Accurate and Explainable Cataract Detection Using Eye Images Taken by Hand-held Slit-lamp Cameras

Daniel Kai Xiang Fung1*, Di Wang2*, Hao Wang2, Yongwei Wang2, Pengcheng Wu2, Yan Yee Hah3, Chee Chew Yip3, Wee Jin Heng4, Tock Han Lim4, Cyril Leung2, Chunyan Miao1,2**
1School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore
2Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, NTU, Singapore
3Khoo Teck Puat Hospital (KTPH), Singapore 4Tan Tock Seng Hospital (TTSH), Singapore
*equal contribution  **corresponding author (ascymiao@ntu.edu.sg)

Abstract—Cataract is a leading cause of visual impairment in the elderly. With a greying population globally, there is a pressing need to improve the accessibility of cataract screening. Hand-held Slit-lamp Cameras (HSCs) are often preferred in community eye screening due to their greater portability and accessibility. However, the image quality from HSCs is generally inferior to that from conventional bulky fundus cameras. In this paper, we extend the pre-trained ResNet-18 neural network to analyze a limited number of HSC images (n=187) for cataract detection. Model accuracy is improved through augmenting training data samples and complementing the visual features with patients’ vision measurements. Explainability (of focal model areas for decision making) is attained via extracting the saliency maps using the Grad-CAM method. Our model achieves a high accuracy of 0.96, on par with state-of-the-art results reported in the literature. Our approach demonstrates the potential of large-scale community-based cataract detection using HSCs and our highly accurate and explainable AI-assisted model.

Index Terms—cataract detection, eye images, hand-held slit-lamp cameras, AI-assisted community health screening

1. INTRODUCTION

Cataract is the clouding of the human crystalline lens. It is the most prevalent and treatable cause of visual impairment and blindness around the globe [1]. Cataract is commonly caused by aging, disease, trauma, medication, and genetic predisposition [2]. Thus, it poses a great challenge to all aging populations including Singapore [3]. Furthermore, because the cataract progresses slowly and painlessly, patients are often unaware of its presence, e.g., most Singaporeans above 40 who have significant cataracts in either eye are unaware of their condition [4]. Generally, the patient is referred to an ophthalmologist when the cataract has progressed significantly. This delay in cataract diagnosis may result in diminished visual quality. Hence, accessible large-scale cataract screening is essential to detect a cataract in its early stage to stop or reverse the progression of visual impairment.

In a typical clinical setting, an ophthalmologist often assesses the patients’ optic nerve and disk using the fundus camera. When the ophthalmologist further checks the existence of cataract, he/she also evaluates the patients’ visual acuity (e.g., Logarithm of the Minimum Angle of Resolution (LogMAR) Chart) and contrast sensitivity (e.g., Pelli-Robson Contrast Sensitivity Chart) using eye charts [2]. Expertise is needed to operate the fundus camera and to compare the patients’ crystal images with the standard graded images such as the Lens Opacity Classification System III (LOCS III) [7]. Fundus cameras are expensive, bulky and requires expertise to operate (see Fig. 1(a)). A few prior studies were carried out to use Hand-held Slit-lamp Cameras (HSCs) as an alternative (see Fig. 1(b)) to capture images of the patients’ eyes and subsequently apply various machine learning algorithms to evaluate cataract (see Section II). The use of HSC enables more healthcare providers to perform cataract screening in community settings. However, HSC image quality is generally much inferior to those taken by fundus cameras (see Fig. 2).

Thus, the accuracy of cataract diagnosis using HSC images is generally lower than using fundus cameras (see Table IV).

To elevate the accuracy of cataract evaluation using HSC images towards large-scale screening in community settings, in this paper, we incorporate two additional procedures into the general framework, namely (i) data augmentation to alleviate the problem of having a small dataset and an imbalanced class distribution, and (ii) fusion of patients’ vision measurements, i.e., visual acuity and contrast sensitivity, as the peripheral information with the deep learning computer vision model.

