regions. Thus, the IoU ranges from 0.0 to 1.0, where larger values indicate higher coincidence between the two regions. The IoU was also calculated between the corneal area annotations made by the two resident ophthalmologists for model–human comparison.

For the staining grading model, the applied performance metrics included a confusion matrix, the classification accuracy, the area under the receiver operating characteristic curve (AUC) and the mean absolute error (MAE). Specifically, the confusion matrix reports information on ground truth results and prediction results for multiclass classification. The classification accuracy is calculated as the ratio of the number of correctly classified samples to the total number of samples in the dataset, where a correctly classified sample is that for which the ground truth score matches the predicted score. When plotting the receiver operating characteristic (ROC) curve, the micro-averaged version was used to compute the AUC. Standard ROC curves are graphical plots used to assess the performance of a binary classifier, where larger AUC values (range between 0.5 and 1.0) indicate better performance. However, the staining grading task is a multiclass classification, and so a variant version of the regular ROC curve was used, that is, the micro-averaged version. The micro-averaged version calculates global metrics by considering each element of the label indicator matrix as a label.[21] The MAE indicates the distance between the ground truth score S^i_{gt} and the corresponding predicted score S^i_{pred}, and is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|S^i_{pred} - S^i_{gt}\right|,$$

where N is the size of the dataset. Considering an extreme case in which a sample with a true label of 0 is classified as having a score of 4 versus score of 1, the MAE metric is more helpful for guiding the clinical application of the grading model.

Statistical analysis

To evaluate the performance of the proposed fully automated grading pipeline, the clinical grading score and the automated grading score were introduced. For each staining image, the sum of ground truth scores of the five subregions was taken as the clinical grading score (0–15). For the fully automated grading system, each input staining image was first fed to the corneal segmentation model to detect the elliptical corneal area, and then its five subregions were extracted and graded by the staining grading model. After the fully automated process was completed, the total score of the input image was obtained and defined as the automated grading score (0–15).

Statistical analysis was performed using SPSS V.18.0 (SPSS, Chicago, IL, USA). The correlation between the clinical grading scores and the automated grading scores was examined using the Pearson test. The interobserver correlation was also calculated between the manual grading scores (0–15) given by the two resident ophthalmologists using the Pearson test. All tests were two-tailed, $p<0.05$ was considered statistically significant, and $p<0.01$ was considered very statistically significant. To evaluate the agreement between the clinical grading scores and automated grading scores, Bland-Altman analysis was used.

RESULTS

The cornea segmentation model and the staining grading model were evaluated on the original testing set, which included 153 staining images and did not overlap with the training data or validation data of the two models. A total of 765 image patches were obtained to assess the grading model. More importantly, to demonstrate the performance of the fully automated grading pipeline, an experiment involving the direct prediction of automated grading scores for the 153 testing images using the two models described above was conducted.

On the corneal segmentation task, the segmentation model achieved an IoU of 0.937, while that between the annotations made by the two resident ophthalmologists was 0.915. Thus, the difference between the detected elliptical cornea and the ground truth elliptical cornea was lower than that between the annotations of the two resident ophthalmologists. Regarding the punctate staining grading task, the confusion matrix is given in figure 5A, and the classification accuracy of the grading model was 76.5%. Figure 5B shows the normalised confusion matrix, showing that class 0 was the easiest to distinguish while more errors were obtained for class 1 and class 2. The micro-averaged ROC curve is shown in figure 6, yielding an AUC of 0.940 (95% CI, 0.932 to 0.949). The grading model also achieved an MAE of 0.280.

For the fully automated pipeline, the Pearson's correlation coefficient between the clinical and automated grading scores was 0.908 ($p<0.01$, figure 7A). For comparison, the Pearson's correlation coefficient between the manual scores of the two resident ophthalmologists was 0.781. A total of 77 observations (dark flowers 36 observations and light flowers 41 observations) showed a good agreement between the clinical and automated grading (figure 7B). In addition, the results of Bland-Altman analysis of the clinical and automated grading scores is given

