J. Health Inform. 2025 Janeiro-Dezembro; 17

Artigo Original

DOI: https://doi.org/10.59681/2175-4411.v17.2025.1441

JHI JOURNAL OF HEALTH INFORMATICS

Artificial intelligence for cataract diagnosis and referral using real-world database

Inteligência artificial para diagnóstico e encaminhamento de catarata usando dados reais

Inteligencia artificial para diagnóstico y remisión de cataratas utilizando datos reales

Marcelo Negreiros[1], Ronaldo Husemann[1], Valter Roesler[1], Aline L. de Araujo[2], Dimitris Rucks Varvaki Rados[2], Felipe C. Cabral[2], Natan Katz[2]

ABSTRACT

The use of artificial intelligence (AI) for ophthalmology applications has shown g promising results worldwide; however, its performance is dependent on population groups and must be evaluated in real-world scenarios. We evaluated the use of AI for cataract diagnosis and referral to specialists using a real-world database consisting of 2642 eye images from a working telemedicine service in South Brazil. Our AI solution adopts an ensemble model to improve classifier performance. The best results showed an accuracy of 90.6% for cataract diagnosis with a corresponding area under the receiver operating characteristic curve (ROC AUC) of 96.7%. The accuracy for surgical referral was 86.5% with a corresponding ROC AUC of 94.3%. The results indicate that the use of an ensemble of models and training with a heterogeneous real-world clinical database enabled our solution to achieve superior performance compared to other works in the literature when evaluated on real-world data.

Descriptors: Artificial Intelligence; cataract; diagnosis.

RESUMO

O uso de inteligência artificial (IA) para aplicações oftalmológicas têm mostrado resultados promissores, entretanto, seu desempenho depende do grupo populacional amostrado e precisa ser avaliado em cenários reais. Propomos o uso de IA para diagnóstico de catarata e encaminhamento para especialista usando uma base de dados real, composta por 2642 imagens oculares, de um serviço de telemedicina no sul do Brasil. Nossa solução adota um modelo composto para aprimorar o desempenho dos classificadores. Os melhores resultados mostram acurácia de 90,6% para diagnóstico, com área correspondente sob a curva de operação característica do receptor (ROC AUC) de 96,7%. A acurácia para encaminhamento de catarata para cirurgia foi de 86,5%, com ROC AUC de 94,3%. Os dados obtidos apontam que o uso de um modelo composto treinado com uma base clínica heterogênea real permitiu que nossa solução atingisse desempenho superior a outros trabalhos da literatura quando avaliados com dados do mundo real.

Descritores: Inteligência Artificial; catarata; diagnóstico.

RESUMEN

El uso de inteligencia artificial (IA) para aplicaciones oftalmológicas ha mostrado resultados prometedores; sin embargo, el desempeño depende de los grupos de población y debe evaluarse en escenarios reales. Evaluamos IA para el diagnóstico de cataratas y derivación a especialistas utilizando una base de datos del mundo real, compuesta por 2642 imágenes oculares, de un servicio de telemedicina del sur de Brasil. Nuestra solución adopta un conjunto compuesto para mejorar el rendimiento de los clasificadores. Los mejores resultados muestran una precisión del 90,6 % para el diagnóstico con área bajo la curva característica operativa del receptor (ROC AUC) del 96,7%. La precisión de la derivación de cataratas fue del 86,5% con AUC ROC del 94,3%. Los resultados indican que el uso de un modelo compuesto entrenado con una base heterogénea real permitió que la solución lograse un rendimiento mayor que otros trabajos cuando fueron evaluados con datos reales.

Descriptores: Inteligencia Artificial; catarata; diagnóstico.

[1]*Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil.*
[2]*Teleoftalmo, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil..*

Autor Correspondente: **Ronaldo Husemann**
e-mail: **ronaldo.husemann@ufrgs.br**

Artigo recebido: 20/12/2024
Aprovado: 12/08/2025

INTRODUCTION

Cataract, the opacification of the lens of the eye, is one of the main causes of vision impairment worldwide and blindness in developing countries[1]. The number of individuals affected grow with the increasing life expectancy and the population aging. This process burdens the already limited ophthalmic services and resources, mainly in developing countries [2]. AI-based applications in ophthalmology promise to address growing healthcare needs. Also Ting et al (2021) [3] presented a systematic review of different AI-based software applications in cataract patient management.. Lin, Liu and Wu (2021) [4] show an overview of AI applications for various diseases in the anterior segment of the eye, including the detection and grading of cataracts. Additional literature reviews of machine learning applications in ophthalmic imaging modalities can be currently found [1, 5, 6].

Fundus eye images and AI techniques have been applied to diagnose diseases such as diabetic retinopathy, age-related macular degeneration, and glaucoma. A combined detection of glaucoma, diabetic retinopathy, and cataracts using fundus images is presented by Orfao and Haar [7]. However, there are still technical and legal challenges to overcome for further AI adoption in ophthalmology [8]. Among these challenges is the restricted population in which the AI studies were conducted: small datasets and a lack of diverse populations hinders generalization. Also, research studies often use data collected in clinical trials or academic databases, which translates into high-standardized imaging datasets. This also limits external validity, data from real-world clinical settings offers varying levels of quality and standardization, which can be challenging for model development and testing [(2].

The use of fundus photography for cataract assessment, common in recent approaches, is unable to diagnose cortical cataracts. Furthermore, these images typically require extra efforts to distinguish poor-quality from normal-quality images in real-world conditions [9]. The use of anterior images achieved good results for cataract detection [10]. However, when applied in real-world cases there is a significant performance reduction, mainly due to the presence of image artefacts, such as illumination problems and noise [10, 11]. Most literature studies on cataracts traditionally focus on cataract detection and grading [12, 13, 14]. From the public health perspective, the main goal in relation to cataracts is to diagnose the condition and define if the patient will benefit from cataract surgery.