Fig. 1. Two types of cameras often used for cataract evaluation. (a) Fundus cameras (picture excerpted from [8]) are often used in clinical settings. (b) Hand-held slit-lamp cameras (picture excerpted from [9]) can be used in non-clinical settings due to their greater portability and accessibility.
We conduct experiments using own-collected dataset comprising 187 eye images (quite small in size comparing to prior studies). The results show that our model achieves an accuracy of 0.96, which is on par with state-of-the-art results reported in the literature. To further validate the model, we extract saliency maps to visualize the areas of attention where the model focuses on while deriving the prediction. The assessment on these saliency maps provide much transparency to the model’s dynamics, which is essential to encourage healthcare providers to adopt our AI-assisted cataract detection model for large-scale screening in community settings.

II. RELATED WORK

Because the images taken by fundus cameras are of high quality (see Figs. 2(a) and 2(b)), many pioneer AI-assisted cataract diagnosis models were developed using fundus images. Especially after the release of the Ocular Disease Intelligent Recognition (ODIR) dataset in 2019 [10], which comprises 5,000 patients’ information and their fundus images of both eyes (other than cataract, disease labels such as diabetes, glaucoma, hypertension, etc., are also provided), a series of studies were conducted using ODIR. Among which, Sudarsono et al. [12] fine-tuned AlexNet with the diffGrad optimizer to achieve an accuracy of 0.975 for cataract detection. Hasan et al. [13] used the pre-trained InceptionResNetV2 to achieve an even higher accuracy of 0.982. These studies showcased the high efficacy of the AI-assisted cataract diagnosis models trained using fundus images.

Nonetheless, as afore-introduced, it is impractical to use bulky fundus cameras for cataract detection in non-clinical settings. Therefore, HSCs were also often used in cataract diagnosis studies. Liu et al. [14] constructed their own CNN model to perform cataract detection and grading. It is worth noting that they focused on pediatric cataracts while all the other studies mentioned in this paper focus on cataracts in adults. Their model achieved an accuracy of 0.971 for cataract detection. Xu et al. [15] first applied Faster R-CNN to locate the nuclear region and subsequently applied ResNet-101 for cataract grading. Their model achieved an accuracy of 0.847 for 5-class cataract grading.

Instead of using commercial-off-the-shelf (COTS) HSCs, a group of researchers attempted to make the hand-held devices more portable and accessible by developing prototypes with COTS smartphones integrated. Zhang et al. [16] used Inception-v3 for cataract detection and achieved an accuracy of 0.948. It is worth noting that they also collected a dataset using COTS HSCs and achieved a higher accuracy of 0.959. Hu et al. [17] employed a combination of YOLOv3, ShuffleNet and SVM to grade cataracts and achieved an accuracy of 0.935 for three classes. Askarian et al. [18] extracted luminance and RGB features to train an SVM, achieving an accuracy of 0.966 for cataract detection. Different from all the other studies mentioned in this paper, Askarian et al. [18] used images of eye models rather than human eyes.

As aforementioned, due to the inferior quality of HSC images, models trained using these images generally achieve a lower accuracy than those trained using fundus images (see Table IV). Nonetheless, we still focus on HSC images in this paper because they represent the real-world quality of eye images acquired in most non-clinical settings.

Surprisingly, these prior studies on cataract diagnosis did not apply any explainable AI (XAI) techniques to visualize or explain the model dynamics. We deem such explanations as critically necessary to provide ‘assurance’ to the medical professionals. Hence, we analyse the explainability of our model (see Section IV-D).

III. METHODOLOGY

Same as a number of prior studies [12]–[14], [16], [18], the key objective of this work is to develop an AI-assisted cataract detection model that is capable of determining whether a given eye image has cataract or not. Nonetheless, comparing to prior studies, our model has the following three innovative aspects, all novel in the field of image-based cataract detection:

1) Augmentation of limited data samples (from our infrared-based HSCs). This alleviates the problems of small data size and imbalanced data distribution in the dataset.

