Corneal Topography Raw Data Classification Using a Convolutional Neural Network

PIERRE ZÉBOULON, GUILLAUME DEBELLEMANIÈRE, MAGALIE BOUVET, AND DAMIEN GATINEL

• PURPOSE: **We investigated the efficiency of a convolutional neural network applied to corneal topography raw data to classify examinations of 3 categories: normal, keratoconus (KC), and history of refractive surgery (RS).**
• DESIGN: **Retrospective machine-learning experimental study.**
• METHODS: **A total of 3,000 Orbscan examinations (1,000 of each class) of different patients of our institution were selected for model training and validation. One hundred examinations of each class were randomly assigned to the test set. For each examination, the raw numerical data from "elevation against the anterior best fit sphere (BFS)," "elevation against the posterior BFS" "axial anterior curvature," and "pachymetry" maps were used. Each map was a square matrix of 2,500 values. The 4 maps were stacked and used as if they were 4 channels of a single image.A convolutional neural network was built and trained on the training set. Classification accuracy and class wise sensitivity and specificity were calculated for the validation set.**
• RESULTS: **Overall classification accuracy of the validation set (n = 300) was 99.3% (98.3%-100%). Sensitivity and specificity were, respectively, 100% and 100% for KC, 100% and 99% (94.9%-100%) for normal examinations, and 98% (97.4%-100%) and 100% for RS examinations.**
• CONCLUSION: **Using combined corneal topography raw data with a convolutional neural network is an effective way to classify examinations and probably the most thorough way to automatically analyze corneal topography. It should be considered for other routine tasks performed on corneal topography, such as refractive surgery screening. (Am J Ophthalmol 2020;219:33–39.**

DEEP LEARNING USING CONVOLUTIONAL NEURAL networks (CNNs) is currently the state-of-the-art technique for image recognition. It has been used successfully in a wide range of applications, both outside of medicine and in various medical fields.[1–3] Modern ophthalmology uses many image rendering tools to analyze various ocular structures and is thus a good candidate to benefit from deep learning. CNNs showed good accuracy in several different ophthalmology tasks, such as fundus photograph–based diagnosis for age-related macular degeneration,[4] diabetic retinopathy,[5] or glaucoma.[6] Moreover, optical coherence tomography–based diagnosis of retinal conditions[7,8] and image segmentation[9–11] has also been successfully achieved. Most of the research using deep learning in ophthalmology currently focuses on retinal diseases, probably because of the precedence of the widespread use of medical imaging in retinal disease compared with anterior segment disease.

Computerized corneal topography is probably one of the oldest digital anterior segment examination tools[12] and is still considered the gold standard for many corneal diseases (eg, keratoconus [KC]). Several screening and diagnostic algorithms have been developed based on corneal parameters engineered by humans.[13–22] In contrast, CNNs automatically generate relevant features from the data to perform a required task and can sometimes outperform algorithms based on human-derived features.

Corneal tomography presents the results as color maps of the different corneal parameters. Although these maps are not images of the eye *per se*, they are nonetheless presented as color images, and it is only natural to try to harness the power of CNNs for corneal tomography analysis. Two studies[23,24] applied such algorithms on corneal tomography color maps to classify examinations into different diagnoses. Although these approaches showed good results, one could argue that using color map images instead of the actual raw numeric values constituting these maps might lead to a loss of information. In addition, color map rendering depends highly on the chosen color scale, which can vary from one tool to another, or even from one patient to the another. This issue might hamper the generalizability of algorithms trained on color maps. Raw topographical numeric data are usually square matrixes of 10,000 elements in which each number is then color coded to render the color map. From a computer's point of view, this is similar to a gray-scale image and can therefore be fed to a CNN. Color images are seen by computers as 3 matrixes of the same size stacked together, each representing a color channel (red, green, and blue). The color of each pixel results from 3 numbers, 1 for each color channel. Therefore, color channels represent different information

AJO.com Supplemental Material available at AJO.com.

