


RESEARCH

Open Access



# Retinal photograph-based deep learning system for detection of hyperthyroidism: a multicenter, diagnostic study

Li Dong<sup>1†</sup>, Lie Ju<sup>2†</sup>, Shiqi Hui<sup>1†</sup>, Lihua Luo<sup>3†</sup>, Xue Jiang<sup>1</sup>, Zihan Nie<sup>1</sup>, Ruiheng Zhang<sup>1</sup>, Wenda Zhou<sup>1</sup>, Heyan Li<sup>1</sup>, Jost B. Jonas<sup>4,5,6</sup>, Xin Wang<sup>2</sup>, Xin Zhao<sup>2</sup>, Chao He<sup>2</sup>, Yuzhong Chen<sup>2</sup>, Zhaohui Wang<sup>7</sup>, Jianxiong Gao<sup>7</sup>, Zongyuan Ge<sup>8,9</sup>, Wenbin Wei<sup>1</sup> and Dongmei Li<sup>1\*</sup> 

<sup>†</sup>Li Dong, Lie Ju, Shiqi Hui and Lihua Luo contributed equally to this study.

\*Correspondence: ldmlily@x263.net

<sup>1</sup> Beijing Tongren Eye Center, Beijing Key Laboratory of Intraocular Tumor Diagnosis and Treatment, Beijing Ophthalmology & Visual Sciences Key Lab, Medical Artificial Intelligence Research and Verification Key Laboratory of the Ministry of Industry and Information Technology, Beijing Tongren Hospital, Capital Medical University, 1 Dong Jiao Min Lane, Beijing 100730, China

<sup>2</sup> Beijing Airdoc Technology Co., Ltd, Beijing, China

<sup>3</sup> Department of Ophthalmology, Beijing Friendship Hospital, Capital Medical University, Beijing, China

<sup>4</sup> Department of Ophthalmology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

<sup>5</sup> Institute of Molecular and Clinical Ophthalmology Basel, IOB, Basel, Switzerland

<sup>6</sup> Privatpraxis Prof Jonas Und Dr Panda-Jonas, Heidelberg, Germany

<sup>7</sup> iKang Guobin Healthcare Group Co., Ltd, Beijing, China

<sup>8</sup> Faculty of Engineering, Monash University, Melbourne, VIC, Australia

<sup>9</sup> Faculty of Engineering, ECSE, Monash University, Melbourne, VIC, Australia

## Abstract

**Background:** Screening for hyperthyroidism using gold-standard diagnostic criteria in the general population is not cost-effective, leading to a relatively high rate of undiagnosed and untreated patients. This study aimed to establish a deep learning-based system to detect hyperthyroidism based on retinal photographs.

**Methods:** The multicenter, observational study included retinal photographs taken from participants in two hospitals and 24 health care centers throughout China. We first trained two models to identify hyperthyroidism: in model #1, the non-hyperthyroidism individuals were randomly selected, while in model #2, the non-hyperthyroidism group was matched for age and gender with the hyperthyroidism group. After internal validation, we selected the better model for further evaluation using external validation datasets.

**Results:** The study included 22,940 retinal photographs of 11,409 participants for the model development, and 3862 retinal photographs (1870 participants) which were obtained from two hospitals and four medical centers as the external validation datasets. Model #1 achieved a higher area under the receiver operator curve (AUC) than model #2 (0.907, 95% CI: 0.894–0.918 versus 0.850, 95% CI: 0.832–0.866) in the internal validation so that model #1 was used for further evaluation. In external datasets, model #1 reached AUCs ranging from 0.816 (95% CI 0.789–0.846) to 0.849 (95% CI 0.824–0.874) and achieved accuracies between 0.735 (95% CI 0.700–0.773) and 0.796 (95% CI 0.765–0.824). Heatmaps showed a focus of the DL-algorithm on large fundus vessels and the optic nerve head.

**Conclusions:** Retinal fundus photographs may serve for DL systems for a cost-effective and non-invasive method to detect hyperthyroidism.

**Keywords:** Artificial intelligence, Deep learning, Hyperthyroidism, Thyrotoxicosis, Retinal photographs, Retina

## Introduction

Thyroid hormones are essential for body growth, neuronal development, and regulation of metabolism [1] Hyperthyroidism is a globally common thyroid dysfunction with potentially devastating health consequences. Hyperthyroidism is a form of thyrotoxicosis due to inappropriately high synthesis and secretion of thyroid hormones [2] In the United States, the prevalence of hyperthyroidism is about 1.2%, with approximately 40% of the patients with clinically apparent symptoms and 60% of the patients having sub-clinical signs [3] Screening for hyperthyroidism in the general population is not a cost-effective and feasible procedure, because of the relatively low prevalence of the disease and since the examination. Only those who are at high risk due to comorbid conditions, family history, or medication use, are recommended to receive physical examinations and laboratory tests [4] Therefore, hyperthyroidism is frequently unrecognized and untreated. It has been estimated that the prevalence of undiagnosed hyperthyroidism is about 1.72% of the general population Europe [5] It may lead to adverse outcomes and increased costs [6] Improved systems for the detection of hyperthyroidism are thus needed.

