Keratoconus detection of changes using deep learning of colour-coded maps

Xu Chen, Jiaxin Zhao, Katja C Iselin, Davide Borroni, Davide Romano, Akilesh Gokul, Charles N J McGhee, Yitian Zhao, Mohammad-Reza Sedaghat, Hamed Momeni-Moghadam, Mohammed Ziaei, Stephen Kaye, Vito Romano, Yalin Zheng

ABSTRACT

Objective To evaluate the accuracy of convolutional neural networks technique (CNN) in detecting keratoconus using colour-coded corneal maps obtained by a Scheimpflug camera.

Methods and analysis We included the images of keratoconic and healthy volunteers’ eyes provided by three centres: Royal Liverpool University Hospital (Liverpool, UK), SeyedTagh Eye Clinic (Mashhad, Iran) and The New Zealand National Eye Center (New Zealand). Corneal tomography scans were used to train and test CNN models, which included healthy controls. Keratoconic scans were classified according to the Amsler-Krumeich classification. Keratoconic scans from Iran were used as an independent testing set. Four maps were considered for each scan: axial map, anterior and posterior elevation map, and pachymetry map.

Results A CNN model detected keratoconus versus health eyes with an accuracy of 0.9785 on the testing set, considering all four maps concatenated. Considering each map independently, the accuracy was 0.9283 for axial map, 0.9642 for thickness map, 0.9642 for the front elevation map and 0.9749 for the back elevation map. The accuracy of models in recognising between healthy controls and stage 1 was 0.90, between stages 1 and 2 was 0.9032, and between stages 2 and 3 was 0.8537 using the concatenated map.

Conclusion CNN provides excellent detection performance for keratoconus and accurately grades different severities of disease using the colour-coded maps obtained by the Scheimpflug camera. CNN has the potential to be further developed, validated and adopted for screening and management of keratoconus.

INTRODUCTION

Keratoconus is a non-inflammatory degeneration that leads to progressive corneal thinning, myopia, irregular astigmatism and scarring, resulting in debilitating disease that may significantly affect patients’ quality of life. The early and accurate detection of keratoconus provides opportunities to reduce risk factors and offer treatments to potentially slow its progression.

Detecting keratoconus in its early stages, however, can be difficult, as patients may have normal visual acuity and an unremarkable slit-lamp examination. The diagnosis often requires an assessment of the topography and/or tomography of the cornea to reveal subtle changes in corneal morphology. The evaluation of corneal topography and tomography may itself be challenging and open to misinterpretation. Several indices have been proposed to facilitate differentiation between keratoconus and normal eyes, such as the zone of increasing corneal power, inferior–superior corneal power asymmetry, steepest radial axes, posterior and anterior corneal and pachymetry.

Utilisation of artificial intelligence (AI) to evaluate and detect early stages of various diseases at much faster rates than physicians has seen a sharp rise in recent years. In particular, AI techniques developed from artificial neural networks, deep learning (DL)-based algorithms show great promise in extracting features and learning patterns from complex data. Indeed, in eye care, DL techniques have helped physicians in detection of anatomical structures and/or lesions, diagnosis of various diseases and provided prognostic insights. Most studies are, however, primarily focused on
relatively common diseases such as diabetic retinopathy, age-related macular degeneration and glaucoma.26

The sensitivity and specificity of machine learning for the detection of keratoconus has been evaluated in several studies.27–40 Among current neural networks, the one which may be more suitable for the evaluation of the keratoconus is the convolutional neural networks technique (CNN), which is one of the main methods of recognising and classifying images through their colour-code pattern.41 CNN, therefore, can be applied to corneal topographic colour maps. Four studies have applied CNNs to patients with keratoconus,27–29 39  but with relatively small numbers, limiting it potential application. We, therefore, aimed to develop a CNN model to automatically detect keratoconus using standard colour-coded corneal maps and to provide a code that could be used in the clinic. This would be beneficial as colour-coded maps provide large amounts of information and clinicians are intrinsically more familiar with interpreting colour-coded maps in comparison to complex topographic indexes.