Accepted for publication Jun 3, 2020.

From the Department of Ophthalmology, Rothschild Foundation, Paris, France (P.Z., G.D., M.B., D.G.); and CEROC (Center of Expertise and Research in Optics for Clinicians), Paris, France (D.G.)

Inquiries to Pierre Zéboulon, Fondation Ophtalmologique Rothschild, 25 rue Manin, 75019 Paris, France; e-mail: pierrezeboulon@gmail.com

0002-9394/$36.00
https://doi.org/10.1016/j.ajo.2020.06.005

of the same spatial location of an image. Similarly, the different corneal tomography maps can also be seen as different information of the same corneal locations. Clinicians routinely compare the same locations on each map to make their diagnoses. For example, in a KC, the steepest area is usually approximately in the vicinity of the thinnest area and the highest elevation of both anterior and posterior corneal surfaces.

Therefore, in this study, we investigated the possibility of using numeric data matrixes instead of color maps to train a CNN for a classification task. Specifically, we used 4 maps that are frequently used in clinical practice, stacked together as if they were 4 color channels of a single image to classify examinations into 3 categories: normal, KC, and history of refractive surgery (RS).

METHODS

• **DATA AND PATIENTS:** This experimental machine-learning study was approved by the Institutional Review Board at the Rothschild Foundation and followed the tenets of the Declaration of Helsinki. Informed consent was obtained from all patients. Our Orbscan (Bausch & Lomb, Bridgewater, NJ) database was exported using the batch export functionality both as image files and the underlying numeric data matrixes represented by each color map. We selected 3,000 examinations in total, 1,000 per class (normal, KC, history of RS). All 3,000 examinations were obtained from different patients, and only 1 eye per patient was selected. We balanced the examinations to have exactly 500 left eyes and 500 right eyes in each class. The selection process was as follows: consecutive examinations were preselected by a resident and reviewed by a corneal tomography expert with at least 5 years of experience in a corneal and RS department. The preselection criteria for KC were: an anterior curvature map showing 1 of the classic KC patterns described by Rabinowitz et al.[25] associated with corneal thinning. The criteria for RS examinations were an oblate anterior surface (flat in its center), a prolate posterior surface (steep in its center), central corneal thinning, and lower central curvature values compared with the periphery (cases of myopic laser surgery). Finally, normal examinations were preselected if no corneal condition could be detected. If the case seemed to match 1 of the 3 desired diagnoses, the patient's electronic file was retrieved to confirm the diagnosis, and the examination was excluded or included in the study dataset. This process was iterated until the target number of 3,000 examinations that were evenly distributed was reached. Bad quality examinations with many artefacts or too many missing values were not included.

• **DATA PREPROCESSING:** The numeric data from 4 commonly used maps in clinical practice were selected for each examination: "elevation against the anterior best fit sphere (BFS)"; "elevation against the posterior BFS"; "axial anterior curvature"; and "pachymetry." In an attempt to mimic the human diagnostic process, we specifically chose these 4 maps because they are the most commonly used routinely in our institution. Furthermore, because CNNs are great tools to detect patterns, we used elevation relative to BFS maps because they tend to highlight diagnosis specific patterns. Each map was a square matrix of 100×100 numerical values. To account for the numerous missing values in the periphery of the curvature maps, all matrixes were cropped to their central 50×50 elements. The 4 maps thus constituted a total of $4 \times 2{,}500 = 10{,}000$ numerical values. The four matrixes were stacked to constitute a 3-dimensional matrix with dimensions of $50 \times 50 \times 4$ elements (Figure 1). Each matrix was padded with 3 rows of zeros on each side to obtain a $56 \times 56 \times 4$ matrix and to match the network architecture. We randomly separated the dataset in a training and a test set of 2,700 and 300 examinations, respectively, with a strict balance of class in each set (900 examinations of each class in the training set and 100 in the test set). We used 10-fold cross validation on the training set to assess internal validity but not for hyperparameter selection. Missing values were replaced from all maps from the training set by the mean value for each matrix element, which was the simplest data replacement strategy. Data from each type of map were scaled between 0 and 1 using data from all patients in the training set. Missing value replacement and data scaling was performed on the test set using the parameters acquired from the training set. This process was performed in the exact same way in each cross-validation set.