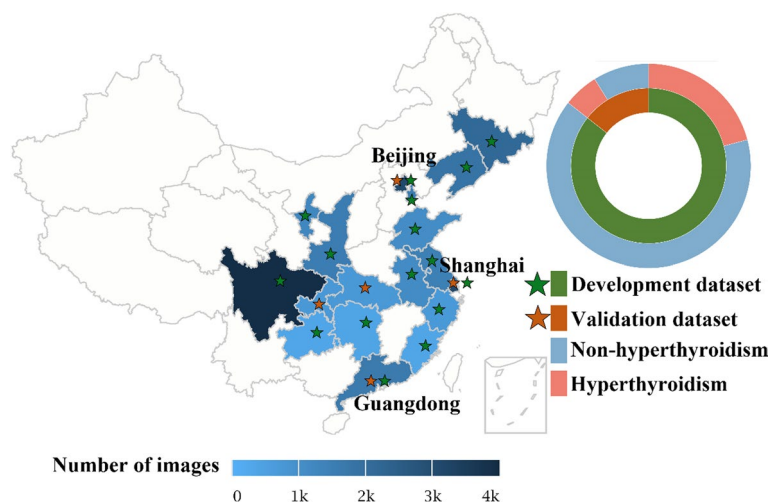
Deep learning (DL) is a state-of-the-art technique that allows computational models to learn representations of data with multiple levels of abstraction [7] In the field of medicine and healthcare, DL has been primarily applied to analyze medical images for automatic measurement, augmentation, classification, diagnosis, and even prediction. [8–11] There is an increasing interest in establishing DL-based low-cost and non-invasive methods trained from color fundus photographs to predict demographic parameters and systematic disorders, including age, gender, cardiovascular risk factors [8], anemia [12] and hepatobiliary diseases [13] Rim and colleagues broadened the applicability of retinal photograph-based DL to predict 47 systemic biomarkers, however, thyroid function has not been well predicted [14] The present study was therefore conducted to develop a DL-based model for the detection of hyperthyroidism based on retinal photographs.

## Methods

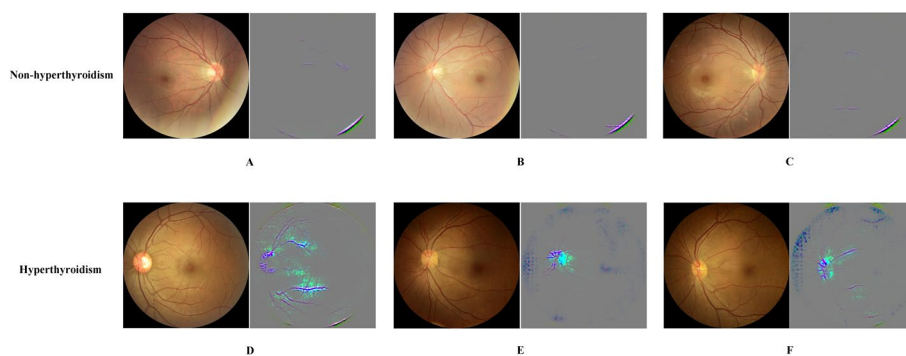
### Study design and participants

The multicenter observational study was conducted in two general hospitals and 24 medical examination centers in 19 cities in China, and registered at ClinicalTrials.gov (NCT04678375). All medical examination centers as independently operating medical units belonged to the same healthcare group under the supervision of the Tongren Hospital. The geographical distribution of all hospitals and medical examination centers included in this study were presented in Fig. 1. The data was retrospectively collected from June 1st, 2018 to July 31st, 2020.

The data obtained from 20 medical examination centers were used for the development of the DL system (Fig. 2). We first randomly extracted data of normal individuals as the control group and data of patients with hyperthyroidism in a ratio of 2:1. To eliminate potentially confounding variables, we established an additional control group, matching for age and gender with the study group. We thus had two different control groups and one study group. Within the control groups and within the study group, the data was then randomly divided into a development dataset and an internal validation



**Fig. 1** Distribution of included participants. Pie charts show the constitution of hyperthyroidism and non-hyperthyroidism images among development and validation dataset



**Fig. 2** Heatmap Visualization of fundus images of non-hyperthyroidism A–C and hyperthyroidism D–F

dataset with a ratio of 8:2. It should be noted that images from the same patient were only included in the development dataset or validation dataset. Each development dataset was again randomly divided into a training dataset and a tuning dataset (7:1 ratio). Therefore, there were two DL algorithms trained from two development datasets and then internally tested in two internal validation datasets, respectively. Only the model with the better performance in the internal validation was chosen to be tested in external validation datasets.

Data from the Beijing Tongren Hospital (BTH, Beijing, China), Beijing Friendship Hospital (BFH, Beijing, China), Chongqing Zhuoyue Medical Centre of iKang Healthcare Group (CZMC, Chongqing, China), Shanghai Yuanhua Medical Centre of iKang Guobin Healthcare Group (SYMC, Shanghai, China), Shenzhen Zhuoyue Medical Centre of iKang Healthcare Group (SZMC, Shenzhen, China), and Wuhan Jindun Road Medical Centre of iKang Healthcare Group (WJMC, Wuhan, China) were used as external validation datasets to further evaluate the performance of the system.

