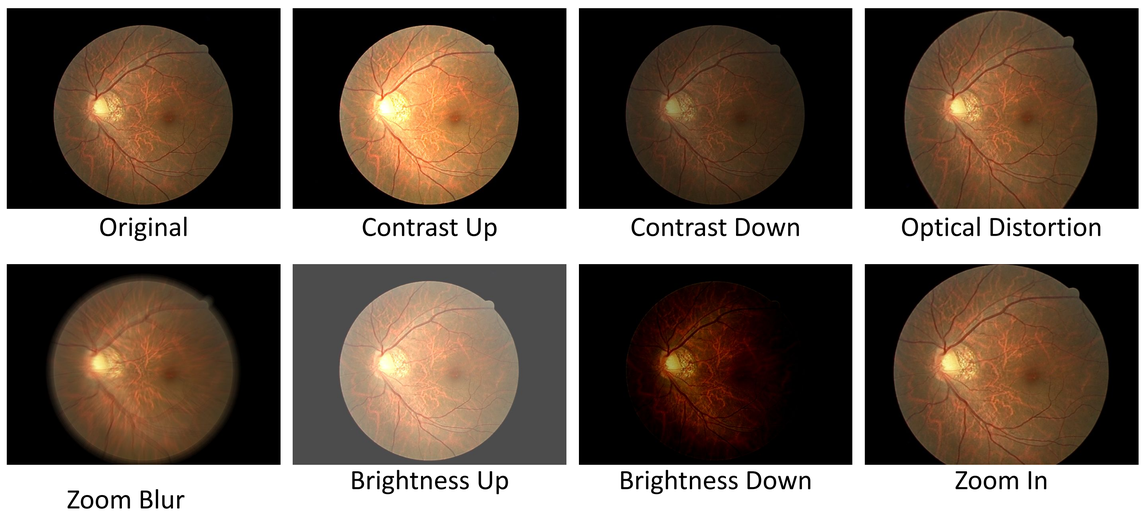
Introduction

- Utilizing AI to read of fundus images, e.g. Diabetic Retinopathy AI Diagnosis Systems such as Idx-DR® and EyeArt, is expected to make the diagnosis process more efficient. [1]
- However, no research has addressed the effect of various noises on fundus images on the performance of AI diagnoses.

Examples of noises



Purpose

1. To verify the **robustness** of multilabel classification AI that detects diseases in fundus images
2. To improve the training protocol based on that, and evaluate the change in the AI's robustness characteristics

Robustness
: the extent to which AI is resilient to input disturbances

Methods

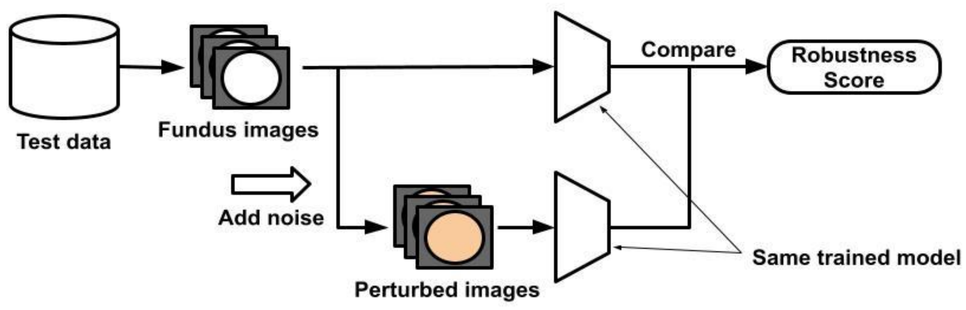
Dataset and Model

- About 400,000 conventional wide-angle color fundus images collected at Jichi Medical University
- Among them, there are approximately 270,000 images without any disease.
- Images with some diseases were divided into train, validate and test dataset, and then a multi-label classification model was trained with the dataset.
- Most images without any disease were used in the training phase by being sampled in each epoch.

Top 10 Disease Based on Image Counts

Disease Name	Image Count
S1	50005
H1	34799
S2	13442
Glaucoma	8380
H2	5411
Drusen	3421
Cataract	3334
Retinal Fibrosis	3053
Optic Disc Cupping	2364
Diabetic Retinopathy	2241

Robustness Score: Metric for Robustness

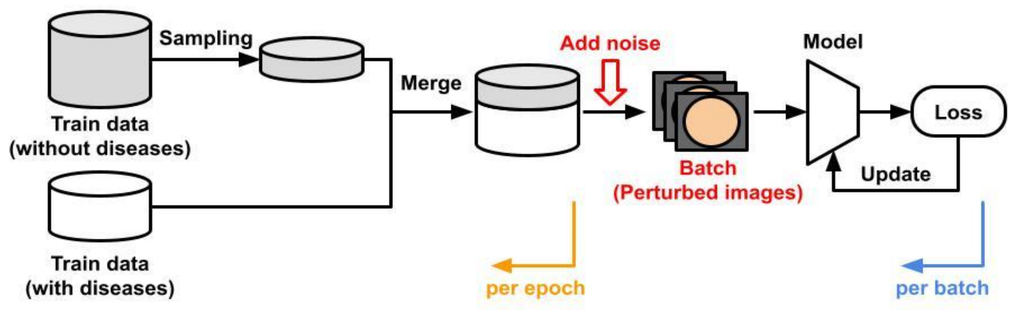


1. Input both an original image and its perturbed variant into a trained model and obtain the prediction for each image about which diseases were detected
2. Calculate the similarity of each prediction using the Jaccard coefficient. Specifically, for the sets of predicted diseases from each image, denoted as A and B respectively, the Jaccard coefficient $J(A,B)$ is calculated as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

3. The Robustness Score is calculated as the averaged Jaccard coefficient across all images in the test data.
   The Robustness Score ranges from 0% to 100, with higher values indicating a stronger resilient to perturbations.