• **MACHINE-LEARNING METHODOLOGY:** We built a CNN with an architecture inspired by the ResNet[26] with fewer hidden layers and 4 input channels. The complete architecture is detailed in Figure 1 (also see Supplemental Material). An Adam[27] optimizer and the binary cross entropy loss function were used. Initialization of the model's weights was performed as described by He et al.[28] The training set was trained during 15 epochs with a learning rate of 0.0001 and a batch size of 2 (see Glossary in Supplemental Material). Overall accuracy was used to assess performance in both the training and validation sets during learning. Training data augmentation was performed as follows-for each epoch, all matrixes were fed to the network both in their original orientation and flipped horizontally. In addition, a random rotation within a −10- to 10-degrees range was applied to the input matrix, changing the angle for each epoch. This augmentation process helps increase accuracy and decrease the risk of overfitting.

• **MEASUREMENTS:** For each cross-validation set, we reported the training and validation results of the epoch yielding the highest validation set accuracy. Class wise sensitivity and specificity of the test set were reported.

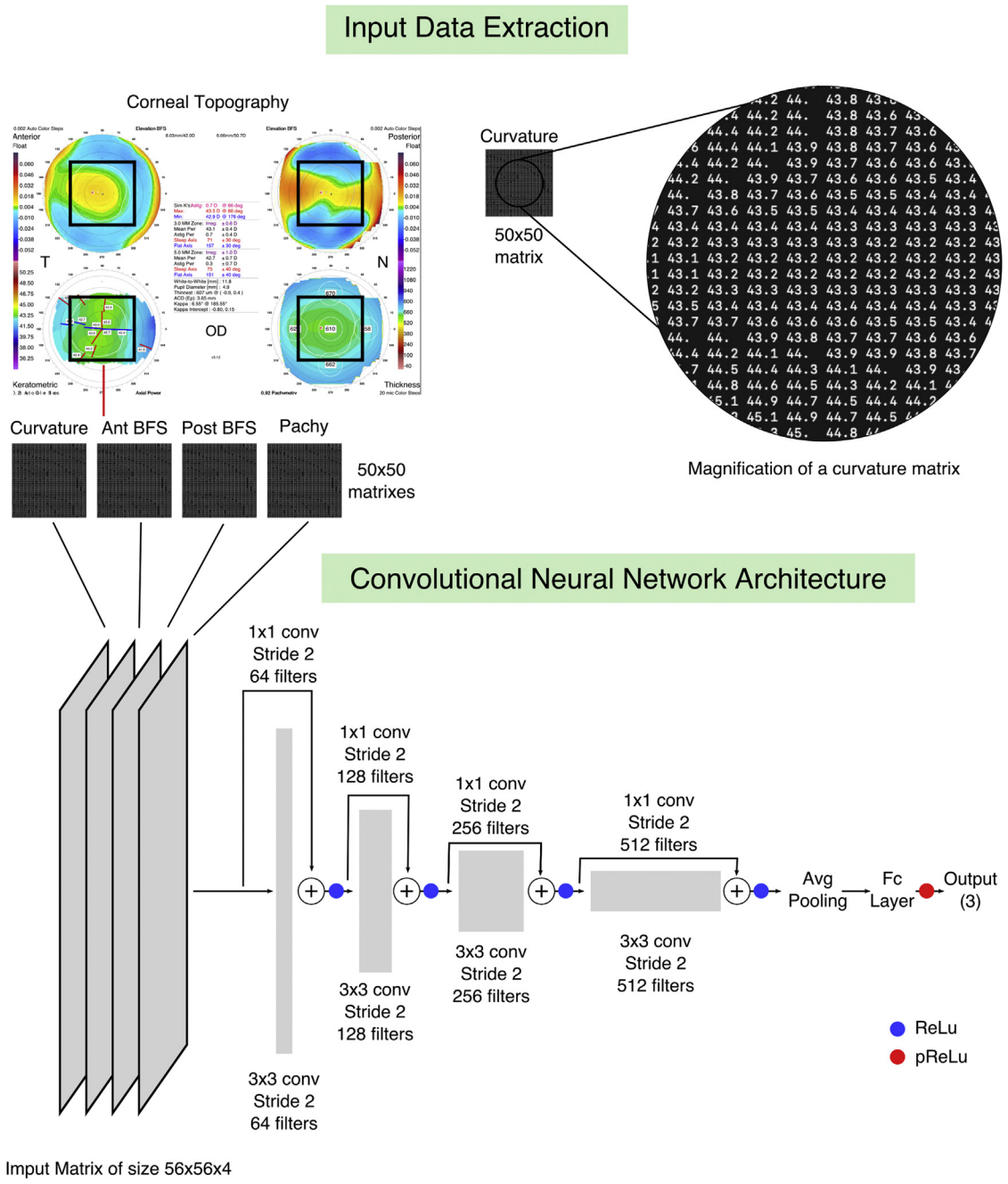

FIGURE 1. Method used to create the input matrixes and of the architecture of the convolutional neural network used. Each matrix was padded by three rows of zeros. Every convolution layer was followed by a batch normalization layer. Ant BFS = anterior best fit sphere matrix, Avg = average; conv = Convolution; Fc = fully connected; Pachy = pachymetry matrix; Post BFS = posteriori best fit sphere matrix; pReLu = leaky ReLu with a learnable parameter; ReLu = rectified linear unit.