Inclusion criteria were the availability of a complete set of clinical data, basic demographic characteristics, medical history, and thyroid function test reports.

Hyperthyroidism was defined by a serum concentration of thyroid-stimulating hormone (TSH) level was lower and serum thyroxine ( $T_4$ ) level, triiodothyronine ( $T_3$ ) level, or both were above the reference range, or by a medical history of hyperthyroidism diagnosed by endocrinologists [2] Participants in control group had normal serum concentration of TSH,  $T_4$ , and  $T_3$  level. After receiving blood test, all subjects underwent fundus examination within 10 min. The Medical Ethics Committee of the Beijing Tongren Hospital, and the Ethics Committee the iKang Corporation approved the study protocol fulfilling the requirements published in the Helsinki declaration. For patients whose fundus images were stored in the retrospective databases at each participating hospital, informed consent was waived by the institutional review boards.

#### **Data acquisition and quality control**

Performed by trained operator and using various types of non-mydratic 45-degree fundus camera (CR-2AF, Canon, Tokyo, Japan; Nonmyd  $\alpha$ -DIII, Kowa, Tokyo, Japan; TRC-NW300, Topcon, Tokyo, Japan; NT-2000, Nidek, Aichi, Japan), retinal photographs were obtained from one or both eyes of the study participants (Additional file 1: Table S1). The photographs were centered on the mid-point between optic disc and macula. All images were stored in a jpeg format. Quality control was unanimously performed by two trained ophthalmologists to remove poor-quality images resulting from halation, blurs, defocus, and non-retinal images. All images were cropped for the removal of black background with only regions of fundus maintained to uniform the styles of fundus images from different fundus cameras. Both eyes of the same participants were included into the model. To investigate the ability of the DL system for hyperthyroidism detection, those images with obvious ophthalmic diseases were also removed to reduce the training bias and validation bias of the model.

#### **Development of the DL system**

For the development of the DL system and analysis of the relationship between hyperthyroidism and retinal photographs, we leveraged the convolutional neural network (CNN) and several state-of-the-art neural network candidature architectures (VGG-19 [15] ResNet-50 [16] InceptionV3 [17] DenseNet-121 [18] etc.) were tested with the same hyper-parameters settings. We selected the model with the best performance for the detection of hyperthyroidism (Additional file 1: Table S2). Among all these candidature architectures, ResNet-50 showed the best performance. In this study, to increase the generalization of the model, we initialized the model with parameters pre-trained from ImageNet. We kept the same configurations for the further comparison analyses.

Before training the networks, we re-sized all digital images to  $512 \times 512$  pixels and some images in low-quality were manually removed. The pixel values of each image were normalized from (0, 255) to (0, 1) with a linear mapping. Some pre-processing strategies were also used for the data augmentation, such as random flip, rotation and crop. The batch size was set as 16. We used a binary cross-entropy loss (BCE loss) and Adam optimizer for stochastic optimization [19] The learning rate started from  $3e-4$  and dropped by tenfold every 10 epochs. We trained 50 epochs in total (Additional file 1: Fig. S1).

All examinations were implemented with Pytorch 1.8.1 DL toolkit platform [20] the build of all backbones are publicly available in torchvision 0.8.2 site-packages. The

algorithm training was performed on NVIDIA RTX 3090 GPU with CUDA version 9.0 and cuDNN 7.0.

#### Validation of the DL system and visualization heatmaps

Two controlled experimental settings were evaluated for the DL system, with the control group being unmatched, or matched for age and gender, with the study group. By comparing the differences between the two models, the influence of age and gender on the detection of hyperthyroidism was assessed. Finally, the model with the better performance was used to for further evaluation in the external datasets. To visualize the decision ways of the model used, we applied the Grad-CAM to generate heatmaps. [21].

#### Statistical analysis

Statistical analyses were performed using the R software (version 4.0.3). The predictive accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score of the system were evaluated. Bootstrapping with 2,000 replications was used to estimate the 95% confidence intervals (CIs) of the performance metrics [22] We used the receiver operating characteristic (ROC) curve to show the predictive ability of the DL system.

#### Results

A total of 22,940 retinal photographs from 11,409 subjects were included in the study (Table 1). Among them, 5601 images from 2767 subjects were labeled as hyperthyroidism. We used 19,078 retinal photographs from 9,539 participants for the development of the models, and 3862 retinal photographs from 1870 participants were used as the external validation datasets. The basic characteristics of the study populations were presented in Table 1. In the study population for the assessment of model 1, the median age was 46 (ranging from 20 to 88) and 45 (ranging from 15 to 85) in the development dataset and in the internal validation dataset, respectively. In the study population for the assessment of model 2, the median age was 40 (ranging from 15 to 85) and 41 (ranging from 16 to 88) in the development dataset and in the internal validation dataset, respectively.

In the internal validation dataset, model 1 achieved a higher area under receiver operator curve (AUC) than model 2 (0.907 (95% CI 0.894–0.918) versus 0.850 (95% CI 0.832–0.866)), and model 1 had a higher accuracy than model 2 (0.809 (95% CI 0.790–0.826) versus 0.780 (95% CI 0.758–0.801)) (Table 2). Compared to model 2, model 1 also reached a higher sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. Subsequently, model 1 was used for the further external evaluation.