In this work, we evaluate the use of AI for cataract diagnosis and referral to a specialized care. Cataract diagnosis is used to classify if the eye has a normal lens or not. The amount of lens opacity that lead to a positive diagnosis is rather subjective and depends on ophthalmologists to define if the cataract is clinically relevant to cause visual problems. Referral is defined as a patient whose level of lens opacity justifies a surgical evaluation. It should be noted that visual acuity loss per se is not sufficient for referral, as other diseases could cause visual impairment. The dataset was extracted from a working telemedicine service in the southern region of Brazil. The research project has been approved by the institutional review board from Hospital de Clínicas in Porto Alegre (under the number 27764620400005327). Written consent was obtained from all patients, parents or guardians.

The diagnosis and the decision for referral were made by the ophthalmologist responsible for each patient. The dataset labels were composed of mydriatic exams (when the pupil of the eye is dilated) from a working database containing patient evaluations reported by one of the several ophthalmologists from the attending team. It represents a real-world heterogeneous database (composed of patients with different ages and ethnicities) with greater label noise than academic databases.

We adopted convolutional neural networks (CNN), using pre-trained models to reduce training time and label noise [15]. Practical results indicated that no single model solution reached adequate accuracy due to the inherent database label noise. In this work, an ensemble of models was used to minimize label noise effects and improve CNN convergence. Additionally, we evaluated effect of extra data, such as visual acuity and patient age, on the models. The best results obtained were superior to other existing works with real-world results and are comparable to approaches using academic databases. Based on that, we consider this proposed hybrid solution (which considers image and additional patient information), trained with real-world database, as a relevant and efficient contribution as practical tools to public health use (i.e., defining the need for ophthalmologic referral and surgical consultation). By using data from a real-world telemedicine service, this study incorporates diagnostic evaluations provided by ophthalmologists who directly assessed the patients. This approach addresses a key gap in the literature by moving beyond binary classifications of "diseased" or "healthy" to identify clinically significant cataracts that is those requiring referral for surgical treatment. This paper is structured as follows: Section 2 provides database details along with the AI techniques used. The results are presented in Section 3, including evaluation of models, image pre-processing effects, and results of experiment that includes clinical features. Section 4 provides a comparison to similar works, followed by conclusions.

MATERIALS AND METHODS

In this section, basic background information used in this work is defined, including performance assessment of binary classifiers. Additionally, the relevant AI methods used are briefly discussed, followed by a description of our working database.

Performance Measures for Binary Classifiers

In this section, we briefly define common performance measures for binary classifiers used in this work. The section assumes a generic binary classifier and the unfamiliarized reader is referred to related works [16, 17].

Considering a population of P positive cases and N negative cases, a classifier will correctly label some of the cases (true positives(TP) and true negatives(TN)) and others (false positives (FP) and false negatives (FN)). The classifier accuracy (ACC) in (1) is the ratio of correct classifications to the total population. The specificity (SPE) (also true negative rate) in (2) is the ratio of correct classifications to negative cases. Sensitivity (SEN) in (3) is the ratio of correct classifications to positive cases. Finally precision (PRE) is the ratio of correct positive classifications to all positive ones.

$$ACC = (TP + TN) / (P + N) \quad (1)$$

$$SPE = TN / (FP + TN) = TN / N \quad (2)$$

$$SEN = TP / (TP + FN) = TP / P \quad (3)$$

$$PRE = TP / (TP + FP) \quad (4)$$

A receiver operating characteristic (ROC) graph is a technique for visualizing and selecting binary classifiers based on their discrimination threshold performance [17,18]. Since AI models produce a probabilistic output to provide a binary output, this technique is used to compare classifiers. The area under the ROC curve (ROC AUC) provides an estimate of classifier's performance using a single number.

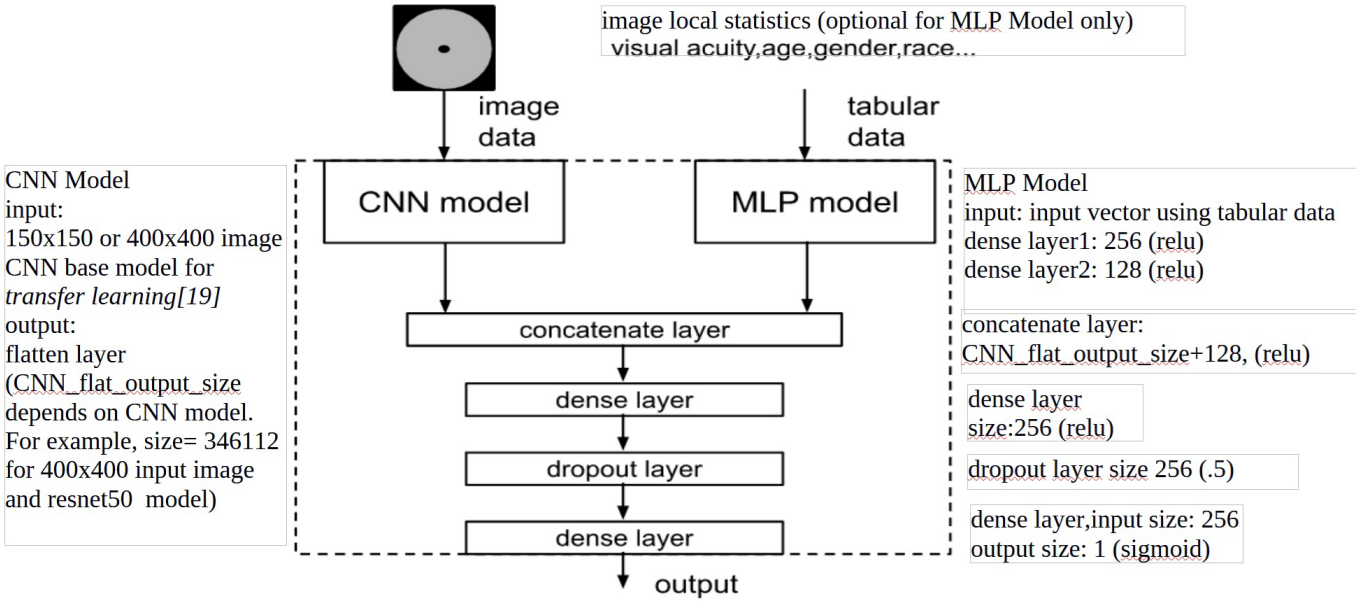
Artificial Intelligence Methods

Two problems were addressed: identifying cataract diagnosis and detecting referable cataracts. For diagnosis, the AI attempts to classify whether an image has a cataract or not. For referable cataracts, the AI classifies whether it is a candidate for surgical treatment or not (as this is the criteria for referral). In the following subsections, several topics regarding models, methods and database development used in this work are explained.