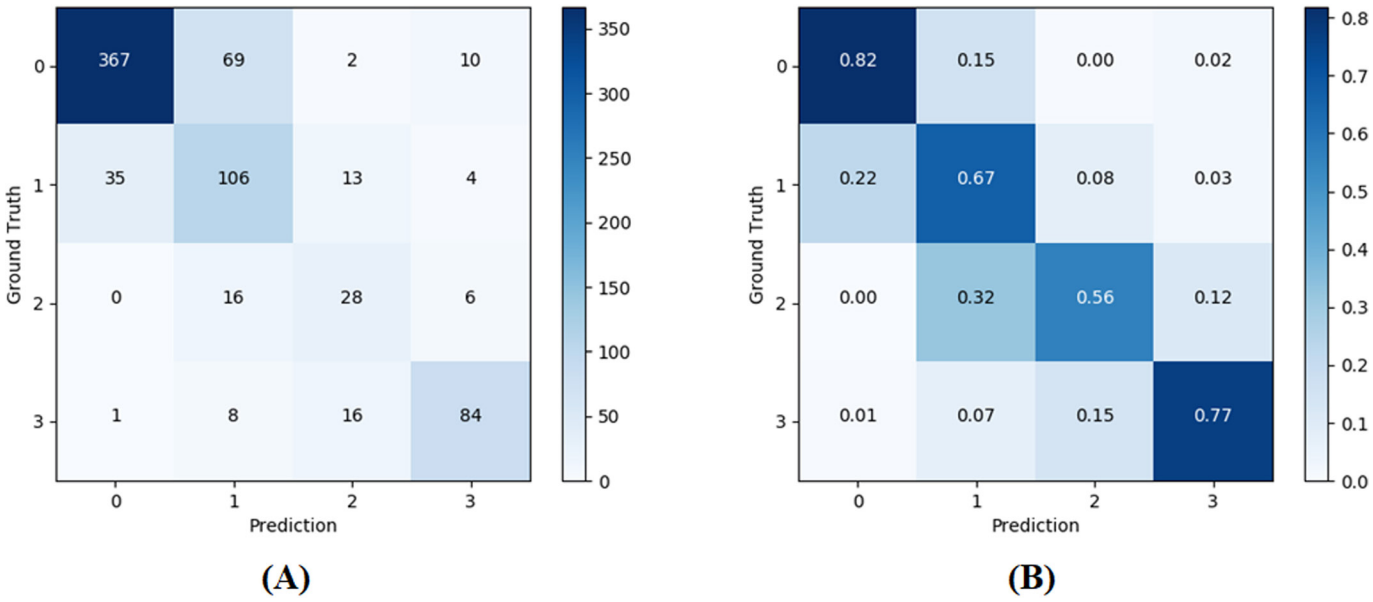

Figure 5 Confusion matrices. (A) Confusion matrix of the staining grading model and (B) normalised confusion matrix.

in figure 8. The 95% limits of agreement between the clinical and automated grading scores were between −4.125 and 3.720 (concordance correlation coefficient=0.904). Furthermore, the average time required for processing a single staining image in the fully automated grading pipeline was 0.58 s. The visual results of some example images from the testing set are shown in figure 9, where figure 9A presents comparisons and IoUs of the detected elliptical corneal boundary (red ellipse) to the ground truth corneal boundary (green ellipse), and figure 9B gives the predicted scores for five extracted subregions.

DISCUSSION

Corneal fluorescein staining is a useful method for evaluating ocular surfaces, and its assessment outcomes are needed for conducting multicentre studies or large-scale clinical trials. Various clinical grading methods for corneal staining have been introduced to compare the images of patients' eyes with reference images. As reported earlier, the Oxford scheme, the NEI scale, the area–density combination index and ocular staining score of Sjögren's International Collaborative Clinical Alliance are all useful methods for evaluating corneal staining.[4] However, all corneal staining systems are difficult to implement in clinical practice. The proposed automated PEEs system is a relatively objective assessment of corneal fluorescein staining. The use of imaging techniques and associated software analysis as a complement or substitute for clinical scoring is already widespread in ophthalmology. These methods provide increased efficiency when processing clinical data and standardise the grading procedure across multicentre clinical trials.

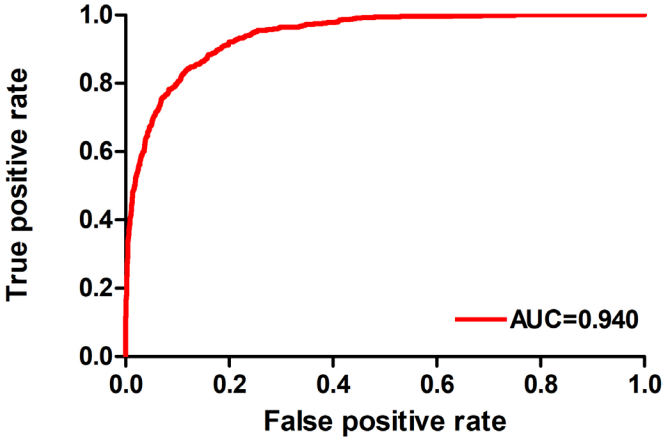

Figure 6 Receiver operating characteristic curve of the staining grading model.