In the external validation datasets, the model reached AUCs ranging between 0.816 (95% CI 0.789–0.846) and 0.849 (95% CI 0.824–0.874) (Fig. 3) and achieved accuracies ranging from 0.735 (95% CI 0.700–0.773) to 0.796 (95% CI 0.765–0.824) (Table 2). Similar values for sensitivity, specificity, PPV, NPV, and F1 score were obtained among all external validation datasets.

In the heatmap visualization, the saliency highlighted the large fundus vessels and optic nerve head areas on the fundus images of patients with hyperthyroidism (Fig. 4). It suggested that the system made the diagnosis based on a fixed pattern.

**Table 1** Characteristics of the study populations

	20 medical centers									
	Development dataset		Internal validation dataset		BTH	BFH	CZMC	SYMJC	SZMC	WJMC
	Model 1	Model 2	Model 1	Model 2	External validation dataset 1	External validation dataset 2	External validation dataset 3	External validation dataset 4	External validation dataset 5	External validation dataset 6
Participants	5080	5080	717	717	227	191	363	363	363	363
Hyperthyroidism	1804	1804	251	251	105	101	116	128	133	129
Non-hyperthyroidism	3276	3276	466	466	122	90	247	235	230	234
Age, years	46 (20–88)	40 (15–85)	45 (19–88)	41 (16–88)	50 (40–60)	43 (34–56)	47 (24–88)	46 (20–73)	47 (24–88)	46 (24–82)
Gender										
Male	2072	1938	344	327	114	73	188	166	168	150
Female	3008	3142	373	390	113	118	178	197	195	213
Body Mass Index, kg/m <sup>2</sup>	22.8 (17.0–28.2)	23.6 (17.2–28.9)	23.3 (16.8–28.9)	23.6 (16.9–28.1)	23.2 (16.8–27.2)	22.5 (16.2–28.0)	23.3 (16.4–28.7)	23.3 (17.2–28.5)	23.9 (15.9–27.8)	23.2 (16.5–28.3)
Systolic blood pressure, mm Hg	130 (88–175)	128 (86–173)	129 (88–173)	128 (83–175)	126 (86–173)	124 (83–175)	130 (87.4–171)	129 (88–179)	130 (86–173)	130 (88–175)
Diastolic blood pressure, mm Hg	78 (49–107)	78 (47–106)	77 (47–105)	78 (49–106)	80 (48–105)	75 (47–106)	78 (50–102)	77 (49–106)	78 (48–106)	79 (48–105)
Retinal photographs	10,160	10,160	1434	1434	592	366	726	726	726	726
Hyperthyroidism	3608	3608	502	502	296	183	232	256	266	258
Non-hyperthyroidism	6552	6552	932	932	296	183	494	470	460	468

Model 1: the control group were randomly selected and unmatched to the hyperthyroidism subjects; Model 2: age and gender of control group were matched to the hyperthyroidism subjects. Hyperthyroidism participants were overlapped in Model 1 and Model 2, whereas non-hyperthyroidism participants were different between Model 1 and Model 2. Data are presented as n, or median values (range) *BFH* Beijing tongren hospital, *BFH* Beijing friendship hospital, *CZMC* Chongqing Zhuoyue medical centre of kang healthcare group, *SYMJC* Shanghai Yuanhua medical centre of kang guobin healthcare group, *SZMC* Shenzhen Zhuoyue Medical Centre of kang Healthcare Group, *WJMC* Wuhan Jindun road medical centre of kang healthcare group

**Table 2** Performance of the DL-based system for the prediction of hyperthyroidism from retinal photographs

	Internal validation dataset		BTH	BFH	CZMC	SYMC	SZMC	WJMC
	Model 1	Model 2	External validation dataset 1 <sup>a</sup>	External validation dataset 2 <sup>a</sup>	External validation dataset 3 <sup>a</sup>	External validation dataset 4 <sup>a</sup>	External validation dataset 5 <sup>a</sup>	External validation dataset 6 <sup>a</sup>
AUC	0.907	0.850	0.816	0.823	0.838	0.822	0.849	0.816
(95% CI)	(0.894–0.918)	(0.832–0.866)	(0.789–0.846)	(0.787–0.858)	(0.810–0.864)	(0.796–0.847)	(0.824–0.874)	(0.791–0.844)
Accuracy	0.809	0.780	0.735	0.735	0.796	0.762	0.789	0.755
(95% CI)	(0.790–0.826)	(0.758–0.801)	(0.706–0.767)	(0.700–0.773)	(0.765–0.824)	(0.729–0.792)	(0.757–0.818)	(0.722–0.785)
Sensitivity (95% CI)	0.761 (0.721–0.796)	0.718 (0.674–0.758)	0.625 (0.567–0.680)	0.617 (0.543–0.687)	0.681 (0.613–0.741)	0.674 (0.609–0.733)	0.738 (0.677–0.792)	0.671 (0.606–0.730)
Specificity (95% CI)	0.885 (0.862–0.904)	0.824 (0.799–0.847)	0.845 (0.797–0.883)	0.852 (0.791–0.899)	0.833 (0.797–0.864)	0.799 (0.761–0.833)	0.814 (0.776–0.847)	0.795 (0.756–0.829)
Positive predictive value (95% CI)	0.791 (0.752–0.825)	0.659 (0.616–0.700)	0.801 (0.742–0.849)	0.807 (0.730–0.867)	0.634 (0.568–0.695)	0.613 (0.550–0.673)	0.658 (0.597–0.714)	0.609 (0.546–0.668)
Negative predictive value (95% CI)	0.866 (0.842–0.887)	0.861 (0.836–0.882)	0.693 (0.642–0.739)	0.690 (0.625–0.749)	0.860 (0.826–0.889)	0.838 (0.801–0.870)	0.865 (0.830–0.894)	0.835 (0.798–0.867)
F1 score	0.837	0.813	0.736	0.735	0.788	0.759	0.789	0.755
	(0.802–0.870)	(0.797–0.831)	(0.706–0.767)	(0.700–0.773)	(0.763–0.813)	(0.731–0.784)	(0.763–0.814)	(0.730–0.781)