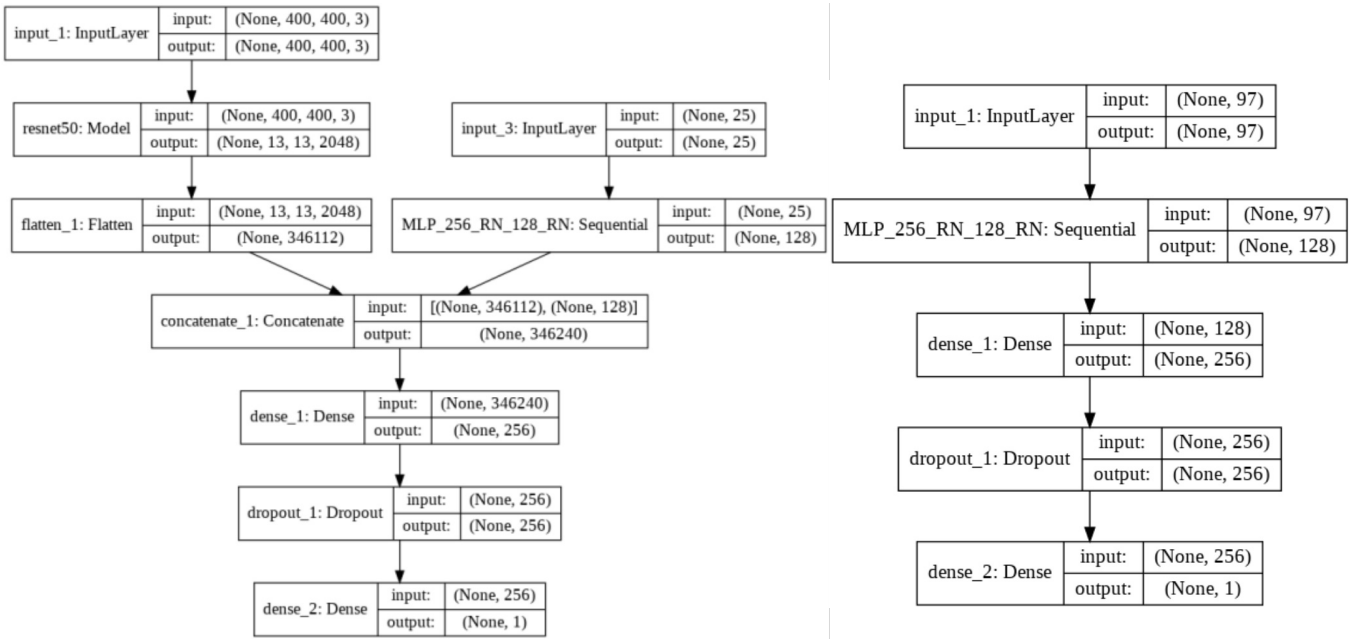
Single models and ensembles

It is well known that CNN models can classify images [19]. Since ophthalmologists use additional patient clinical information such as age and visual acuity level to improve accuracy of medical diagnosis, we hypothesized that an AI model designed for diagnosis could increase performance by processing such additional data. For this study, we utilized an AI solution that combines a pre-trained CNN with an additional multi-layer perceptron (MLP) model (Figure 1).

Figure 1. (a) Structure of the mixed-mode model used in this work and (b) example report from Keras for a mixed-model using resnet50 CNN and associated MLP for tabular data, and (c) report from an MLP-only model.



(a)

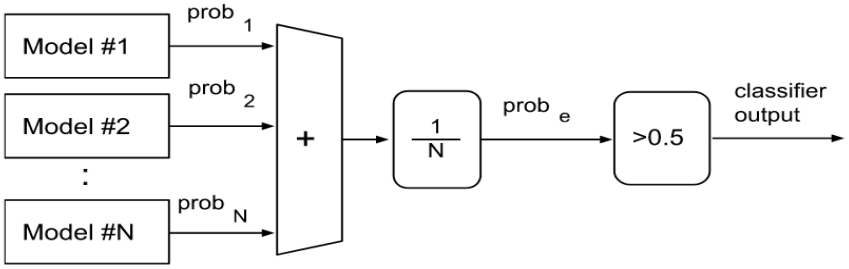


The AI models were developed based on pre-trained CNN models available in the Keras software library [19]. Different image resolutions and CNN models were used for pupil-only (150x150) and complete images (400x400).

An MLP-only model was also used, in which the input image was divided into 17 zones defined by a circular grid mask, and local statistics were evaluated, without the need for a CNN model. The image statistics are transformed to a tabular data format and fed into the MLP model along with other tabular data (Figure 1). This approach is similar to that used by Gao et al. (2013) [20] but utilizes AREDS mask [12].

Multiple models were used in this work in order to improve results [18]. A diagram of the used ensemble model in this work is presented in Figure 2. The probability outputs of the *N* individual models in the ensemble are added with equal weights (1/N). The combined probability is then thresholded to the final classifier output A similar approach was adopted by Chauhan et al (2018) [18], where distinct pre-trained models (VGG-19, ResNet101V2 and InceptionV3) were combined for cataract diagnosis using eye fundus images. In this work an exhaustive search was performed in order to evaluate the model combination providing better accuracy in the training data.

