Appendix S1. Details of Strategies

(a) RDN: Cataract images contain many redundant parts (pixels). Therefore, if the CNN is trained by the original images without any pre-processing, the network reflects unnecessary or unimportant visual information for classifying the image. Consequently, a suboptimal representation might be learned, leading to lower classification performance. Hence, we extracted the core region from the images using RDN and then classified them using the classification network. To automatically localize the necessary region, we applied the faster R-CNN method[1] which is widely used for object detection tasks in the computer vision field. For the feature learner of the RDN and classification network, we used ResNet-50 and ResNet-18, respectively. To accelerate the training procedure, we used the ImageNet 1k pre-trained weights for the RDN, obtained from the official Pytorch repository (See Figure S3 for cropped images by RDN).

(b) Data augmentation and transfer learning: Most medical image datasets[2,3] contain a smaller amount of data compared to standard vision datasets (e.g., ImageNet 1K[4]), which can be obtained easily such as through crowd-sourcing or artificial synthesis. However, it has been shown that neural networks tend to memorize specific details of data samples (that is, overfitting problem[5]) rather than learn salient fragments or visual patterns of data. Because these factors cause the neural network to generalize in other unseen data, we used generalization techniques to ensure that networks did not overfit in the training data. The data augmentation method uses more training data by adding slightly modified copies of original data or newly created synthetic data from the original data. The purpose of data augmentation is to provide neural networks the ability to generalize. Among the various types of augmentation methods, we exploited the general ones, such as cropping, blurring, flipping, and rotating (For specific techniques and their hyper-parameters, see Table S1). On the other hand,

the transfer-learning method is to train the "initialized" network whose weights are pre-trained in another dataset. The purpose of transfer learning is to accelerate the training procedure. Additionally, it prevents overfitting in the training data, especially when the size of the training data is very small. In our system, we pre-trained the first to fourth residual layers and fully connected layers of ResNet-18 using the ImageNet 1k dataset.

(c) GCE Loss: Clinicians' diagnoses reflect the individual's bias; they may grade different severity levels for the same image. This problem, known as the noise label problem[6] in the field of machine learning, leads to the model being trained with inconsistent (or wrong) information, making it difficult to classify the correct diagnosis. The standard loss function in image classification, cross-entropy (CE) loss, is known to be vulnerable to this setting. Therefore, to alleviate this problem, we used the GCE[6] function, which enables robust learning by making the training procedure less prone to dataset bias; it incorporates the MAE loss[7], which alleviates information bias of labels from individual to CE loss.

(d) CB Loss: The medical image dataset usually contains a larger number of images of healthy persons than of persons with severe conditions. When a neural network is trained on an imbalanced dataset, it tends to classify images into classes that occupy a large portion of the training data.[8] Therefore, if the model is trained on the cataract dataset with standard CE loss, the network can be trained to classify patients in highly severe conditions as healthy ones, which can lead to undesirable situations in practice. To address this, we used CB loss[8], which is proposed for robust training against the class imbalance of the dataset (See Figure S6 for details of the DL system's concept figures).

Appendix S2. Training and Optimization of DL system

We built a two-step process to train our system. In the first step, we trained the ImageNet-1K pre-trained RDN. In the second step, we freeze the trained RDN and train the classification network. For the first step, we regressed the bounding box location and trained the network using standard smooth L1 losses[1]. For the second step, we predicted two types of grades for each image type and trained the network using the combined GCE loss and CB method (CB-GCE loss). To predict two types of grades for each image type, we split the classifier neurons in half to predict one type of grade and the other half to predict the other type of grade. For example, for the retro-illumination grading task, the output dimension of a fully connected layer is 12, where the former six neurons are linked to the CO grade and the latter six neurons to a PSC grade. Then, we calculated the final loss by summing two CB-GCE losses that correspond to each prediction (NO and NC for slit-lamp images; CO and PSC for retro-illumination images). We used different strategies to optimize each step. For the first step, we used the SGD optimizer with a learning rate of 0.005, momentum of 0.9, and weight decay coefficient of 0.0005. Because the localization procedure is a simple task, a short learning schedule (ten epochs with mini-batch size of eight) is sufficient. No augmentation strategies were used in this study. Finally, we resized and normalized cropped images from the RDN from 0 to 1. For the second step, we used the Adam optimizer with default hyperparameters, except for the weight decay coefficient and the learning rate. For the weight decay coefficient, we set the value to 0.0005. The learning rate was initially set to 0.001 for the fully connected layer and 0.0001 for the last residual block. Then, we decayed the learning rate by a factor of 0.95 for every ten epochs for convergence of the training. We trained the network for 200 epochs with a mini-batch size of 32. To achieve the cataract classification task, we first regressed the location of the cropped image that contained only salient regions and then predicted its grading.

Therefore, the regression and classification labels were required for network training, the ground truths for classification were obtained by physicians, and the ground truths for detection were obtained by cropping only the nuclear part of the image using the ‘Labeling’ package, which is a graphical image annotation tool provided in Python.

References

1. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE transactions on pattern analysis and machine intelligence 2017;39:1137-49.
2. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2097-106.
3. Vázquez D, Bernal J, Sánchez FJ, et al. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. Journal of healthcare engineering 2017;2017:4037190.
4. Krizhevsky A, Sutskever I, Hinton GEJAinips. Imagenet classification with deep convolutional neural networks. 2012;25:1097-105.
5. Ying X. An overview of overfitting and its solutions. Journal of Physics: Conference Series; 2019: IOP Publishing. p. 022022.
6. Zhang Z, Sabuncu MRJapa. Generalized cross entropy loss for training deep neural networks with noisy labels. 2018.
7. Ghosh A, Kumar H, Sastry P. Robust loss functions under label noise for deep neural networks. Proceedings of the AAAI Conference on Artificial Intelligence; 2017.
8. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 9268-77.