Data are presented as estimate (95% Confidence interval)

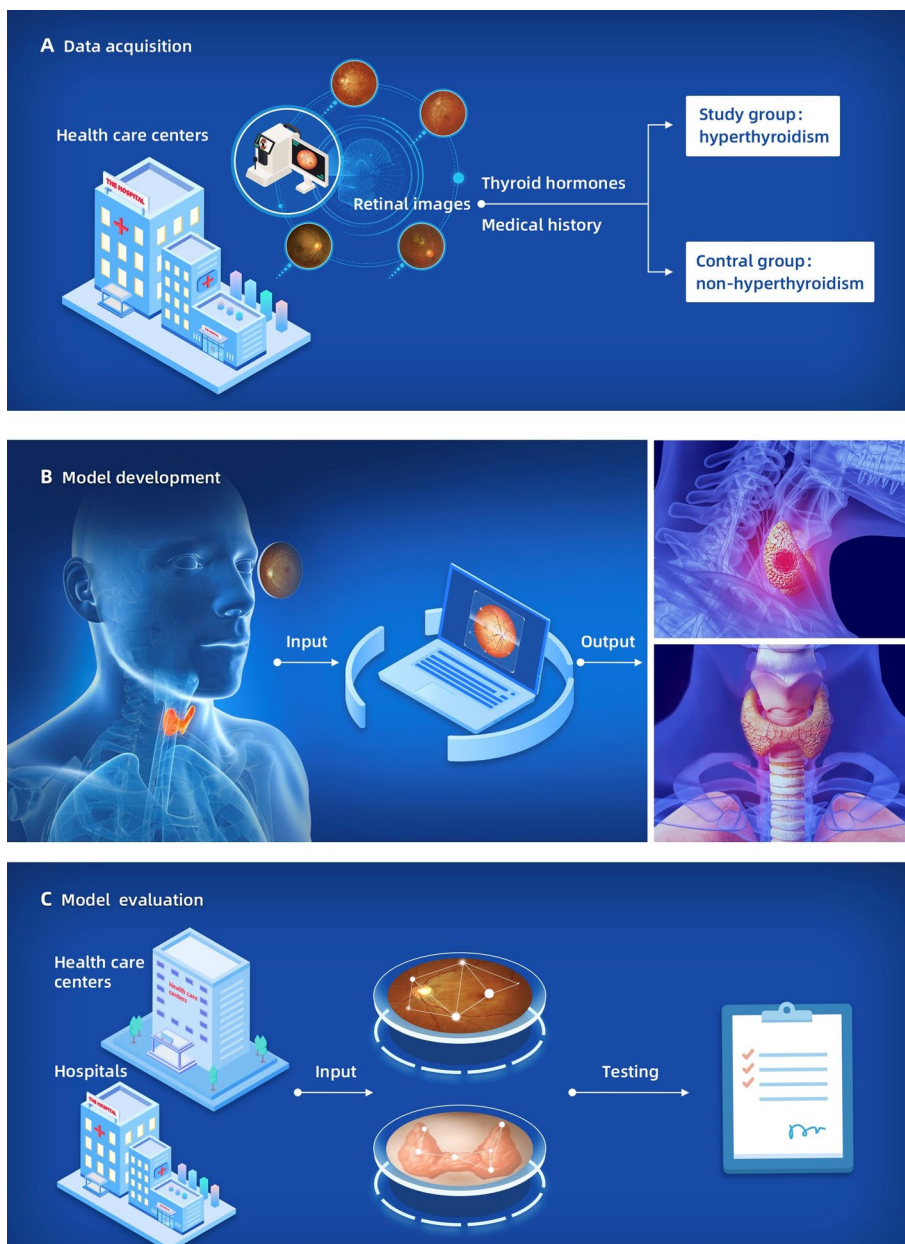
AUC area under the curve, CI confidence interval, BTH Beijing tongren hospital, BFH Beijing friendship hospital, CZMC: chongqing zhuoyue medical centre of ikang healthcare group, SYMC Shanghai Yuanhua Medical Centre of iKang Guobin Healthcare Group, SZMC: Shenzhen Zhuoyue medical centre of ikang healthcare group, WJMC Wuhan Jindun Road Medical Centre of iKang Healthcare Group

<sup>a</sup> Only model 1 was applied in external validation

## Discussion

In this study, more than 20,000 retinal photographs taken using various camera types from 26 centers were used. DL algorithms developed in this study were able to detect hyperthyroidism on retinal photographs with an AUC of ranging between 0.82 and 0.85 and accuracies ranging from 0.74 to 0.80. The key finding shows that a DL-based system can well detect hyperthyroidism from retinal photographs. These results suggest that an automated screening for thyroid diseases may be possible based on routinely taken ocular fundus photographs. The heatmap revealed that the DL algorithm was based on the large retinal blood vessels and on the optic nerve head region.