Figure 2. Diagram of used ensemble model.



Database development

In this section, details of the database used in this work are presented. The database is a subset of a working database from a telemedicine service in southern Brazil. The "Teleoftalmo" initiative in South Brazil is briefly introduced, describing the equipment and procedures used in data acquisition. The process of image selection and classification used in this work is also presented.

The Teleoftalmo initiative in South Brazil

Teleoftalmo [21] is a telediagnostic service that operates in Rio Grande do Sul, the southernmost state of Brazil. The main objective of this service is provide ophthalmologic diagnosis closer to individuals homes, reducing displacement and waiting times. A central office is located in the city of Porto Alegre. Patient data collection and examination sites are distributed over different health districts in seven cities of the state. The Teleoftalmo initiative has performed more than 30,000 telediagnoses since 2017. The examinations include visual acuity measurements, eye refraction, intraocular pressure measurement, anterior segment and fundus photography, eyelid

evaluation, ocular motility, and pupillary tests. Specialized ophthalmologic equipment is used for fundus and anterior eye segment photography (Zeiss Visucam 224 Fundus Imaging) and for eye refraction measurements (Zeiss Visuref 100 Autorefractor/Keratometer).

Databases developed

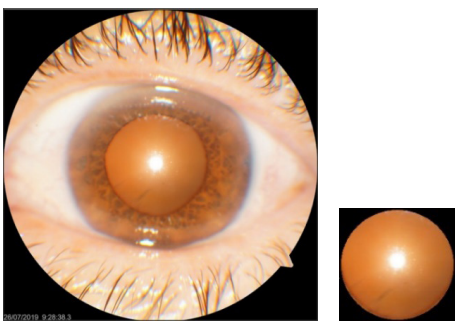
Data-oriented approaches require specific databases to evaluate different problems. The problems addressed include whether AI can operate using complete images or if it needs pre-segmented images of the pupil. Additionally, we evaluated the impact of extra patient data to perform cataract diagnosis and referral.

To evaluate if AI can identify automatically the region of interest in the images, two types of images were used: complete eye and pupil-only images. To check the performance for cataract diagnosis and referable cataract, a database was developed for each problem. Finally, to evaluate the performance when extra patient data is used, tabular data and image data were used in the developed databases. The databases were constructed using digital photographs of eye anterior segment, patient demographic information (gender, race, age), data acquisition details (date, left or right eye, field angle), diagnostic information, and the final medical report of every patient.

Image acquisition and pre-processing

The images used in this work are photographs of the eyes' anterior segment obtained using a professional retinal camera. An image typically includes the pupil, iris, part of the sclera, the eyelids, and eyelashes. Images were captured using a strong light source (flash) focused on the pupil and a field angle of 30°. Flash reflection is always present in the images. Images were retrieved in PNG format from the Teleoftalmo database with a resolution of at least 2448x2448 pixels. Figure 3a shows a typical complete image. To evaluate the performance of the AI with and without prior pupil segmentation, two different image datasets were generated from the selected images of the database: complete images and pupil images (pupillary region only).

Figure 3. Example of the same patient photograph in two formats: (a) complete image and (b) cropped pupil-only image, as used in this work.



(a) **(b)**

The complete images were rescaled to a resolution of 400x400. For the pupil images dataset, the original (unscaled) images were processed by a dedicated pupil area detection algorithm designed in Matlab, considering segmenting pupil color and shape to allow automatic area selection. The pupil area was cropped from the original image, and a circular mask was used to exclude non-pupil regions. The average resolution of the cropped pupil images was 800x800. The pupil images were rescaled to a resolution of 150x150. The pupil-only and complete image sizes were chosen to keep the average pupil size as close as possible in both sets. Figure 3b shows a segmented pupil image obtained by cropping and masking the complete image.

Image selection

Image selection for the cataract diagnosis database was based on annotations in the Teleoftalmo working database indicating the presence of cataracts in the selected eye of each patient. Each image was then reviewed by the authors to ensure minimum pupil dilation and image quality (acceptable focus and flash intensity), while also avoiding occlusions (partially closed eyes, eyelashes) or the presence of artifacts caused by other eye pathologies.

In this work, the minimum accepted pupil diameter was arbitrarily set to one-third of the external iris diameter. Images of non-dilated pupils were discarded during the creation of the database. The medical report was also examined to confirm the cataract diagnosis and determine whether it was a referable cataract or not (indicating whether the patient's eye should be directed for surgical treatment). Inconsistent cataract images and medical reports may occur when other eye pathologies that require urgent treatment, such as age-related macular degeneration (AMD), are present. Such cases were not included in the database. Eyes that had already undergone cataract surgery and had artificial intraocular lenses were also excluded from the database.

The resulting selection of images contains three types: eyes without cataracts, eyes with a positive diagnosis of cataracts but not eligible for correction surgery (non-referable), and eyes with a positive diagnosis of cataracts eligible for surgery (referable). Two different databases were generated: one for cataract diagnosis evaluation (images with and without cataracts) and one for referable cataract evaluation (images with referable cataracts and images with non-referable cataracts). In total, 2642 eyes from 1544 different patients were used in the construction of the two databases, with only a single image of each eye being used.

Additional patient data in tabular form

When examining a patient, an ophthalmologist routinely uses information such as the patient's age and visual acuity when analyzing eye photographs. It is expected

that incorporating additional data would enhance the performance of an AI classifier. In this study, both image-only classifiers and classifiers that utilized both images and additional tabular patient data were employed. The additional patient data included the patient's age, gender, self-declared race (skin color), visual acuity measurement tests (including the best visual acuity measurement), and measurements from the autorefractor eye exam.