Experiments



1. Using **Citadel Lens**, we calculated the Robustness Score for various perturbations.
2. By training the model with perturbed fundus images in categories that had low robustness scores, we confirm that the model develops resistance to those perturbations.

栃木県眼科集談会
利益相反開示

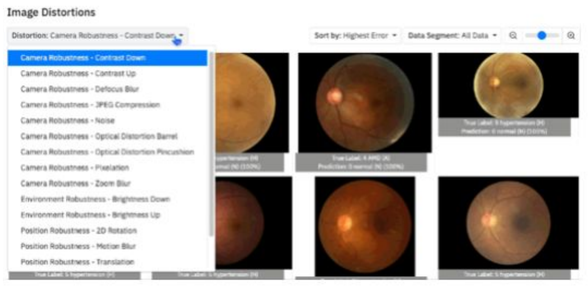
筆頭演者：松本　大蔵

高橋秀徳：【I】DeepEyeVision株式会社、【P】

Citadel Lens

- A software provided by Citadel AI that performs technical verification of AI models and datasets, to measure and accelerate AI quality improvement.

Testing Capabilities of Citadel Lens



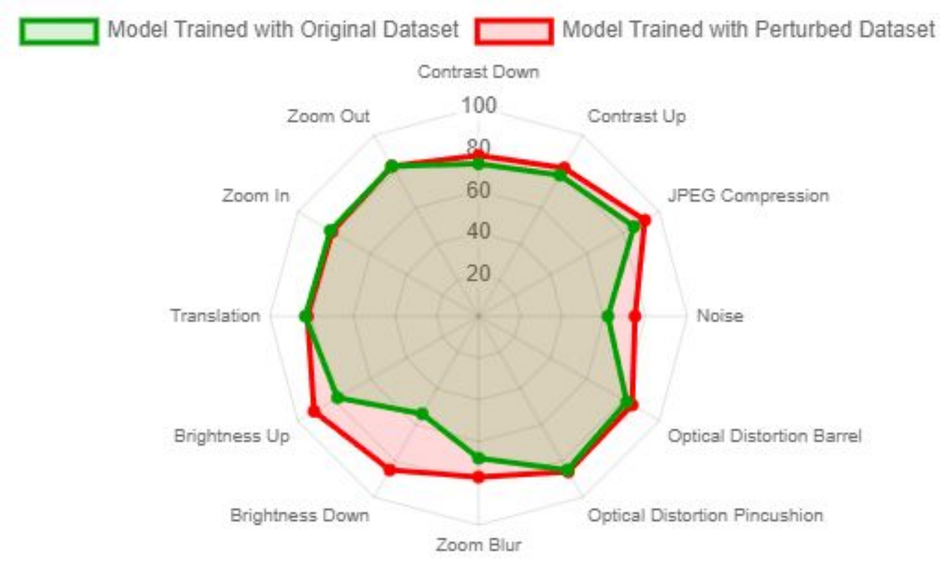
- Comprehensive and automated AI quality tests such as:
  - Noise robustness testing
  - Untrained/biased area detection
  - Dataset label error estimation
  - Fairness evaluation
  - Visualization of explainability
- Reports to assess compliance with legal regulations and international standards related to AI
- Supports testing a wide variety of AI model and dataset formats

Results

Experiment 1

- We discovered that the robustness score is the lowest in the **Brightness Down** perturbation, indicating that the fundus image model's prediction quality is strongly influenced by low brightness.

Experiment 2



- We augmented the training data of fundus images through a process of randomly changing the brightness level of each image.
- As a result, for most including Brightness Down, the model trained with the perturbed training dataset showed a better result.

Discussion

Zoom Blur

Brightness	[0.01, 0.12]	(0.12, 0.15]	(0.15, 0.22]	(0.22, 0.29]	(0.29, 0.7]
With disturbance	62%	65%	71%	73%	72%
Without disturbance	**79%**	**81%**	**75%**	**77%**	**75%**

Gaussian Noise

Brightness	[0.01, 0.12]	(0.12, 0.15]	(0.15, 0.22]	(0.22, 0.29]	(0.29, 0.7]
With disturbance	39%	46%	57%	82%	**85%**
Without disturbance	**66%**	**68%**	**71%**	**85%**	84%

- We focus on Zoom Blur and Gaussian Noise here, which showed improved performance. Using Citadel Lens, we divided the dataset evenly based on brightness, and observed the changes in Robustness Scores for each subset.
- Notably, there was a significant improvement in low brightness images, suggesting that learning improved for low brightness images, and also resulted in an increased robustness to other types of perturbations as a secondary effect.

Conclusion

- We quantitatively confirmed that the current model is susceptible to changes in brightness of input images.
- Subsequently, by introducing brightness-related fundus image perturbations during training, we verified that the model gained robustness to variation in brightness.
- This approach is adaptable to other kinds of perturbations to fundus images.

[1] Abràmoff, M.D., Lavin, P.T., Birch, M. *et al.* Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Med* **1**, 39 (2018).