Previous studies have demonstrated a good performance of artificial intelligence-related and fundus image-based algorithms to differentiate both sexes with an AUC higher than 0.90, and to estimate the age with a coefficient of determination ( $R^2$ ) higher than 0.83 [8, 14]. Considering that age and gender might act as confounding factors for the performance of DL system for the detection of hyperthyroidism [23] we trained and validated two models using randomly selected controls (model 1) and age and gender-matched controls (model 2), respectively. The two models both showed high accuracy

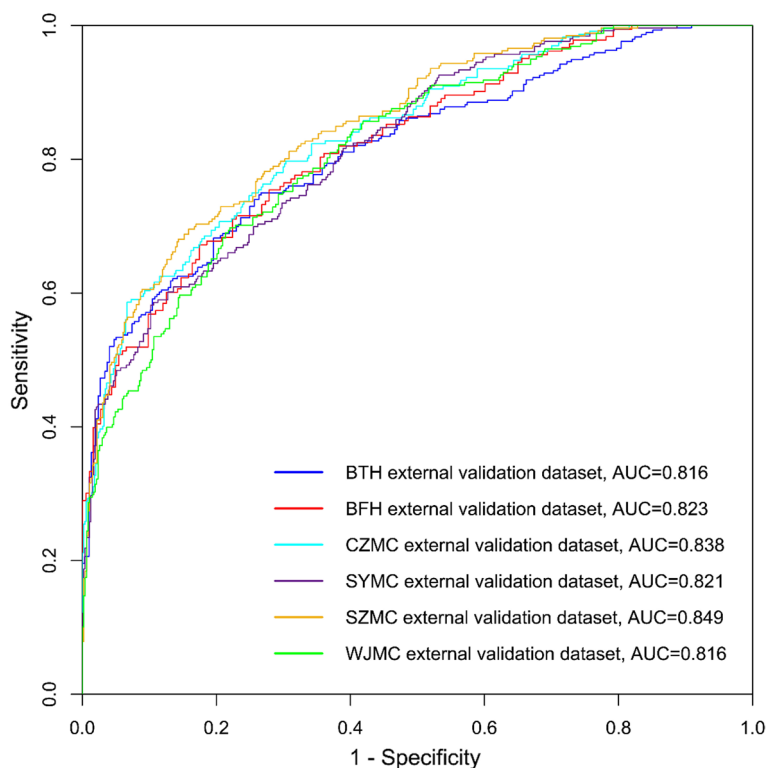


**Fig. 3** Overview of the study process

for the detection of hyperthyroidism, with AUCs higher than 0.85. It suggests that the DL systems could detect hyperthyroidism independently of demographic parameters such as age and sex.

Various efforts have been made to screen and diagnose hyperthyroidism in populations with convenient and low-cost approaches. Sato et al. proposed a novel screening method to assist the diagnosis of Graves' hyperthyroidism applying Bayesian-type and Self-organizing map-type neural networks which used routine test data including serum concentrations of alkaline phosphatase, creatinine, and total cholesterol [24] The model was trained in 120 women (35 patients with Graves' hyperthyroidism and 85 healthy





**Fig. 4** Receiver operating characteristic curves illustrate this algorithm's ability to detect hyperthyroidism in external validation datasets

volunteers) and then tested in 171 female individuals. It achieved a correct diagnosis in 81–94% for the hyperthyroid patients and in 94% to 98% for the healthy individuals. Similar models were formed based on the examination of 78 male individuals and reached AUCs higher than 0.95 for screening hyperthyroidism in 133 subjects [25]. Our study cannot be directly compared to these studies since this is the first investigation to diagnose hyperthyroidism from non-invasive data rather than traditional methods such as biochemical blood examinations and systemic biomarkers.