Database statistics

A total of 2642 images were selected from the Teleoftalmo database. A single image of each eye was used, and only images with a field angle of 30 degrees were selected. Two different databases were generated: one for cataract diagnosis and one for referable cataracts. Each database contains two classes with the same number of images in each class. The total number of images for cataract diagnosis was 2160 (1080 normal eyes and 1080 cataract eyes), and the total number of images for cataract referral was 1824 (912 referable cataracts and 912 non-referable cataracts and normal eyes). Among images with a positive diagnosis, 68% were from female patients, while for surgical referral, 63% were from female patients.

RESULTS

In this section, the results obtained are presented using accuracy and ROC AUC for model comparison. In this work, we aimed at two target outputs, cataract diagnosis and referral. For each target, complete images and pupil-only images were used to evaluate the need for previous segmentation of the images. We also studied the effect of additional patient features such as visual acuity and patient age on the models.

Test environment

Google Colaboratory [22] was used as the development platform for this work. The dataset was divided into six folds, where five folds were used for training and validation, and the remaining fold was kept as a holdout test set. After validation, the models were discarded, and a final model was trained using all training/validation data. The final model was then tested on the holdout test set.

During the development of this work, the versions of Python, Keras, TensorFlow (TF), and Scikit-learn were 3.7.10, 2.3.1, 1.15.2, and 0.22.2.post1, respectively. TensorFlow 1.15.2 was selected as it was the only TF 1.x version available on Google Colaboratory, and different versions of TF 2.x were noticed in different virtual machines during initial tests.

Pre-trained CNN models available in Keras were used in this study, including DenseNet201, DenseNet169, DenseNet121, ResNet152, ResNet101, ResNet50, VGG16, VGG19, MobileNet, InceptionResNetV2, Xception, and InceptionV3. A standard transfer learning approach of the CNN ImageNet models provided in Keras was used [19]. An MLP-only model was also used as described earlier in Section 2. For cases where extra data was provided, an MLP network was used in parallel with the pre-trained CNNs. The models were fine-tuned for up to 100 epochs. The average accuracy during validation was used to select the best model for each scenario.

During single model validation and testing, all individual model responses and probabilities were recorded. During the model ensemble evaluation, results from each individual model were loaded. Since 13 different models were available, the total number of model combinations is 2^13=8192. To ease this evaluation, all individual model responses and probabilities were recorded in a file when the single model validation and testing was performed. This way, model ensemble evaluation search used previously evaluated results from each individual model. As presented in Figure 2, the single model probabilities were added together and divided by the number of models in the ensemble to generate a new ensemble probability. The new ensemble probability was then thresholded to generate a binary classification. The average accuracy during validation was used to select the best ensemble model combination for each scenario.

Overall results

Since this was as exploratory work, several models were evaluated and compared as described in Table 1. For example, using a single model is described as "model type: single". Two columns are used to indicate the use of tabular data in addition to image data ("image only" and "image + extra data").

Table 1 presents accuracy and ROC AUC results for single and ensembles models. To evaluate the results, a specific scenario must be selected: the model type (single or ensemble), the database (diagnosis or referral), the image type (pupil or complete image), and the use of extra data ("image_only" or "image+extra data"). For each scenario, the number of models, accuracy, and ROC AUC are provided. All data was evaluated on the holdout test set. Observing Table 1, it is evident that the use of extra data enhances accuracy and ROC AUC results in all scenarios compared to the image-only case, with an average improvement of 7.7% for accuracy and 5.4% for ROC AUC. Ensembles outperformed single models in the same scenarios. Additionally, ensembles reduced the performance gap between complete image and pupil-only images, with an average difference of 0.8% for accuracy and 0.2% for ROC AUC.

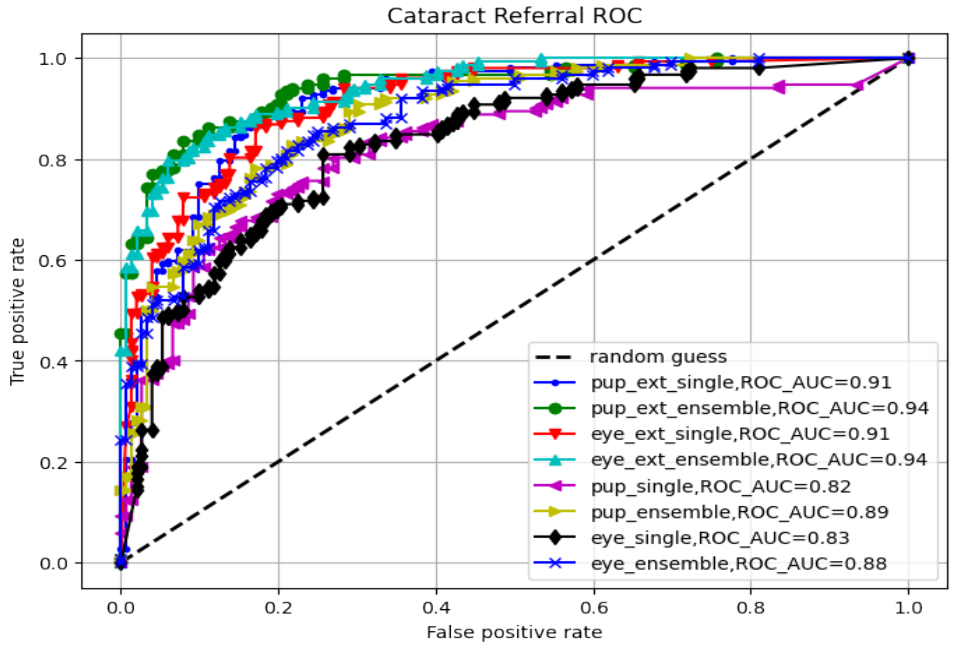
Table 1. Number of models, Accuracy and ROC AUC of the final models evaluated in the test set (holdout set)

			Number of models		Accuracy (ACC)		ROC AUC	
Model type	**Database**	**Image**	**Image only**	**Image + extra data**	**Image only**	**Image + extra data**	**Image only**	**Image + extra data**
Single	referral	complete eye	1	1	0.740	0.836	0.831	0.914
		pupil	1	1	0.760	0.852	0.823	0.914
	diagnosis	complete eye	1	1	0.836	0.886	0.920	0.950
		pupil	1	1	0.867	0.906	0.945	0.954
Ensemble	referral	complete eye	6	6	0.789	0.865	0.883	0.943
		pupil	7	3	0.799	0.852	0.885	0.944
	diagnosis	complete eye	5	7	0.872	0.906	0.950	0.967
		pupil	7	5	0.867	0.919	0.947	0.973

Cataract referral and diagnosis results

Figure 4 presents ROC curves for cataract referral. The plotted curves correspond to the eight scenarios for each classifier (complete or pupil-only image, use of extra data, single or ensemble model) and a random guess classifier for reference.