MATERIAL AND METHODS

Study population

We included tomographic images of keratoconic and healthy eyes provided by three centres: The Royal Liverpool University Hospital (UK); Sedaghat Eye Clinic, Mashhad (Iran) and The New Zealand National Eye Center (New Zealand). Patients with all stages of keratoconus were included. Images were obtained between January 2013 and January 2020. Scans were obtained using the standard 25 scans setting without pupil dilation, under scotopic conditions by trained technicians using the Pentacam HR (Oculus Optikgeräte, Wetzlar, Germany). Patients were asked to fixate on the target light and asked to blink completely just before each measurement to allow for adequate tear film coverage over the corneal surface. The examiner checked each scan and its quality before recording it and only scans of acceptable quality, reporting OK at quality score section, were included. For the control group subjects with a healthy eyes provided by three centres: The Royal Liver- pool Center (New Zealand). Patients with all stages of keratoconus and in order to preserve the maximum possible level of information for the prediction, we trained four models with one for each type of map images mentioned above, and a fifth model which used a concatenation of 4 map images (axial, corneal thickness, front and back elevation) as input. During the prediction, the output of each CNN model was the predicted class label of the input map images so that each of the trained models would give a predicted class label based on its own opinion.

We also made a sixth model by using the majority voting strategy from the above mentioned five models for each given task. For this model, a final prediction was made on the class of the input samples through a voting strategy as each of the five models will produce a prediction. More specifically, for each image sample the categorical labels from the above five different models were pooled together the argument of the maxima (abbreviated argmax) operation was applied to determine the best prediction (online supplemental figure 2). Without first performing the argmax operation, the predicted results would be directly influenced by the addition of the probability distribution.

Data preparation

In this work, different classification strategies were developed. Considering that each of the four maps contained numerous parameters that could be used to detect keratoconus and in order to preserve the maximum possible level of information for the prediction, we trained four different classification strategies were developed. Considering that each of the four maps contained numerous parameters that could be used to detect keratoconus and in order to preserve the maximum possible level of information for the prediction, we trained four

![Table 1](https://example.com/table1.png)

<table>
<thead>
<tr>
<th>Class name</th>
<th>Training</th>
<th>Testing</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy (n=134)</td>
<td>82</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>Stage 1 (n=282)</td>
<td>159</td>
<td>40</td>
<td>83</td>
</tr>
<tr>
<td>Stage 2 (n=425)</td>
<td>211</td>
<td>53</td>
<td>161</td>
</tr>
<tr>
<td>Stage 3 (n=208)</td>
<td>115</td>
<td>29</td>
<td>64</td>
</tr>
<tr>
<td>Stage 4 (n=877)</td>
<td>548</td>
<td>137</td>
<td>192</td>
</tr>
<tr>
<td>Total (n=1926)</td>
<td>1115</td>
<td>279</td>
<td>532</td>
</tr>
</tbody>
</table>
Natick, USA). In brief, all of the white pads around the four individual maps and the colour bars were removed respectively and then the four cropped individual maps were resized into 224 by 224 as the final inputs for the model learning. For the purpose of learning, the four individual maps together with the four resized individual maps were concatenated back in the original order.

**Neural network architecture**

We adopted a VGG16 model\(^4\) for this study, but in order to prevent overfitting, the number of connected weights of the top layer were reduced. Everything before the flattened layer was the same as with the standard VGG16. After the flattened layer for the 2-class task and in order to fully connect the (FC) layer with 128 outputs with a rectified linear unit (ReLU), an FC layer with 64 outputs with ReLU and an FC layer with two outputs (the final output layer representing two classes) was used with a SoftMax activation function. After the flattened layer for the 5-class task, the order was an FC layer with 128 outputs, a ReLU dropout layer with probability 0.5, FC layer with 64 outputs with ReLU dropout layer with probability 0.5, FC layer with five outputs (the final output layer representing five classes) were used with a softmax activation function. Please see online supplemental figure 2 for more information.