2) Feature complement with patient’s vision measurements, namely visual acuity measured by the LogMar Chart [5] and contrast sensitivity measured by the Pelli-Robson Contrast Sensitivity Chart [6], respectively.

3) Visualisation of the model’s decision process by extracting the saliency maps using the Grad-CAM method [19].

We deem these three innovative aspects should be adopted by subsequent studies in this field. Specifically, data augmentation should be applied because this process would save cost in data collection; vision measurements including visual acuity and contrast sensitivity should be used because they are always collected during a standard cataract screening process;
and the model’s dynamics should be visualized to assess the
rightfulness of the decisions made. In summary, adopting these
three procedures will not lead to any increase in the amount
of effort required for data collection, while at the same time,
they would help to elevate the model performance and
provide the necessary transparency of the model.

The architecture of our cataract detection model is shown
in Fig. 3. As shown, there are five key procedures in our
model, namely (a) image preprocessing and data augmen-
tation, (b) parameter adoption of ResNet-18 pre-trained on
ImageNet, (c) complements with patient’s vision measure-
ments, (d) model fine-tuning, and (e) visualization of the
decision process. We introduce these five key procedures in
the following subsections, respectively.

A. Image Preprocessing and Data Augmentation

We preprocess all the available HSC images as follows:
1) The white space on the image border, if any, is cropped.
2) The contrast of each cropped image is enhanced by ap-
plying the Contrast-Limited Adaptive Histogram Equalization
(CLAHE) [20] technique.
3) The RGB pixel values for all images are normalized
across the dataset towards zero-mean. This step is usually
applied to improve the deep learning model’s convergence.

Thereafter, we generate synthetic samples to deal with the
following two common problems exist in this research field:

**Problem 1:** The sample size of the collected images is
generally small, which may lead to model underfitting.

**Problem 2:** The number of healthy eye samples (without
cataract) is generally much smaller than that of cataract
samples, because it is exceptionally uncommon for people
with healthy vision have their eye(s) imaged. Nonetheless,
this imbalanced class distribution may pose challenges to the
classification model.

Specifically, we apply random scaling, random cropping,
random rotation, and random brightness scaling to generate
synthetic samples based on the original ones. In addition, we
purposely apply another round of random rotation (ranging in
[-180°, +180°]) to all images in the training dataset. This is
to let the model learn that the visual artifacts (glares and their
reflections) are irrelevant to the presence of cataracts. In the
end, all images are center cropped to the size of 224×224
pixels to fit with the input size of ResNet-18 [11].

B. Adoption of Pre-trained ResNet-18

Because the size of the HSC image datasets is normally
small, we choose not to train a deep learning model from
scratch. In this work, we select ResNet-18 [11] (see Fig. 4)
pre-trained on the ImageNet dataset. As aforementioned, this
selection is based on preliminary results that ResNet-18
achieves competitive performance comparing to larger models
in the ResNet family. In addition, we deem the small model
size will lead to a faster processing speed in future real-world
deployments for cataract screening.

C. Complements with Patient’s Vision Measurements

Patient’s visual acuity and contrast sensitivity are always
measured as part of the cataract detection and grading pro-
cedures. Such measurements can be easily and are often
conducted in non-clinical settings, such as in the spectacles
shops. In our work, we complement the HSC images with
such vision measurements as the model’s inputs, aiming to
elevate the model performance by utilizing the highly relevant,
existing information.

Specifically, we extend the Fully Connected (FC) layer in
the original ResNet-18 (see Fig. 4) with two FC layers (see
Fig. 3). The inputs and outputs of the two FC layers are
described in Table I. Specifically, the patients’ visual acuity
Fig. 4. The architecture of ResNet-18. As shown, multiple convolutional layers are grouped into convolutional blocks: Conv_1, Conv_2, Conv_3, and Conv_4.