These findings of the present study illustrate that a fundus photo-based DL system can detect hyperthyroidism. These findings agree with clinical studies showing the involvement of the ocular system in patients with hyperthyroidism. To cite examples, thyroid-associated ophthalmopathy (TAO) occurs mainly in the patients with Graves' disease. Patients with TAO had an increased choroidal thickness [26]. Increased retinal microvascular density was detected in patients with active TAO, while the retinal vessel density in the peripapillary area was decreased in eyes with a dysthyroid optic neuropathy [27]. In addition, inactive TAO also showed an altered retinal perfusion as assessed in optical coherence tomography angiography [28, 29]. To our knowledge, understanding is still limited regarding retinal change in hyperthyroidism without TAO. Previous study indicated patients with thyroid dysfunction had wider retinal arterioles [30]. To better understand the mechanism of the DL-algorithm and to minimize the "black-box effect", we collected ocular fundus images taken in two general hospitals and four medical examination centers located

in North, East, South and West China. We used four different types of fundus cameras (TRC-NW200, CR-2 AF, nonmyd  $\alpha$ -DIII, and NT-2000). The DL-algorithm showed reliable and reproducible results in all these six external validation datasets. Although human experts cannot recognize the changes from retinal images in hyperthyroidism, the heatmap visualization revealed that the large retinal vessels and the optic nerve head area were the regions preferentially assessed by the DL-system. This may be explainable since hyperthyroidism is hypermetabolic condition, which leads to the increase of blood flow rate as well as the dilation of peripheral blood vessels. [31].

The DL system found in the present study may have clinical implications. The use of a computational evaluation of fundus photographs may be promising in screening individuals for hyperthyroidism and other thyroid diseases, so that unnecessary cost may be avoided and social burden is reduced. Future studies may evaluate whether the DL system can be integrated into mobile terminals, such as smart phone apps, to identify hyperthyroidism individuals from populations. Identification of hyperthyroidism at an early stage can assist clinicians to better organize future management strategies and provide more treatment options. Patients undergoing ophthalmic examinations and receiving eye surgeries may benefit from better understanding their general conditions such as thyroid function.

When the results of our study are discussed, its limitation should be taken into account. First, the DL system in this study was trained in participants with overt hyperthyroidism, so that the system was not tested to identify subclinical cases. Second, we failed to develop models to predict the precise levels of serum thyroid hormones, including TSH,  $T_3$ , and  $T_4$ . Third, we did not include participants with other thyroid diseases, such as hypothyroidism, thyroiditis, and thyroid cancer, due to the relatively low prevalence and few cases. Fourth, the study population consisted mostly of Chinese, so that the results may not directly be transferable on individuals of other ethnicities. Fifth, we collected data mainly from physical examinations, which included a large number of healthy subjects and relatively few hyperthyroidism patients. It caused the imbalance between hyperthyroidism cases and control groups. Sixth, although we established two models, only one model was applied in the external validation, and the performance of the model in external datasets were not perfect (all AUCs < 0.90). Seventh, in this study, we only considered thyroid diseases while other systemic diseases (e.g. hypertension) were not excluded. Although it might add some confounding factors, the results might be more close to real-world application. Eighth, although some confounding factors such as age and sex were matched, other parameters such as diastolic/systolic blood pressure with potential correlations between fundus images were not fully evaluated in this study. The strengths of our study include that it is the first artificial intelligence-based investigation to assess the association between retinal features and hyperthyroidism; that the DL system was developed from data obtained in 26 various data sources in China, with different settings such as hospitals and healthcare center; and that a total of four types of fundus cameras were used in the training and validation of the models, suggesting a wide applicability of the DL system.

## Conclusions

Retinal fundus photographs may serve for DL systems for a cost-effective and non-invasive method to detect hyperthyroidism.

## Abbreviations

AUC	Area under the receiver operator curve
BCE loss	Binary cross-entropy loss
CIs	Confidence intervals
CNN	Convolutional neural network
DL	Deep learning
PPV	Positive predictive value
NPV	Negative predictive value
ROC	Receiver operating characteristic
TAO	Thyroid-associated ophthalmopathy
TSH	Thyroid-stimulating hormone
T <sub>3</sub>	Triiodothyronine
T <sub>4</sub>	Thyroxine

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40537-023-00777-6>.

**Additional file 1: Table S1.** List of the location of each health screening centers and the digital retinal cameras used in each center. **Table S2.** The performance of several different state-of-the-art neural network architectures. We selected ResNet-50 for further external validation. **Figure S1.** The AUC changing curves on internal validation dataset for Model 1 and Model 2 using different neural network architectures.

## Acknowledgements

Not applicable

## Author contributions

DML and WW contributed to conception of the study. LD, LJ, SH, LL, JBJ, WW and DL contributed to study design. LD, LJ, SH, LL, XJ, ZN, RZ, WZ, HL, JD, JZ, ZH, YL, XW, XZ, CH, YC, ZW, JG, ZG, WW and DL contributed to acquisition of the data. LD, LJ, SH, LL, XJ, ZN, WW, and DL contributed to analysis and interpretation of the data. LD, LJ, SH, LL, ZN, RZ, WZ, HL, JD, JZ, ZH, YL, XW, XZ, CH, YC, ZW, JG, ZG, WW and DL contributed to drafting and revising the manuscript. All authors had access to all the data and DL were responsible for the decision to submit the manuscript. All authors read and approved the final manuscript.

## Funding

This study was supported by the National Natural Science Foundation of China (82071005, 82220108017, 82141128); The Special Fund of the Pediatric Medical Coordinated Development Center of Beijing Hospitals Authority (XTCX201824); the Research Foundation of Beijing Friendship Hospital, Capital Medical University (yyqdk2018-33); the Capital Health Research and Development of Special (2020–1-2052); Science & Technology Project of Beijing Municipal Science & Technology Commission (Z201100005520045, Z181100001818003).