Bootstrapping was performed with 2,000 iterations on the test set of 300 examinations to compute 95% confidence intervals for each measurement.

For demographics purpose, KC examinations were separated into 4 severity categories according to the mean keratometry, using the cutoffs of the Amsler-Krumeich classification. The χ^2 test was used to compare the number of examinations of these categories on the training and test sets. A P value $<.05$ was considered significant.

• **CLASS ACTIVATION MAPPING:** We performed class activation mapping as described previously[29] to visualize which

TABLE 1. Patients and Topography Main Characteristics

	Steep Axis Keratometry (D) (mean ± SD)	Flat Axis Keratometry (D) (mean ± SD)	Astigmatism (D) (mean ± SD)	Minimum Pachymetry (μm) (mean ± SD)	Patient Age (y) (mean ± SD)	Fraction of Males (%)
Normal (n = 1,000)	44.7 ± 1.7	43.4 ± 1.6	1.2 ± 1.0	550.0 ± 38.4	47.0 ± 16.5	42.9%
Keratoconus (n = 1,000)	50.3 ± 5.3	45.3 ± 4.3	4.3 ± 2.6	431.4 ± 68.3	39.6 ± 12.0	63.9%
RS (n = 1,000)	40.2 ± 2.1	39.3 ± 2.1	0.9 ± 0.6	453.7 ± 74.9	46.4 ± 12.6	39.3%

D = diopter; RS = history of refractive surgery.

part of the input matrix was used to predict each class and present graphic examples. The results are presented as a color heatmap in which hotter colors represent the matrix elements that were the most useful to the model for making the prediction.

• **PROGRAMMING LANGUAGE AND LIBRARIES:** All calculations, algorithms, and figures were done in Python 3.6. PyTorch library was used to build and train the neural network. Matplotlib library was used for plotting of class activation maps.

RESULTS

PATIENTS' AGE, SEX, AND MAIN TOPOGRAPHICAL CHARACTERISTICS are reported in Table 1.

Severity repartition of the KC examinations in the training and test sets are shown in Table 2. KC severity distribution was comparable in both sets ($P = .99$).