The proposed automated PEEs grading system was shown to be capable of grading the PEEs condition automatically. A high correlation coefficient (r=0.908, p<0.01) was observed between the estimated and clinical grades. Pearson's correlation coefficient between the manual scores of the two resident ophthalmologists was 0.781, which shows that in the clinic, the manual grading process is easily influenced by subjectivity. Additionally, the prediction results from the proposed system are strongly related with the clinical results from the specialist. The IoU of the segmentation model was much higher than that of resident ophthalmologists (0.937 vs 0.915). This indicates that the corneal segmentation model can extract accurate corneal areas for further grading tasks and play an important role in the fully automated system, where manual input or intervention is not required at all. As seen in the normalised confusion matrix (figure 5B), class 1 and class 2 were more easily misclassified. One reason for this may be that the details in images with a score of 1 (slight punctate staining) may be ignored, leading to a prediction of a score of 0. Another reason could be that images with a score of 1 and a score of 2 were seen as too similar by the staining grading model. Furthermore, the average time required to process a single image was 0.58 s, faster than that achieved by the method proposed in a previously published study (6.25 s).[7] The results of these analyses illustrate that the fully automated grading system could potentially assist ophthalmologists in performing faster and more accurate diagnoses.

The correlation between the estimated grades and the clinical grades is higher than that in the studies by Chun *et al*,[7] Rodriguez *et al*[8] and Bunya *et al*[9] (r=0.90, r=0.88 and r=0.83, respectively), all of which involved semiautomated systems. The paper written by Su *et al*[13] was the only one focusing on an automatic PEEs grading system using deep neural networks and achieved a

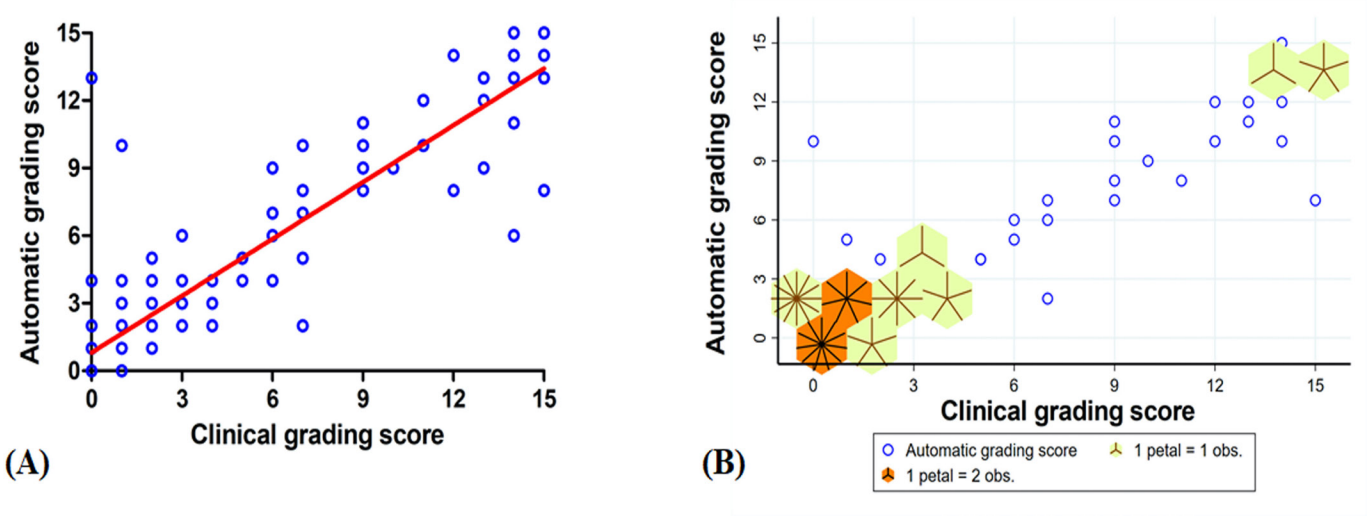

Figure 7 Relationship between the clinical and automated grading scores. (A) Correlation between the clinical and automated grading scores and (B) density-distribution sunflower plot between the clinical and automated grading scores.

correlation between the estimated grades and the clinical grades of 0.85. The corneal grading system proposed in above paper may have shortcomings. Because the corneal limbus is not fully exposed in the images, the corneas may be inaccurately graded. As the grades are highly correlated with the clinical results, our grading results may serve as a reference for ophthalmologists in clinical trials and residency training procedures.