TABLE I
THE TWO FC LAYERS IN OUR MODEL

<table>
<thead>
<tr>
<th>Layer</th>
<th>Input(s)</th>
<th>Shape</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC_1</td>
<td>Latent visual features from ResNet-18</td>
<td>1×128</td>
<td>1×128</td>
</tr>
<tr>
<td></td>
<td>Patient’s visual acuity score</td>
<td>1×128</td>
<td>1×128</td>
</tr>
<tr>
<td>FC_2</td>
<td>Output of FC_1</td>
<td>1×128</td>
<td>1×2</td>
</tr>
</tbody>
</table>

(measured by LogMAR [5]) and contrast sensitivity (measured by Pelli-Robson Contrast Sensitivity [6]) are mapped to vectors with the size of 1×128, respectively. Such mapping is learned (with batch normalization) by the model (see the following subsection).

It is worth mentioning that for the augmented images, the associated vision measurements are adopted from the corresponding original image without any random value alteration.

D. Model Fine-tuning

Because the key objective of our model is to determine whether cataract exists in the input HSC image with the complement of patient’s vision measurements, the model essentially performs binary classifications that 0 represents the absence of cataract and 1 represents presence. Therefore, we use the following cross-entropy loss ($J(\theta)$) to train all learnable parameters:

$$J(\theta) = -\frac{1}{N} \sum y \log(g(z)) + (1 - y) \log(1 - g(z)),$$

(1)

where $\theta$ denotes the set of all learnable parameters, $N$ denotes the total number of data samples used for training, $y$ denotes the ground-truth label, $z$ denotes the model inputs, and $g(z)$ denotes the model output. To speed up the learning process, we optimize $\theta$ using the stochastic gradient descent algorithm.

E. Saliency Maps for Model Dynamics Visualization

To provide transparency on how our model detects the presence of cataract in an input image, we apply the Grad-CAM algorithm [19]. Specifically, for each input HSC image, the gradients flowing into the final convolutional layer of each convolutional block (see Fig. 4) are used to produce a coarse saliency map to visualize and examine the attention of the model (see Section IV-D).

IV. EXPERIMENTS

In this section, we first introduce the own-collected dataset. We then describe how we define the various model configurations and set up the experiments. After presenting and discussing the experimental results, we visualize the saliency maps extracted from both healthy and cataract images.

A. Own-collected HSC Image Dataset

The study data were collected from Khoo Teck Puat Hospital (KTPH) and Tan Tock Seng Hospital (TTSH) in Singapore over the period of Q2 2022 to Q2 2023. After removing samples with missing patient data, a total number of 187 samples comprising both the HSC images and patient vision measurements were collated. This dataset is highly imbalanced where only 19 samples (10.16%) are healthy, while the other 168 samples all have cataracts in varying grades.

B. Model Configurations and Experimental Set-ups

Because the key objective of this study is to develop an AI-assisted cataract detection model for future deployment in community settings, we focus on the binary classification problem to determine whether a sample has the presence of cataract or not. Nonetheless, to alleviate the problem of having a highly imbalanced class distribution, we generate two augmented datasets following the steps introduced in Section III-A. As shown in Table II, the Baseline dataset only comprises the original 187 samples. For the Balanced-small dataset, we only augment the healthy class to have a larger dataset. It is worth mentioning that the augmented samples are generated independently for the latter two datasets.

TABLE II
DATASETS USED IN OUR EXPERIMENTS

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image Type</th>
<th>Train</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Healthy</td>
<td>14</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Cataract</td>
<td>163</td>
<td>5</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>177</td>
<td>10</td>
<td>187</td>
</tr>
<tr>
<td>Balanced-small</td>
<td>Healthy</td>
<td>168</td>
<td>5</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Cataract</td>
<td>163</td>
<td>5</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>331</td>
<td>10</td>
<td>341</td>
</tr>
<tr>
<td>Balanced-large</td>
<td>Healthy</td>
<td>499</td>
<td>5</td>
<td>505</td>
</tr>
<tr>
<td></td>
<td>Cataract</td>
<td>489</td>
<td>5</td>
<td>494</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>988</td>
<td>10</td>
<td>998</td>
</tr>
</tbody>
</table>
### Table III: Performance of our cataract detection model