**Visualisation**

Saliency maps\(^4\) and Gradient-weighted Class Activation Mapping (Grad-CAM)\(^1\) were developed for visualisation of the learning, because the outputs of saliency maps and Grad-CAM were the same shape as the input image and could provide some intuition of attention. For saliency maps, the gradient of output with respect to input image from our models were computed. These gradients can highlight certain regions which contribute the most. Higher gradient values mean more contribution towards the output. In Grad-CAM, the computed gradients of output with respect to input image, produce a coarse localisation map or heat map highlighting the important regions in the image to show where the network has been focusing on. This provides a bit more explanation on the black-box characteristic of DL.

**Implementation**

We implemented our classification models by using Python V.3.7, Keras V.2.3.1 and Tensorflow V.1.14 as backends. All of the models were trained by using the Adam optimiser. We searched the optimal learning rate from the set of (10^-3, 10^-4, 10^-5, 10^-6) for each model based on the validation set. Binary cross-entropy was the loss function used for the 2-class classification models, and the categorical cross-entropy loss function for the 5-class classification models. The batch size was set to 20. All the training processes were performed on a TESLA V100 GPUs and all the testing experiments were conducted using a 2080Ti GPU.

**Evaluation metrics**

In this work, accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC) were used. In brief, Accuracy (ACC)=(TP+TN)/(TP +TN + FP+FN); Sensitivity: TP/(TP+FN); Specificity: TN/(TN+FP), where TP is a true positive, FP a false positive, TN a true negative and FN a false negative. AUC is valued between 0 and 1, where 1 means perfect classification. AUC will be more informative when the number of samples in different classes are imbalanced. De Long’s method was introduced to construct confidence intervals for accuracy, sensitivity, specificity and AUC, by bootstrapping with 2000 samples to calculate 95% CI.\(^4\)

**RESULTS**

A total of 1926 corneal tomography scans were obtained comprising of 134 healthy controls and 1702 patients with keratoconus. A total of 1394 scans were collected from UK/NZ including 102 healthy, 199 stage 1, 264 stage 2, 144 stage 3 and 685 stage 4. An additional 532 keratoconus scans (32 healthy, 83 stage 1, 161 stage 2, 64 stage 3, and 192 stage 4) collected from IRAN between February 2017 to January 2020 were used a validation set.

We will first present the results on the testing datasets of the UK/NZ datasets for the five different classification tasks, and then present the results on the IRAN dataset as the external validation set. Due to the imbalance of the classification problems, we will focus on the AUC results.

**Healthy (n=20) vs keratoconus (n=259)**

All the models using the thickness map, front elevation map, the back elevation map, the concatenated four maps and majority voting performed well, all with AUC higher than 0.80 (table 2). The model using the concatenated four maps as the input achieved the best AUC of 0.9423 (95% CI 0.8773 to 0.9942), followed by the model using the back elevation map (0.9173, 95% CI 0.8453 to 0.9793) and the majority voting model (0.8942, 95% CI 0.8115 to 0.9641).

**Healthy (n=20) vs stage 1 (n=40)**

The majority voting model achieved the best AUC of 0.90 (95% CI 0.8242 to 0.9653), followed by the model using concatenated 4 maps 0.8875 (95% CI 0.8103 to 0.9556) (table 3).

**Stage 1 (n=40) vs stage 2 (n=53)**

All models except the axial and front elevation maps performed well with AUCs higher than 0.88. The model using the back elevation map had the best performance with an AUC of 0.9153 (95% CI 0.8618 to 0.9592), followed by the majority voting model 0.9092 (95% CI 0.8557 to 0.9569). The difference, however, was marginal (table 4).
Stage 2 (n=53) vs stage 3 (n=29)
The model using the concatenated four maps achieved the best AUC of 0.8165 (95% CI 0.7398 to 0.8908) (table 5).