Overall accuracy for each cross-validation training and validation set is shown in Table 3. Overall accuracy of the test set (with its 95% confidence interval) was 99.3% (98.3%-100%). (Learning curves are available in the Supplemental Material).

Class-wise sensitivity and specificity with their 95% confidence intervals are shown in Table 4 (receiver-operating characteristics curves for each class are available in the Supplemental Material). Only 2 examinations of the 300 test sets were misclassified, which were 2 RS examinations that were classified as normal.

Class activation mapping (Figure 2) showed that the region used by the network to predict the KC class was the region that combined the steepest curvature, the thinnest point, and the highest elevations of the anterior and posterior surfaces. For RS examinations, the network used the region that combined the flattest curvature, the thinnest point, and the lowest and highest elevations of the anterior and posterior surfaces, respectively.

TABLE 2. Keratoconus Severity Repartition in the Training and Test Sets Based on Mean Keratometry

	Mean Keratometry			
Keratoconus examinations, D	<48	48-53	53-55	>55
Training set (n = 900)	543	232	46	79
Test set (n = 100)	64	24	6	6

D = diopter.

DISCUSSION

WE FOUND THAT A CNN USED ON THE COMBINED NUMERIC data matrixes from 4 corneal tomography maps could accurately classify examinations of 3 diagnoses: normal, KC, and history of RS.

To our knowledge, only 2 other studies applied CNNs to corneal tomography map classification. In both studies, different technologies than Orbscan were used for data acquisition. Because each tool had its own spatial resolution and coverage, it could affect the results and their comparability to this work. In the first study,[23] authors used only curvature color maps and trained a model on a 3,000 examinations dataset, half of which were KC and the other half normal corneas. The model reached an accuracy of 99.33% on the test set after 38 epochs.

In the second study,[24] 6 different color maps for each of the 543 participants were used to train 6 different models and combined the results of each model to classify examinations between normal and KC. The resulting model reached an accuracy of 99.1%.

When considering only KC and normal examinations, our model reached a 100% accuracy after only 9 epochs. Also, we used input matrixes with >10 times fewer elements (50 × 50 × 4 elements in our case compared with 180 × 240 × 3 and 224 × 224 × 3 elements, respectively). We showed we could achieve a similar accuracy with an additional third diagnostic class, smaller input matrixes, and a smaller model architecture. This was mainly due to

TABLE 3. Accuracy of the Training Set and Validation Set of Each Cross-Validation Set

	Cross-Validation Set Number									
	1	2	3	4	5	6	7	8	9	10
Training set accuracy (n = 2,430), %	98.8	99	98.8	99.2	99	99.4	99.2	99.2	98.5	99
Validation set accuracy (n = 270), %	98.9	99.6	99.6	98.9	98.8	99.6	98.9	99.6	98.5	99.6

TABLE 4. Class Wise Results in the Validation Set With 95% Confidence Intervals

	Sensitivity	Specificity
Keratoconus (n = 100), %	100 (100-100)	100 (100-100)
Normal (n = 100), %	100 (100-100)	99 (94.9-100)
History of refractive surgery (n = 100), %	98 (97.4-100)	100 (100-100)

the fact that we were able to use the numeric data matrixes instead of color map images and combined them in a single matrix. This approach had several advantages. First, it required a much smaller computational cost and was more efficient. Second, it is probably easier to generalize to other topographs. We could probably use transfer learning to retrain our algorithm on other numeric data matrixes of topographs with a little number of examinations. Variations in color scales in the different topographs might make this process more complicated in the other cited approaches. Finally, our model used spatial concordance of features on each map as human experts would, when analyzing an examination. In addition, using only the central part of each map was at least as efficient as using the whole map for this classification task.

We believe that using the numeric data matrixes instead of the color maps derived from them was a more accurate and natural approach. Color maps are simplified representations of the matrixes that necessarily imply a loss of data and may be altered by the image compression process if they are exported in a compressed format. Moreover, resizing the image to fit the CNN architecture can alter the original data. Therefore, when numeric matrixes are available, they should be used instead of the color maps to train CNN models.