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Dataset</th>
<th>Accuracy</th>
<th>F1-score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>image-only</td>
<td>Baseline</td>
<td>0.740 ± 0.038</td>
<td>0.795 ± 0.041</td>
<td>0.960 ± 0.046</td>
<td>0.520 ± 0.122</td>
</tr>
<tr>
<td></td>
<td>Balanced-small</td>
<td>0.890 ± 0.035</td>
<td>0.892 ± 0.032</td>
<td>0.900 ± 0.053</td>
<td>0.880 ± 0.077</td>
</tr>
<tr>
<td>image+metadata</td>
<td>Baseline</td>
<td>0.845 ± 0.048</td>
<td>0.863 ± 0.040</td>
<td>0.950 ± 0.056</td>
<td>0.740 ± 0.103</td>
</tr>
<tr>
<td></td>
<td>Balanced-small</td>
<td>0.905 ± 0.039</td>
<td>0.908 ± 0.037</td>
<td>0.900 ± 0.053</td>
<td>0.910 ± 0.072</td>
</tr>
<tr>
<td></td>
<td>Balanced-large</td>
<td>0.960 ± 0.039</td>
<td>0.959 ± 0.041</td>
<td>0.960 ± 0.046</td>
<td>0.960 ± 0.036</td>
</tr>
</tbody>
</table>

### Table IV: Performance comparisons with prior studies (models ranked by accuracy)

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Image type</th>
<th>Sample size</th>
<th>Model</th>
<th>Accuracy</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basai et al. [13]</td>
<td>2021</td>
<td>Fundus</td>
<td>1,088</td>
<td>Inception+ResNetV2</td>
<td>0.982</td>
<td>Using images selected from ODIR [10]</td>
</tr>
<tr>
<td>Sudarsono et al. [12]</td>
<td>2020</td>
<td>Fundus</td>
<td>200</td>
<td>AlexNet</td>
<td>0.975</td>
<td>Using images selected from ODIR [10]</td>
</tr>
<tr>
<td>Liu et al. [14]</td>
<td>2017</td>
<td>HSC</td>
<td>886</td>
<td>CNN</td>
<td>0.971</td>
<td>Focused on pediatric cataract</td>
</tr>
<tr>
<td>Askarian et al. [18]</td>
<td>2021</td>
<td>HSC</td>
<td>100</td>
<td>SVM</td>
<td>0.966</td>
<td>Using own HSC prototype and eye models, using extracted luminance and RGB features</td>
</tr>
<tr>
<td>Ours</td>
<td>2024</td>
<td>HSC</td>
<td>187</td>
<td>ResNet-18</td>
<td>0.990</td>
<td>Image augmentation + vision measurements</td>
</tr>
<tr>
<td>Zhang et al. [16]</td>
<td>2020</td>
<td>HSC</td>
<td>2,516</td>
<td>Inception+3</td>
<td>0.948</td>
<td>Using own HSC prototype</td>
</tr>
</tbody>
</table>

The size of the Baseline dataset is small, especially because it only has 19 healthy samples. Therefore, instead of splitting the samples with a consistent ratio into the train and test subsets, we maintain a consistent size of the test subset. Specifically, for all experiments, we always keep five samples from each class in the test subset while use all the other samples for training. It is worth highlighting that the samples in the test subset are all the original ones without augmentation and for all samples selected in the test subset, their correspondingly augmented samples, if any, are excluded from the train subset. These precautions are implemented to ensure there is no data leakage between the train and test subsets.

To remove the random effects, for each experiment setting, we report the averaged performance with standard deviations (std) across 20 independent runs.