Five-class classification
The majority voting model achieved the best AUC of 0.888 (95% CI 0.8677 to 0.907), followed by the concatenated 4 maps model 0.8835 (95% CI 0.8641 to 0.9029) (online supplemental table 1). The confusion matrices of the above models are shown in online supplemental table 2.

From the above results of different classification tasks on the testing sets, it is apparent that the top three models are always concatenated map model, major voting model and the model using back elevation maps. The differences in performance between the aforementioned models, however, were marginal. The concatenated map model performed consistently across all classification tasks when compared with the voting model but importantly required almost 80% less computational resources. We, therefore, selected only the model using concatenated four maps for the validation testing of the IRAN dataset.

Results on the external IRAN validation dataset
In the IRAN dataset, there were 532 images comprising: Healthy (n=32); stage 1 (n=83); stage 2 (n=161); stage 3 (n=64) and stage 4 (n=192). As all the images were from patients classified as keratoconus, we were not able to evaluate the sensitivity and specificity of this model on this dataset.

The AUC was 0.9737 (95% CI 0.9653 to 0.9821) for the classification of healthy (n=32) vs keratoconus (n=500). The AUC was 0.7304 (95% CI 0.6817 to 0.7783) for the classification of healthy (n=32) vs stage 1 (n=83). The AUC was 0.8988 (95% CI 0.8649 to 0.9305) for the classification of stage 1 (n=83) vs stage 2 (n=161). The AUC was 0.8285 (95% CI 0.7793 to 0.8754) for the classification of stage 2 (n=161) vs stage 3 (n=64). The AUC was 0.859 (95% CI 0.8438 to 0.8749) for the 5-class classification.

### Table 2
<table>
<thead>
<tr>
<th>Classification results of health volunteers (n=20) and patients (n=259) with various stages of keratoconus on the testing sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy versus rest</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Axial map</td>
</tr>
<tr>
<td>Thickness map</td>
</tr>
<tr>
<td>Front elevation map</td>
</tr>
<tr>
<td>Back elevation map</td>
</tr>
<tr>
<td>Concatenated</td>
</tr>
<tr>
<td>Majority voting</td>
</tr>
</tbody>
</table>

95% CI in brackets.
AUC, area under the curve.

### Table 3
<table>
<thead>
<tr>
<th>Classification results of health volunteers (n=20) and patients with stages 1 keratoconus (n=40) on the testing sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy versus stage 1</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>Axial map</td>
</tr>
<tr>
<td>Thickness map</td>
</tr>
<tr>
<td>Front elevation map</td>
</tr>
<tr>
<td>Back elevation map</td>
</tr>
<tr>
<td>Concatenated</td>
</tr>
<tr>
<td>Majority voting</td>
</tr>
</tbody>
</table>

95% CI in brackets.
AUC, area under the curve.
The confusion matrices of the above models are shown in online supplemental table 2.

**Visualisation of the learnt features**

We also studied the area of interest (focus) of the models during its learning process. Saliency maps and Grad-CAM were developed to obtain heat maps to show where the network highlighted the important regions in the image. The following are some good examples (online supplemental figure 3).

**DISCUSSION**

Our findings suggest that accurate automated detection of keratoconus and its evolution are possible using a CNN. When provided with all four maps (axial, corneal thickness, front and back elevation) the model is automatically able to discern between keratoconus and healthy eyes with an accuracy of 99.07% (97.57%–99.43%). Moreover, it can detect different stages of keratoconus with an accuracy of 93.12% (86.75%–93.98%). We postulate that the CNN model has better accuracy than the current gold standard of human interpretation. The use of Amsler-Krumeich classification of keratoconus in four stages is related to its worldwide use in daily practice.

We believe that DL will aid in screening and in staging of keratoconus in a clinical setting, because the precise detection of early keratoconus is still challenging in daily practice.