Figure 4. ROC curves for cataract referral.



The higher ROC_AUC values presented indicate the curves closer to the ideal binary classifier (true positive rate = 1 and false positive rate = 0). The plots obtained using extra data and ensemble models (pup_ext_ensemble and eye_ext_ensemble) have the highest ROC_AUC (0.94) and the best performance.

Figure 5 presents ROC curves for cataract diagnosis. The plotted curves correspond to the eight scenarios for each classifier and a random guess classifier for reference. The plots obtained using extra data and ensemble models (pup_ext_ensemble and eye_ext_ensemble) have the highest ROC_AUC (0.97).

Figure 5. ROC curves for cataract diagnosis.

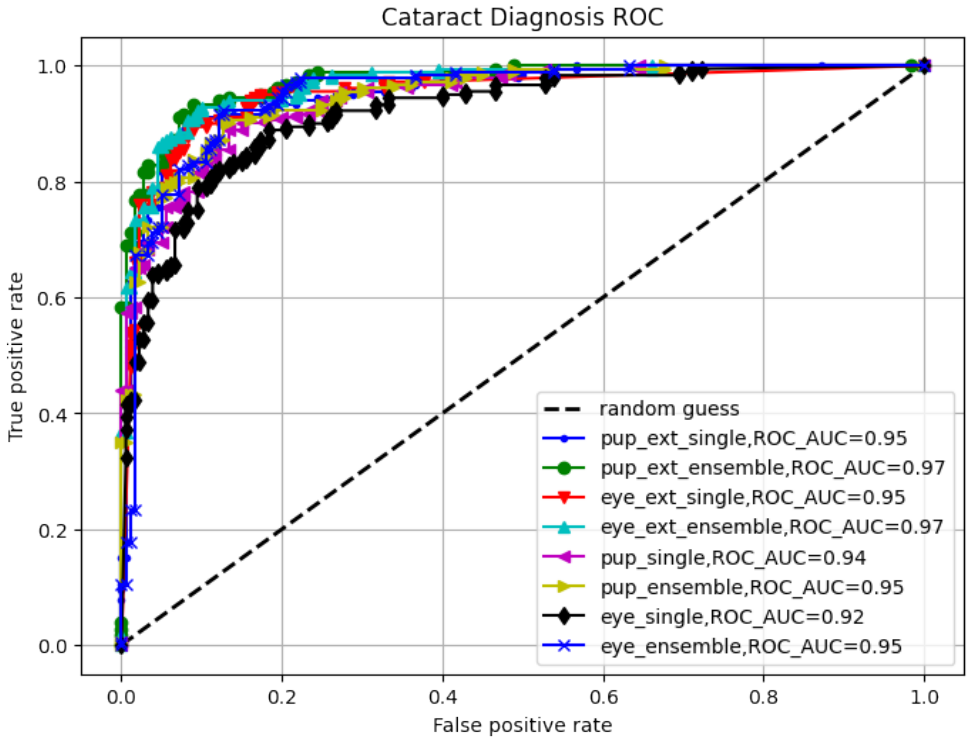


Table 2 presents the best cataract referral results obtained using an ensemble of models, while Table 3 presents the best cataract diagnosis results. The results indicate that extra data improves accuracy, ROC_AUC, sensitivity, and specificity results. The extra data, in column 1 of both tables, represent tabular data in addition to image data such as visual acuity and patient age.

Table 2. Best ensemble model results for cataract referral and the use of extra tabular data from the patient (extra data)

Use of extra data	Accuracy (ACC)	ROC_AUC	Sensitivity (SEN)	Specificity (SPE)	Models composing the ensemble
no	0.789	0.883	0.717	0.862	DenseNet121, ResNet101, InceptionResNetV2, Xception, InceptionV3, MLP_256_RN_128_RN_A
yes	0.865	0.943	0.842	0.888	DenseNet121-256_RN_128_RN_A, ResNet101-256_RN_128_RN_A, MobileNet-256_RN_128_RN_A, InceptionResNetV2-256_RN_128_RN_A, InceptionV3-256_RN_128_RN_A,MLP_256_RN_128_RN_A

Table 3. Best ensemble model results for cataract diagnosis and the use of extra tabular data from the patient (extra data)

Use of extra data	Accuracy (ACC)	ROC_AUC	Sensitivity (SEN)	Specificity (SPE)	Models composing the ensemble
no	0.872	0.950	0.833	0.911	DenseNet201, DenseNet121, ResNet50, MobileNet, MLP_256_RN_128_RN_A
yes	0.906	0.967	0.883	0.928	DenseNet201-256_RN_128_RN_A, DenseNet121-256_RN_128_RN_A, MobileNet-256_RN_128_RN_A, InceptionResNetV2-256_RN_128_RN_A, Xception-256_RN_128_RN_A, InceptionV3-256_RN_128_RN_A, MLP_256_RN_128_RN_A

DISCUSSION

In this work, complete images of the anterior part of the eye and pupil-only images were used to determine if the AI classifier would be able to work without the need for previous pupil segmentation. Although more information is present in the complete images, the use of an ensemble of models allows their use without the need of segmentation. This way, the requirement for costly pre-segmentation of the region of interest (pupil) in the images may be disregarded.