The use of class activation mapping (Figure 2) helped to understand how the network made its prediction. It showed that our model classified KC and RS examinations by using the same regions of the input matrix as a human would have on the color maps. Because the 4 matrixes were combined in 1 single input matrix, the class activation mapping result was the same for each map and did not distinguish which information from each map had been used by the network. Nonetheless, the area of activation included the regions of interest from each map. For the normal class, most of the input seemed to activate the network; this probably occurred because there was no specific characteristic of a normal examination.

Regarding the classification results of KC examinations, we found 100% accuracy despite having 64% of KCs with a mean keratometry of <48 diopters. This result suggested that the algorithm was efficient even for mild stages of the disease.

Many other algorithms have been developed to classify tomography examinations.[13–22] Most of them use human defined or selected features. CNNs create features automatically and are not biased by human assumptions, which make them more likely to discover unknown relationships in the data. Previous studies that described algorithms to classify normal and KC examinations were hardly comparable due to differences in methodology, population, and sample size. Nonetheless, most models were trained and tested on smaller samples, and reported accuracies were between 80%[14] and 100%.[15]

This study had several limitations. First, the matrixes were cropped to their 50 × 50 central elements to account for the limitation in coverage of the Placido of the Orbscan and the resulting frequent missing values. This might impair the model's performance if the region of interest (eg, the apex of a KC) is located outside the analyzed zone. We did not exclude those patients, but we did not specifically evaluate the classification performance in these cases, either. It could be interesting to see if the model performs as well in eccentric KC. In addition, of the relatively small area analyzed, Orbscan is known to generate artefacts of the posterior surface after RS. Nonetheless, the good performance of our model suggested that combining the numeric data of 4 different maps to the power of CNN might be sufficient to overcome those limitations. Also, the initial selection of cases was made by reading the tomographic maps. This could have induced a selection bias, and harder or subtle cases might have been overlooked. Finally, we did not evaluate the performance of the model on forme fruste KC (FFKC). This kind of model could nonetheless be a powerful tool for FFKC detection. Therefore, we plan to retrain the model, with FFKC as an additional class, using the pretrained model as it is now. We could probably retrain the last layer only because most features must have been learned from the current dataset. This would dramatically improve training time.