The current gold standard of keratoconus diagnosis and staging relies on human interpretation of biomicroscopy features and tomography scans, and has previously been shown to be limited by poor reproducibility. The most commonly used parameter to determine keratoconus progression has been maximum keratometry (Kmax). Kmax is a single point reading representing the maximum curvature typically taken from the axial or sagittal anterior corneal curvature map. Kmax has numerous limitations as a single point reading is a poor descriptor of the cone morphology, a change in cone morphology (eg, a nipple cone progressing to a globular cone) can sometimes be associated with a reduction in

**Table 4** Classification results of patients with stages 1 (n=40) and 2 keratoconus (n=53) on the testing sets

<table>
<thead>
<tr>
<th>Stage 1 versus stage 2</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axial map</td>
<td>0.5699</td>
<td>1.0 (1.0 to 1.0)</td>
<td>0.0 (0.0 to 0.0)</td>
<td>5.0 (5.0 to 5.0)</td>
</tr>
<tr>
<td>Thickness map</td>
<td>0.8925</td>
<td>0.9245 (0.8571 to 0.9808)</td>
<td>0.85 (0.75 to 0.9429)</td>
<td>0.8873 (0.8281 to 0.9383)</td>
</tr>
<tr>
<td>Front elevation map</td>
<td>0.7957</td>
<td>0.8302 (0.74 to 0.913)</td>
<td>0.75 (0.6316 to 0.8751)</td>
<td>0.7901 (0.7178 to 0.8592)</td>
</tr>
<tr>
<td>Back elevation map</td>
<td>0.914</td>
<td>0.9057 (0.8276 to 0.9636)</td>
<td>0.925 (0.8462 to 0.9796)</td>
<td>0.9153 (0.8618 to 0.9592)</td>
</tr>
<tr>
<td>Concatenated</td>
<td>0.9032</td>
<td>0.9245 (0.86 to 0.9811)</td>
<td>0.875 (0.7805 to 0.9535)</td>
<td>0.8998 (0.8423 to 0.9505)</td>
</tr>
<tr>
<td>Majority voting</td>
<td>0.914</td>
<td>0.9434 (0.8837 to 1.0)</td>
<td>0.875 (0.7812 to 0.9535)</td>
<td>0.9092 (0.8557 to 0.9569)</td>
</tr>
</tbody>
</table>

95% CI in brackets. AUC, area under the curve.

**Table 5** Classification results of patients with stages 2 (n=53) and 3 keratoconus (n=29) on the testing sets

<table>
<thead>
<tr>
<th>Stage 2 versus stage 3</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axial map</td>
<td>0.6463</td>
<td>1.0 (1.0 to 1.0)</td>
<td>0.0 (0.0 to 0.0)</td>
<td>0.5 (0.5 to 0.5)</td>
</tr>
<tr>
<td>Thickness map</td>
<td>0.7683</td>
<td>0.4138 (0.2727 to 0.5769)</td>
<td>0.9623 (0.9184 to 1.0)</td>
<td>0.6880 (0.6151 to 0.7708)</td>
</tr>
<tr>
<td>Front elevation map</td>
<td>0.7561</td>
<td>0.4483 (0.2903 to 0.6071)</td>
<td>0.9245 (0.8571 to 0.9808)</td>
<td>0.6864 (0.5977 to 0.7712)</td>
</tr>
<tr>
<td>Back elevation map</td>
<td>0.7927</td>
<td>0.5172 (0.3571 to 0.6774)</td>
<td>0.934 (0.8846 to 0.9825)</td>
<td>0.7303 (0.6463 to 0.8128)</td>
</tr>
<tr>
<td>Concatenated</td>
<td>0.8537</td>
<td>0.6897 (0.5455 to 0.8333)</td>
<td>0.9434 (0.8846 to 0.9836)</td>
<td>0.8165 (0.7398 to 0.8908)</td>
</tr>
<tr>
<td>Majority voting</td>
<td>0.7317</td>
<td>0.2759 (0.1389 to 0.4194)</td>
<td>0.9811 (0.9444 to 1.0)</td>
<td>0.6285 (0.5577 to 0.7012)</td>
</tr>
</tbody>
</table>

95% CI in brackets. AUC, area under the curve.