The results of models using complete images were selected for cataract referral and diagnosis. The combined use of extra patient data and model ensembles provided the best performance metrics. A cataract diagnosis accuracy of 90.6% and ROC_AUC of 96.7% were achieved in this work considering complete images, model ensembles, and the use of extra patient data. A cataract referral accuracy of 86.5% and ROC_AUC of 94.3% were achieved for the same conditions. Therefore, the cataract referral classification system had lower performance than the cataract diagnosis system, as indicated by the best accuracy for each task.

It is interesting to notice that when additional patient data is not included, the performance decreases. A cataract diagnosis accuracy of 87.2% and ROC_AUC of 95.0% were achieved considering complete image images and model ensembles. A cataract referral accuracy of 78.9% and ROC_AUC of 88.3% were achieved for the same conditions.

Table 4. Comparison of related cataract diagnosis and referral approaches.

	Database type	**Cataract Diagnosis**		**Cataract Referral**		
Reference	**Training/Testing**	**ACC (%)**	**ROCAUC (%)**	**ACC (%)**	**ROC_AUC (%)**	**Image Type**
(10)	academic/academic	98.68	99.93	88.00	94.88	anterior segment, mydriatic, diffuse
THIS WORK (image and extra data)	real-world/real-world	90.60	96.70	86.50	94.30	anterior segment, mydriatic, flash
(10)	academic/real-world a	88.79	95.96	79.50	91.51	anterior segment, non-mydriatic, diffuse
THIS WORK (image only)	real-world/real-world	87.20	95.00	78.90	88.30	anterior segment, mydriatic, flash
(25)	real-world/real-world		90.8 (frames) 85.7(videos)			mydriatic , non-mydriatic, slip-lamp
(9)	academic/academic	84.3	91.62			fundus
(11[	academic/academic	93.9				fundus
(13)	academic/academic	97.0				fundus
(14)	academic/academic		94.80			fundus
(18)	academic/academic	95.0				fundus
(23]	academic/academic	98.0				anterior segment
(24)	academic/academic	96.3				fundus

ROC_AUC=area under the ROC curve;

A selection of related works is presented in Table 4. Several works are based on fundus images for cataract detection [9, 13, 14, 23, 24, 25]. Although high accuracy in cataract diagnosis is achieved, cataract referral is not addressed. Wu et al. (2019) [10], like our solution, presents results for cataract detection and referral, being tested in academic and real-world scenarios. The obtained results showed a significant reduction in performance when a deep learning approach, only trained with academic databases, is applied to real-world cases. The main reason for that performance difference is the incidence of distinct noise interferences in real-world images, like problems with inadequate focus, reflections, flash and bad illumination artefacts and others.

It should be noted, however that Wu et al. (2019) [10] uses non-mydriatic images and our work is based on mydriatic images. Therefore, we alternately compare our results withShimizu et al (2023) [25] which is, to the authors' knowledge, the only published AI work with real-world results of cataract detection, using mydriatic eye images. Shimizu et al (2023) [25] evaluates a novel solution of machine learning (ML) to diagnose cataracts comparing results when the ML algorithm is trained using isolated images (frames) or recorded videos, using real data collected from a Japanese ophthalmology institution (Yokohama Keiai Eye Clinic). Table 5 presents a comparison of real-world results including sensitivity and specificity.

Table 5. Summary statistics comparison for real-world data

	Cataract diagnosis				Cataract referral			
Reference	ACC (%)	ROC_AUC (%)	SEN (%)	SPE (%)	ACC (%)	ROC_ AUC (%)	SEN (%)	SPE (%)
THIS WORK (image only)	**87.2**	**95.0**	**83.3**	**91.1**	**78.9**	**88.3**	**71.7**	**86.2**
(25) a		90.8						
(25) b		85.7						
(10) c	88.79	95.96	92.00	83.85	79.50	91.51	73.0	86.0
THIS WORK (image and extra data)	**90.6**	**96.7**	**88.3**	**92.8**	**86.5**	**94.3**	**84.2**	**88.8**

ROC_AUC=area under ROC curve;

a real-world experiment using isolated mydriatic images (frames)

b real-world experiment using sequence of mydriatic images (videos)

c real-world test in AI ambulatory experiment using non-mydriatic images

To compare our results with those of Shimizu et al (2023) [25] we only considered cataract detection results obtained using mydriatic images (the whole work of Shimizu et al (2023) [25] presents results for cataract detection and grading using mydriatic and non-mydriatic images). Considering that, we can clearly note that our solution was able to reach superior performance (ROC_AUC) in both scenarios (when Shimizu et al (2023) [25] use isolated images or videos).

Furthermore, when comparing our solution with that of Wu et al. (2019) [10], we can observe that our first version (which uses only images to performs cataract diagnosis) obtained inferior results. However, when image and additional patient data are incorporated, we achieved accuracy, ROC_AUC and specificity in both purposes (diagnosis and referral).

It important to cite that the use of an ensemble of models will have a high computational cost. The use of single models and additional data was shown in Table 1 to provide similar results to ensembles and additional data and would be a viable option for the implementation in a practical scenario, that have a limited computational infrastructure. The use of pupil images would require additional computational cost to isolate the pupil, so a solution that uses the complete eye may be preferred.

CONCLUSION

In this work, an AI approach to cataract diagnosis and referral was performed using a real-world data. By utilizing a real-world database, we inherently assume that there will be noise in the dataset labels. To address the label noise issue, fine-tuned pre-trained ImageNet CNN models were used. To improve the individual models performance, our work adopted an ensemble of models. The use of additional patient data, such as visual acuity and age, combined with eye image, considerably improved the results, particularly for cataract referral. These results can inform future decision-making when designing eye care strategies augmented by AI. Other works trained on academic databases present high accuracy, but their performance significantly decreases when applied in real--world scenarios[10]. Furthermore, our results are superior to those obtained by recent AI solutions, which use mydriatic and non-mydriatic images, when considering their application in a real-world setting.