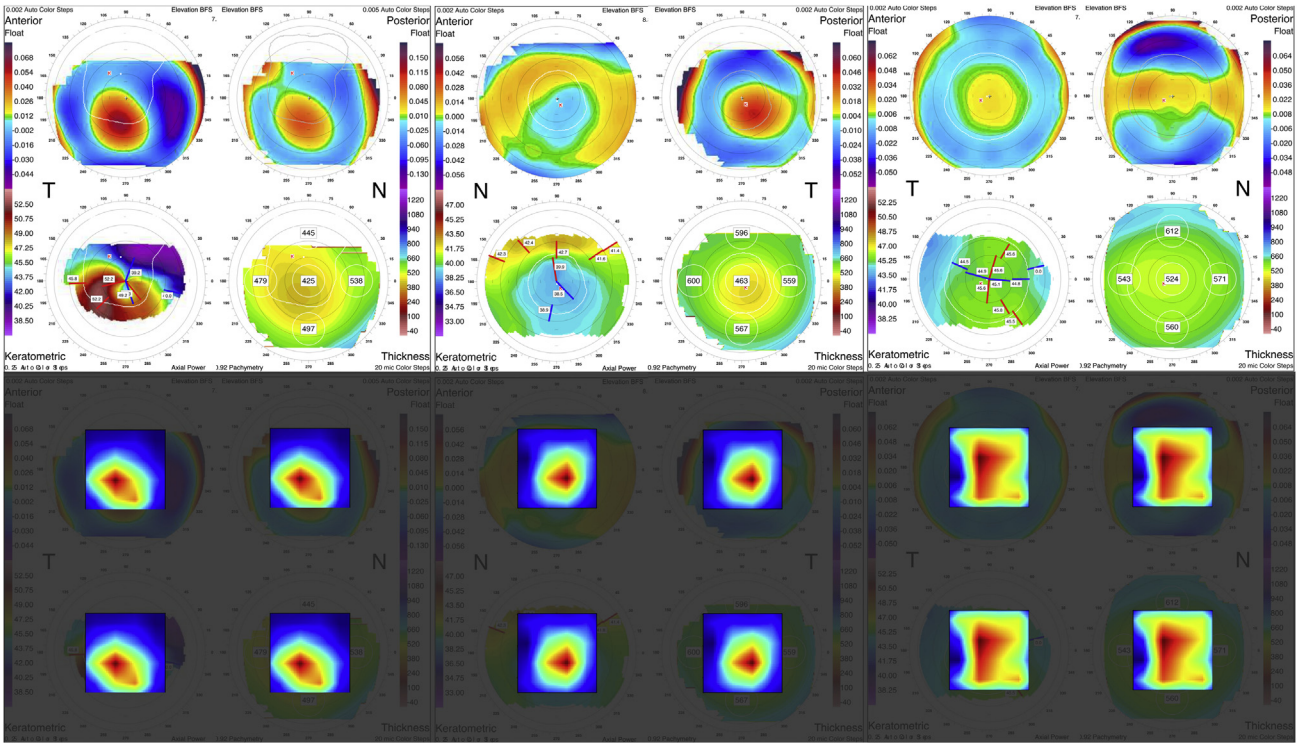

FIGURE 2. Examples of examinations of each class with the corresponding class activation maps generated from the model. Left: keratoconus; center: history of refractive surgery; right: normal; bottom row: class activation maps overlay. Hotter colors on the activation maps means more activation.

In summary, we showed that analyzing combined corneal tomography numeric data matrixes with a CNN was effective to classify examinations. We believe that it is the most comprehensive way to automatically analyze corneal tomography, and that it should be extended to the diagnosis and screening of other diseases and be used routinely in clinical practice.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

PIERRE ZÉBOULON: CONCEPTUALIZATION, METHODOLOGY, Data curation, Software, Writing - original draft. **Guillaume Debellemanière:** Data curation, Writing - review & editing. **Magalie Bouvet:** Data curation, Writing - review & editing. **Damien Gatinel:** Supervision, Methodology, Writing - review & editing.

FUNDING/SUPPORT: NONE.

Financial Disclosures: None for P.Z., G.D. and M.B. Consultant for Physiol, NIdek, Heidelbert Ing, Alcon Wavelight (D.G.). All authors attest that they meet the current ICMJE criteria for authorship.

Other Acknowledgments: None.

REFERENCES

1. Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. *Jpn J Radiol* 2019;37(1):15–33.
2. Xu B, Kocyigit D, Grimm R, Griffin BP, Cheng F. Applications of artificial intelligence in multimodality cardiovascular imaging: a state-of-the-art review. *Prog Cardiovasc Dis* 2020;63(3):367–376.
3. Chan S, Reddy V, Myers B, Thibodeaux Q, Brownstone N, Liao W. Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatol Ther* 2020;10(3):365–386.
4. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology* 2019;126(4):565–575.
5. Raman R, Srinivasan S, Virmani S, Sivaprasad S, Rao C, Rajalakshmi R. Fundus photograph-based deep learning