Kmax, single point readings tend to have poor reproducibility, changes in Kmax do not correlate to changes in visual function and Kmax is limited to the anterior corneal surface, ignoring the posterior cornea, thereby having no ability to detect early or subclinical disease or early disease progression.13-37

The ability of an algorithm to detect keratoconus is based on an operationalisable distinction between a normal and ectatic cornea. A Global Consensus Panel in 2015 was able to agree on a definition of keratoconus in terms of abnormal posterior ectasia, abnormal corneal thickness distribution and clinical non-inflammatory corneal thinning.13 Various indices have been evaluated with regard to their ability to discriminate an ectatic from a normal cornea. Among these, the Smolek/Klyce and the Klyce/Maeda (KCI) have been shown to possess a good specificity and sensitivity in distinguishing between keratoconus and healthy eyes.15 The Tomographic and Biomechanical Index uses AI to combine Scheimpflug tomography and corneal biomechanical parameters to optimise ectasia detection with good sensitivity and specificity.58 This index has been shown to be even more accurate than Corneal Biomechanical Index.59

Regarding the detection of keratoconus progression, however, the Global Consensus Panel noted that specific quantitative data were lacking and, moreover, would most likely be device specific. Determinants for assessing keratoconus progression have been reviewed by Duncan et al.45 They concluded that this multitude of suggested progression parameters highlights the need for a new or standardised method to document progression.11-65

The use of colour-coded maps for DL provides more complete information with the global status of the cornea, instead of using topographic and/or tomographic numeric indices as done in the past.32 34 35 37 38 40 66-69 Numeric values can easily exemplify corneal shape, but they fail to represent the spatial gradients and distributions of corneal curvature, elevation, refractive power and thickness.

However, it is not possible to demine the superiority of the CNN over a single numerical index, as the Kmax, in view of the overall learning process, which required four maps: axial, anterior elevation, posterior elevation and pachymetry. Furthermore, for the evaluation of keratoconus, a single parameter is not sufficient.45

In this study, the images of four colour-coded maps (axial, corneal thickness, front and back elevation) were used for DL, instead of topographic and tomographic numeric indices. The reason of our choice is based on the capability of colour-coded maps to hold a larger amount of corneal information than these numeric values for this learning. The maps were obtained via tomography Scheimpflug imaging which has an advantage over Placido disk-based corneal topography as it is able to evaluate both the anterior and posterior surfaces of the cornea. Evaluation of the posterior corneal surface is essential as both curvature and elevation of the posterior corneal surface have to be considered for the detection of early keratoconus.70-72

A multiplicity of machine-learning techniques such as neural network, support vector machine, decision tree, unsupervised machine learning, custom neural network, feedforward neural network and CNN have been used in previous studies but only in four studies has a combination of colour-coded maps and CNN been used.27-29 39 The reason for opting for CNN over other machine learning methods in this study was based on the ability of the CNN to directly extract the morphological characteristics from the obtained images without preliminary learning, subsequently providing higher classification precision, especially in the field of image recognition. This study is different from the aforementioned ones as Lavric and Valentin28 did not use real clinical data, Zéboulou et al.27 focused on refractive surgery; Kamiya et al.29 used a no tomography device (AS-OCT), and Kuo et al.33 had a smaller sample size.

Our study has a number of limitations. First, the number of eyes is still modest in nature and there are only a small number of eyes in some specific groups and this may create a classification model bias. A potential strategy for this is to use generative adversarial networks to synthesise images from a small number of real images but such models would require further external validation. Second, other risk factors of keratoconus were not included in the used prediction models and such factors will be useful for further refinement of the prediction performance. Future studies incorporating such risk factors such as family history, atopy or ethnicity,73-75 may improve the overall function of the model.

Moreover, given the multicentre nature of the study a number of technicians were used to obtain the scans but the Scheimpflug tomographer has previously been reported to have good intra and interobserver repeatability in healthy patients76 and those with keratoconus.77 Finally, the adoption of more recent CNN models and tricks (eg, attention and customised loss functions) can potentially further enhance the performance of the model.