## Availability of data and materials

The dataset generated and/or analysed during the current study are not publicly available due to patients' privacy, but are available from the corresponding author on reasonable request.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

Authors Lie Ju, Xin Wang, Xin Zhao, Chao He, Yu Zhong Chen, were employed by the company Beijing Airdoc Technology Co., Ltd. Beijing, China. Authors Zhao Hui Wang and Jian Xiong Gao were employed by the company iKang Guobin Healthcare Group Co., Ltd, Beijing, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential competing of interests.

Received: 20 January 2022 Accepted: 17 May 2023

Published online: 29 August 2023

## References

1. Mullur R, Liu YY, Brent GA. Thyroid hormone regulation of metabolism. *Physiol Rev*. 2014;94(2):355.
2. De Leo S, Lee SY, Braverman LE. Hyperthyroidism. *Lancet*. 2016. [https://doi.org/10.1016/S0140-6736\(16\)00278-6](https://doi.org/10.1016/S0140-6736(16)00278-6).
3. Hollowell JG, Staehling NW, Flanders WD, et al. Serum TSH, T(4), and thyroid antibodies in the United States population (1988 to 1994) national health and nutrition examination survey (NHANES III). *J Clin Endocrinol Metab*. 2002;87(2):489.
4. Ross DS, Burch HB, Cooper DS, et al. 2016 american thyroid association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. *Thyroid*. 2016;26(10):1343.
5. Garmendia Madariaga A, Santos Palacios S, Guillén-Grima F, Galofré JC. The incidence and prevalence of thyroid dysfunction in Europe: a meta-analysis. *J Clin Endocrinol Metab*. 2014;99(3):923.
6. Asban A, Chung SK, Tresler MA, et al. Hyperthyroidism is underdiagnosed and undertreated in 3336 patients: an opportunity for improvement and intervention. *Ann Surg*. 2018;268(3):506.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436.
8. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomed Eng*. 2018;2(3):158.
9. Varadarajan AV, Poplin R, Blumer K, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Visual Sci*. 2018;59(7):2861.
10. Shkolyar E, Jia X, Chang TC, et al. Augmented bladder tumor detection using deep learning. *Eur Urol*. 2019;76(6):714.
11. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402.
12. Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus images via deep learning. *Nat Biomed Eng*. 2020;4(1):18.
13. Xiao W, Huang X, Wang JH, et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *Lancet Digit Health*. 2021. [https://doi.org/10.1016/S2589-7500\(20\)30288-0](https://doi.org/10.1016/S2589-7500(20)30288-0).
14. Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health*. 2020. [https://doi.org/10.1016/S2589-7500\(20\)30216-8](https://doi.org/10.1016/S2589-7500(20)30216-8).
15. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ArXiv*. 2014;1409:1556.
16. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016.
17. Szegedy C, Vanhoucke V, Ioffe S, et al. 2016. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
18. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* 2017.
19. Kingma DP, Ba J. Adam a method for stochastic optimization. *arXiv*. 2014;1412:6980.
20. Ketkar N. Introduction to pytorch Deep learning with python. Berlin: Springer; 2017.
21. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision* 2017.
22. Chihara LM, Hesterberg TC. *Mathematical statistics with resampling and R*. Hoboken: John Wiley & Sons; 2018.
23. Rim TH, Lee G, Kim Y, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *Lancet Digit Health*. 2020;2(10):e526.
24. Sato W, Hoshi K, Kawakami J, et al. Assisting the diagnosis of graves' hyperthyroidism with bayesian-type and som-type neural networks by making use of a set of three routine tests and their correlation with free T4. *Biomed Pharmacother*. 2010;64(1):1.
25. Aoki S, Hoshi K, Kawakami J, et al. Assisting the diagnosis of Graves' hyperthyroidism with pattern recognition methods and a set of three routine tests parameters, and their correlations with free T4 levels extension to male patients. *Biomed Pharmacother*. 2011;65(2):95.
26. Lai FHP, lao TWU, Ng DSC, et al. Choroidal thickness in thyroid-associated orbitopathy. *Clin Exp ophthalmol*. 2019;47(7):918.
27. Zhang T, Xiao W, Ye H, et al. Peripapillary and macular vessel density in dysthyroid optic neuropathy an optical coherence tomography angiography study. *Invest Ophthalmol Vis Sci*. 2019;60(6):1863.
28. Mihailovic N, Lahme L, Rosenberger F, et al. Altered retinal perfusion in patients with inactive graves ophthalmopathy using optical coherence tomography angiography. *Endocr Pract*. 2020;26(3):312.
29. Yu L, Jiao Q, Cheng Y, et al. Evaluation of retinal and choroidal variations in thyroid-associated ophthalmopathy using optical coherence tomography angiography. *BMC Ophthalmol*. 2020. <https://doi.org/10.1186/s12886-020-01692-7>.
30. Teo L, Cheung C, Tay WT, Wong TY. Associations between thyroid dysfunction and retinal microvascular changes. *Invest Ophthalmol Vis Sci*. 2011;52(14):5106.
31. Devereaux D, Twelde SZ. Hyperthyroidism and thyrotoxicosis. *Emerg Med Clin North Am*. 2014;32(2):277.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.