algorithms in detecting diabetic retinopathy. *Eye (Lond)* 2019;33(1):97–109.
6. Asaoka R, Tanito M, Shibata N, et al. Validation of a deep learning model to screen for glaucoma using images from different fundus cameras and data augmentation. *Ophthalmol Glaucoma* 2019;2(4):224–231.
7. Lee CS, Baughman DM, Lee AY. Deep learning is effective for the classification of OCT images of normal versus age-related macular degeneration. *Ophthalmol Retina* 2017;1(4):322–327.
8. Li F, Chen H, Liu Z, Zhang X, Wu Z. Fully automated detection of retinal disorders by image-based deep learning. *Graefes Arch Clin Exp Ophthalmol* 2019;257(3):495–505.
9. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology* 2018;125(4):549–558.
10. Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express* 2017;8(7):3440–3448.
11. Maloca PM, Lee AY, de Carvalho ER, et al. Validation of automated artificial intelligence segmentation of optical coherence tomography images. *PLoS One* 2019;14(8):e0220063.
12. Klyce SD. Computer-assisted corneal topography. High-resolution graphic presentation and analysis of keratoscopy. *Invest Ophthalmol Vis Sci* 1984;25(12):1426–1435.
13. Maeda N, Klyce SD, Smolek MK, Thompson HW. Automated keratoconus screening with corneal topography analysis. *Invest Ophthalmol Vis Sci* 1994;35(6):2749–2757.
14. Maeda N, Klyce SD, Smolek MK. Neural network classification of corneal topography. Preliminary demonstration. *Invest Ophthalmol Vis Sci* 1995;36(7):1327–1335.
15. Smolek MK, Klyce SD. Current keratoconus detection methods compared with a neural network approach. *Invest Ophthalmol Vis Sci* 1997;38(11):2290–2299.
16. Accardo PA, Pensiero S. Neural network-based system for early keratoconus detection from corneal topography. *J Biomed Inform* 2002;35(3):151–159.
17. Souza MB, Medeiros FW, Souza DB, Garcia R, Alves MR. Evaluation of machine learning classifiers in keratoconus detection from orbscan II examinations. *Clinics* 2010;65(12):1223–1228.
18. Arbelaez MC, Versaci F, Vestri G, Barboni P, Savini G. Use of a support vector machine for keratoconus and subclinical keratoconus detection by topographic and tomographic data. *Ophthalmology* 2012;119(11):2231–2238.
19. Smadja D, Touboul D, Cohen A, et al. Detection of subclinical keratoconus using an automated decision tree classification. *Am J Ophthalmol* 2013;156(2):237–246.e1.
20. Kovács I, Miháltz K, Kránitz K, et al. Accuracy of machine learning classifiers using bilateral data from a Scheimpflug camera for identifying eyes with preclinical signs of keratoconus. *J Cataract Refract Surg* 2016;42(2):275–283.
21. Ruiz Hidalgo I, Rodriguez P, Rozema JJ, et al. Evaluation of a machine-learning classifier for keratoconus detection based on Scheimpflug tomography. *Cornea* 2016;35(6):827–832.
22. Ruiz Hidalgo I, Rozema JJ, Saad A, et al. Validation of an objective keratoconus detection system implemented in a Scheimpflug tomographer and comparison with other methods. *Cornea* 2017;36(6):689–695.
23. Lavric A, Valentin P. KeratoDetect: keratoconus detection algorithm using convolutional neural networks. *Comput Intell Neurosci* 2019;2019:8162567.
24. Kamiya K, Ayatsuka Y, Kato Y, et al. Keratoconus detection using deep learning of colour-coded maps with anterior segment optical coherence tomography: a diagnostic accuracy study. *BMJ Open* 2019;9(9):e031313.
25. Rabinowitz YS, Yang H, Brickman Y, et al. Videokeratography database of normal human corneas. *Br J Ophthalmol* 1996;80(7):610–616.
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Available at: ; 2015. http://arxiv.org/abs/1512.03385;. Accessed November 30, 2019.
27. Kingma DP, Ba J. Adam: a method for stochastic optimization. *ArXiv14126980 Cs*. Available at: ; 2017. http://arxiv.org/abs/1412.6980;. Accessed May 13, 2020.
28. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. Available at: ; 2015. http://arxiv.org/abs/1502.01852;. Accessed May 13, 2020.
29. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Available at: ; 2015. http://arxiv.org/abs/1512.04150;. Accessed April 11, 2020.