In summary, our results demonstrate that AI models provide excellent detection performance for keratoconus and can accurately grade different severities of disease and therefore have the potential to be further developed, validated and adopted for screening and management of keratoconus. Clinical implication of automated detection and screening are of considerable importance in view to their ability to provide diagnosis in shorter time, increasing in this way the patient care. Indeed, it could be deployed particularly in regions with a high burden of disease so that the CNN model may have the potential to provide earlier diagnosis of keratoconus, improve access to treatments such as corneal cross-linking and potentially reduce preventable visual loss. A larger external validation study with another study population including healthy controls is required to confirm this study’s preliminary findings.

Author affiliations
1Department of Eye and Vision Science, Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, UK
REFERENCES

Identification of more reliable parameters for the detection of change during the follow-up of mild to moderate keratoconus patients. Eye Vis 2017;4:24.


Supplemental Figure 1. Example images of four stages of keratoconus. I: Stage 1; II: Stage 2; III: Stage 3 and IV: Stage 4.
**Supplemental Figure 2.** Diagram of the proposed network architecture (Top) based on the classic VGG16 model (Bottom). We showed 5 models that use different maps as the input for the prediction. Each of these five models can work on their own for the prediction tasks, as well as can form a voting model to produce a prediction from majority voting strategy.
Supplemental Table 1. Classification results of healthy (n=20) and every stage of keratoconus (n=40, 53, 29, and 137 for stage 1, 2, 3, and 4 respectively) on the testing sets. 95% confidence interval (CI) in brackets.

<table>
<thead>
<tr>
<th>5-class</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axial Map</td>
<td>0.4910 (0.4409 0.5412)</td>
<td>0.8 (0.8 0.8)</td>
<td>0.2 (0.2 0.2)</td>
<td>0.6819 (0.6565 0.7074)</td>
</tr>
<tr>
<td>Thickness Map</td>
<td>0.8065 (0.767 0.8459)</td>
<td>0.9443 (0.9336 0.9552)</td>
<td>0.6999 (0.6507 0.7504)</td>
<td>0.879 (0.8591 0.8989)</td>
</tr>
<tr>
<td>Front Elevation Map</td>
<td>0.7061 (0.6595 0.7527)</td>
<td>0.9145 (0.9016 0.9267)</td>
<td>0.5706 (0.5163 0.6238)</td>
<td>0.8163 (0.7921 0.8404)</td>
</tr>
<tr>
<td>Back Elevation Map</td>
<td>0.8065 (0.767 0.8459)</td>
<td>0.9145 (0.9016 0.9267)</td>
<td>0.6923 (0.6473 0.7349)</td>
<td>0.879 (0.8591 0.8987)</td>
</tr>
<tr>
<td>Concatenated</td>
<td>0.8136 (0.7742 0.8530)</td>
<td>0.9484 (0.9379 0.9588)</td>
<td>0.7364 (0.6909 0.7797)</td>
<td>0.8835 (0.8641 0.9029)</td>
</tr>
<tr>
<td>Majority Voting</td>
<td>0.8208 (0.7814 0.8566)</td>
<td>0.9458 (0.9348 0.9567)</td>
<td>0.7109 (0.6686 0.7531)</td>
<td>0.888 (0.8677 0.907)</td>
</tr>
</tbody>
</table>
**Supplemental Table 2.** Confusion matrix of the testing (Top row) and Iran dataset (Bottom row).

From left to right: Health vs Keratoconus, Health vs stage 1, Stage 1 vs 2, Stage 2 vs 3, and 5-class classification. TP is true positive, FP is false positive, TN is true negative, and FN is false negative.
Supplemental Figure 3. Learning visualization. From top to bottom, four examples of concatenated inputs of stage 1, stage 2, stage 3 and stage 4 of keratoconus are displayed, respectively. From left to right: example images, saliency maps, images-overlapped saliency maps, Grad-CAMs and images-overlapped Grad-CAMs are displayed, respectively (hot (red) regions mean large contribution and cold (blue) regions means small contribution).