CONCLUSIONS

Using deep neural networks, a fully automated grading system was developed for evaluating PEEs. The automated PEEs grading system could serve as an excellent assistant in clinical and multicentre clinical trials.

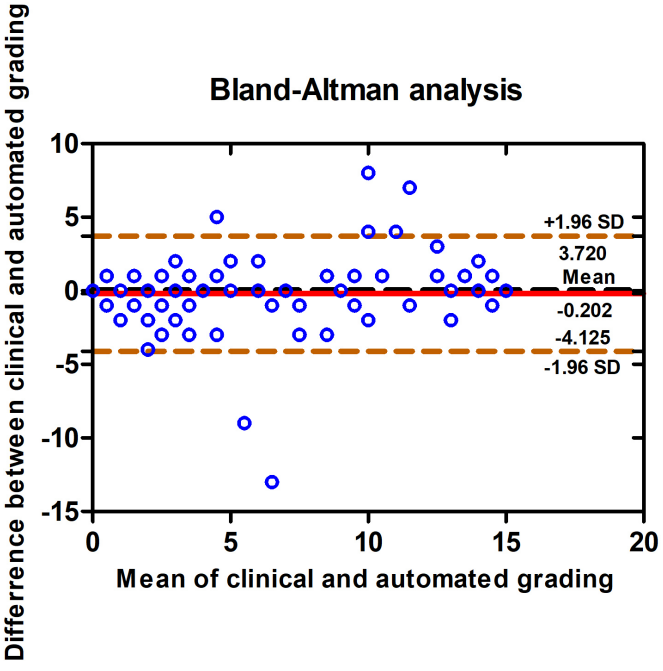

Figure 8 Bland-Altman plot comparing the clinical and automated grading scores.

Limitations

The BX 900 photo slit-lamp system uses blue light to excite fluorescein molecules to highlight damage to the ocular surface after their instillation. According to the previous research, the wavelength of blue light and the use of a yellow cut-off filter to remove extraneous blue light are critical to the optimal visualisation of ocular surface staining.[22] If subjective grading was performed with images using a yellow cut-off filter, the correlation between subjective and objective assessment might be stronger. What's more, the NEI scale is not linear (with the amount and type of staining combined in one scale) and limited in sensitivity which will have impacted the comparison with objective grading. The

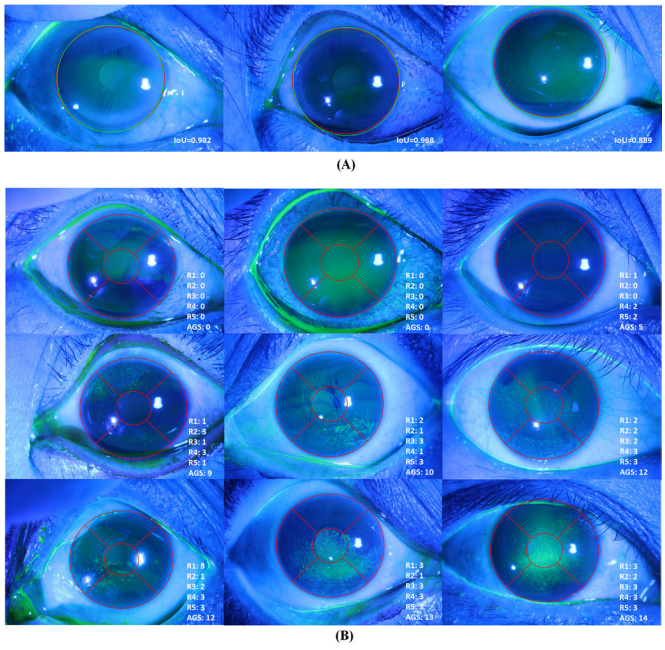

Figure 9 Example images from the testing set. (A) Comparisons and IoUs of the detected elliptical corneal boundary (red ellipse) to the ground truth corneal boundary (green ellipse); (B) the predicted scores for the five extracted subregions. AGS is the sum of five predicted scores. AGS, automated grading score; IoU, intersection over unions.

images were obtained from individuals from a single ethnic background in a single centre; the corneal fluorescein staining dataset thus needs to be expanded. The results obtained using the proposed PEEs grading system should be further confirmed through large-scale clinical trials.