In addition, to evaluate the effectiveness of having the complements of patient’s vision measurements, we conduct an ablation study to examine the performance of our model without taking in the vision measurements as inputs. We refer to the former configuration as image-only and the latter as image+metadata. Specifically, for the image-only configuration, instead of the two FC layers (see Table I), we adopt the original single FC layer of ResNet-18 (see Fig. 4).

### C. Experimental Results

The experimental results are presented in Table III. It is not surprising to see that the image+metadata model trained on the Balanced-large dataset achieves the best performance across all performance metrics. Moreover, the image+metadata model is shown as outperforming the image-only model, showcasing the effectiveness and importance of the complement of patient’s vision measurements. Furthermore, it is observed that with the increase of the dataset size, the performance of both models generally improves, showcasing the effectiveness and importance of image augmentation. Specifically, with the augmentation of more and more healthy samples, both models’ specificity keeps improving.

Because all HSC image datasets used and the models constructed in prior studies were not made publicly available, it is not viable to make fair comparisons using the same dataset. Alternatively, we compare the performance of our model against the prior studies’ results on cataract detection (binary classification on the presence of cataract) simply using their reported model performance while listing the contextual information in Table IV. Although our model comes in second last in terms of accuracy among the six compared in Table IV, its performance is still competitive especially among the studies using HSC images. The top-two models [13], [12] used high-quality fundus images (see Fig. 2) that let them outperform the rest. The third model [14] used a larger dataset with a balanced class distribution (476 healthy and 410 cataract). The fourth model [18] used eye models to take high-quality images with an equal class distribution. As such, considering all these factors that greatly affect the model performance, we deem ours achieves the same level of high performance as these state-of-the-art prior studies.

### D. Using Saliency Maps to Visualize Attention

Using the best-performing model trained on the Balanced-large dataset, we extract saliency maps of all the original images for visualization of the model’s attention (see Section IV-A). In Fig. 5, we present the saliency maps extracted from two randomly selected images: one with cataract (patient ID: A031, right eye) and one healthy sample (patient ID: K009, right eye). These two sets of saliency maps well demonstrate the model’s attention at different depth. Specifically, at the relatively shallower layers (i.e., Conv_1 and Conv_2, see Fig. 4), the model pays more attention to the specific and scattered regions, e.g., the small regions brightened by the HSC and around the peripheral; while at the deeper layers...
Fig. 5. Saliency maps extracted from two randomly selected images. The red regions indicate where the model’s attention lies the most, the black regions indicate where the model’s attention lies the least, while the other colors interpolate the level of attention in between.

(i.e., Conv_3 and Conv_4), the model pays more attention to the abstract and connected regions, e.g., the large regions around the centerline of the iris. Such observations are consistent across most saliency maps that cataracts are generally identified at either Conv_3 or Conv_4 (for A031, the most noticeable part of the cataract is accurately identified in the bottom-right quadrant at Conv_3), while for healthy eyes, the attention is generally spread out at Conv_3 and Conv_4 (see Conv_4 of K009).

Furthermore, to provide a visualized overview of all the extracted saliency maps, we average them at each convolutional block according to the images’ class labels and present the averaged results in Fig. 6. As shown, there is little difference between the three sets of averaged saliency maps, suggesting our model has a consistent focus. The HSC glare and the outline of the iris are clearly visible at Conv_1, while they get blurred at Conv_2 and Conv_3. The area of attention at Conv_4 is enlarged and centers around the centerline of the iris with no specific focus.

Observations from both Figs. 5 and 6 suggest that our cataract detection model successfully emulates how an ophthalmologist checks the various regions of an eye step-by-step to arrive at the final decision.

V. CONCLUSION

This study shows that our AI-assisted cataract detection model using HSC images has high accuracy and explainability (for transparency of the decision process). It may potentially be deployed for large-scale cataract detection in communities.

Going forward, we plan to 1) develop own hand-held cameras to capture high-quality images and 2) automate the screening procedures towards real-world deployments.

REFERENCES