Regarding potential applications, our system can have clinical and public health utilities. Clinically, the system could enable automated screenings in both primary care settings and through remote consultations. By assisting physicians and healthcare professionals with clinical decision support and patient prioritization based on the need for surgical consultation, our system can contribute to healthcare efficiency. Ensuring timely intervention of cataracts can help prevent vision impairment or blindness associated with advanced cataracts. Public health initiatives could leverage AI for large-scale screening programs and epidemiological studies, contributing significantly to the reduction of healthcare costs through efficient resource utilization.

REFERENCES

1. Gutierrez L, Lim JS, Foo LL, Ng WY, Yip M, Lim GYS, et al. Application of artificial intelligence in cataract management: current and future directions. Eye Vis. 2022;9(1):1–11.
2. Gunasekeran DV, Wong TY. Artificial intelligence in ophthalmology in 2020: A technology on the cusp for translation and implementation. Asia-Pacific J Ophthalmol. 2020;9(2):61–6.
3. Ting DSJ, Foo VHX, Yang LWY, Sia JT, Ang M, Lin H, et al. Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology. British Journal of Ophthalmology. 2021.

4. Lin H, Liu L, Wu X., Artificial Intelligence for Cataract Management. Artificial Intelligence in Ophthalmology. 2021:203–6.
5. Tong Y, Lu W, Yu Y, Shen Y. Application of machine learning in ophthalmic imaging modalities. Eye Vis. 2020;7(1).
6. Zhang XQ, Hu Y, Xiao, ZJ, Fang JS, Higashita R, Liu J, Machine learning for cataract classification/grading on ophthalmic imaging modalities: A survey. Mach. Intell. Res. 2022;19(3):184–208.
7. Orfao J, van der Haar D. A Comparison of Computer Vision Methods for the Combined Detection of Glaucoma, Diabetic Retinopathy and Cataracts. In: Lecture Notes in Computer Science. 2021.
8. Alam M, Hallak JA. AI-automated referral for patients with visual impairment. Lancet Digit Heal [Internet]. 2021;3(1):e2–3
9. Wu X, Xu D, Ma T, Li ZH, Ye Z, Wang F, et al. Artificial Intelligence Model for Antiinterference Cataract Automatic Diagnosis: A Diagnostic Accuracy Study. Front Cell Dev Biol. 2022;10(July):1–11.
10. Wu X, Huang Y, Liu Z, Lai W, Long E, Zhang K, et al. Universal artificial intelligence platform for collaborative management of cataracts. British Journal of Ophthalmology. 2019;103(11):1553–60.
11. Pratap T, Dhulipalla VR, Kokil P. Computer-aided Cataract Diagnosis with Fundus Retinal Images under Noisy Conditions. IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) 2024:1–18.
12. Kassoff A, Kassoff J, Mehu M, Buehler JA, Eglow M, Kaufman F, et al. The Age-Related Eye Disease Study (AREDS) system for classifying cataracts from photographs: AREDS Report No. 4. Am J Ophthalmol. 2001;131(2):167–75.
13. Maaliw RR, Alon AS, Lagman AC, Garcia MB, Abante MV, Belleza RC, et al.. Cataract Detection and Grading Using Ensemble Neural Networks and Transfer Learning. IEEE Annual Information Technology, Electronics and Mobile Communication Conference. 2022:1-9.
14. Tham YC, Anees A, Zhang L, Goh JHL, Rim TH, Nusinovici S, et al. Referral for disease-related visual impairment using retinal photograph-based deep learning: a proof-of-concept, model development study. Lancet Digit Heal [Internet]. 2021;3(1):e29–40.
15. Jiang L, Huang D, Liu M, Yang W. Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In: 37th International Conference on Machine Learning, ICML 2020. 2020.
16. Yahata E, Winnikow EP, Suyama R, Simoes PW. Explainability in Machine Learning Predictive Models in Breast Cancer. Journal of Health Informatics, 2022;15(Especial).
17. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74.
18. Chauhan K, Kashish, Dagar K, Yadav RK. Cataract detection from eye fundus image using an ensemble of transfer learning models. Intern Conference on Advance Computing and Innovative Technologies in Engineering. 2022:2194-98.
19. Chollet F. Deep Learning with Python Manning. 2018.
20. Gao X, Wong DWK, Ng TT, Cheung CYL, Cheng CY, Wong TY. Automatic grading of cortical and PSC cataracts using retroillumination lens images. Lect Notes Comput Sci . 2013;7725 LNCS(PART 2):256–67.
21. de Araujo AL, Moreira TC, Rados DRV, Gross PB, Bastos CGM, Katz N, et al. The use of telemedicine to support Brazilian primary care physicians in managing eye conditions: The Teleoftalmo project. PLoS One. 2020;15(4):1–12.
22. Carneiro T, Da Nobrega RVM, Nepomuceno T, Bian G Bin, De Albuquerque VHC, Filho PPR. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. IEEE Access. 2018;6:61677–85.
23. Ramlan LA, Zaki WMDW, Mutalib HÁ, Hussain A, Mustapha A. Cataract Detection using Pupil Patch Classification and Ruled-based System in Anterior Segment Photographed Images. Symposium on Computer Applications & Industrial Electronics. 2023: 124-9.
24. Kaur N, Gupta G. Cataract Disease Diagnosis Using SURF Features and Pre-Trained Variants of an EfficientNet Model: Comparative Analysis. Inter Conf Comp, Autom and Know Manag (ICCAKM) 2023:1-6.
25. Shimizu, E, Tanji M, Nakayama S, Ishikawa T, Agata N, Yokoiwa R , Nishimura H, et al. AI-based diagnosis of nuclear cataract from slit-lamp videos. Sci Rep 13, 22046 (2023).