Contributors Design of the study (G-GX, JH); conduct of the study (J-HQ, X-RQ, C-DL); collection and management of data (J-HQ, C-DL, R-MP); image capturing (S-FG, H-KW); analysis and interpretation of data (X-RQ, JC); writing of manuscript (J-HQ, X-RQ); responsible for the overall content as the guarantor (JH).

Funding This study was supported by the National Natural Science Foundation of China under grant nos 81970768 and 81800801. This study was also financially supported by the China National Key Research and Development Program no. 2020AAA0105004.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was performed according to the tenets of the Declaration of Helsinki and was approved by the local ethics committee (S2021117).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. None.

ORCID iDs
Jing-Hao Qu http://orcid.org/0000-0003-2368-3708
Jing Hong http://orcid.org/0000-0002-8079-2073

REFERENCES

1 Bron AJ, Argüeso P, Irkec M, *et al*. Clinical staining of the ocular surface: mechanisms and interpretations. *Prog Retin Eye Res* 2015;44:36–61.
2 Begley C, Caffery B, Chalmers R, *et al*. Review and analysis of grading scales for ocular surface staining. *Ocul Surf* 2019;17:208–20.
3 Bron AJ, Evans VE, Smith JA. Grading of corneal and conjunctival staining in the context of other dry eye tests. *Cornea* 2003;22:640–50.
4 Sook Chun Y, Park IK. Reliability of 4 clinical grading systems for corneal staining. *Am J Ophthalmol* 2014;157:1097–102.
5 Lemp MA. Report of the National eye Institute/Industry workshop on clinical trials in dry eyes. *Clao J* 1995;21:221–32.
6 Peterson RC, Wolffsohn JS. Objective grading of the anterior eye. *Optom Vis Sci* 2009;86:273–8.
7 Chun YS, Yoon WB, Kim KG, *et al*. Objective assessment of corneal staining using digital image analysis. *Invest Ophthalmol Vis Sci* 2014;55:7896–903.
8 Rodriguez JD, Lane KJ, Ousler GW, *et al*. Automated grading system for evaluation of superficial punctate keratitis associated with dry eye. *Invest Ophthalmol Vis Sci* 2015;56:2340–7.
9 Bunya VY, Chen M, Zheng Y, *et al*. Development and evaluation of semiautomated quantification of lissamine green staining of the bulbar conjunctiva from digital images. *JAMA Ophthalmol* 2017;135:1078–85.
10 Son J, Shin JY, Kim HD, *et al*. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology* 2020;127:85–94.
11 Medeiros FA, Jammal AA, Mariottoni EB. Detection of progressive glaucomatous optic nerve damage on fundus Photographs with deep learning. *Ophthalmology* 2021;128:383–92.
12 Liefers B, Taylor P, Alsaedi A, *et al*. Quantification of key retinal features in early and late age-related macular degeneration using deep learning. *Am J Ophthalmol* 2021;226:1–12.
13 Su T-Y, Ting P-J, Chang S-W. Superficial punctate keratitis grading for dry eye screening using deep Convolutional neural networks. *IEEE Sens J* 2020;20:1672–8.
14 Deng L, Lyu J, Huang H, *et al*. The SUSTech-SYSU dataset for automatically segmenting and classifying corneal ulcers. *Sci Data* 2020;7:23.
15 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015.
16 Zhou L, Zhang C, Ming W. D-LinkNet: LinkNet with Pretrained Encoder and dilated convolution for high resolution satellite imagery road extraction. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.
17 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. IEEE Conference on Computer Vision & Pattern Recognition, 2016.
18 Yu F, Koltun V. Multi-Scale context aggregation by dilated Convolutions. ICLR, 2016.
19 Russakovsky O, Deng J, Su H. ImageNet large scale visual recognition challenge. *Int J Comput Vision* 2014:1–42.
20 He K, Zhang X, Ren S. Delving deep into Rectifiers: Surpassing Human-Level performance on ImageNet classification. *CVPR* 2015.
21 Scikit-learn developers (BSD License). Compute area under the receiver operating characteristic curve (ROC AUC) from prediction scores, 2021. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score
22 Peterson RC, Wolffsohn JS, Fowler CW. Optimization of anterior eye fluorescein viewing. *Am J Ophthalmol* 2006;142:572–5.