Screening Diabetic Retinopathy Using an Automated Retinal Image Analysis System in Mexico: Independent and Assistive use Cases

Alejandro Noriega, Daniela Meizner, Dalia Camacho, Jennifer Enciso, Hugo Quiroz-Mercado, Virgilio Morales-Canton, Abdullah Almaatouq, Alex Pentland

Submitted to: JMIR Medical Informatics
on: November 18, 2020

Table of Contents

Screening Diabetic Retinopathy Using an Automated Retinal Image Analysis System in Mexico: Independent and Assistive use Cases

Alejandro Noriega[1, 2] PhD; Daniela Meizner[3] MD; Dalia Camacho[2, 4] MSc; Jennifer Enciso[2, 5] PhD; Hugo Quiroz-Mercado[3] MD; Virgilio Morales-Canton[3] MD; Abdullah Almaatouq[6] PhD; Alex Pentland[1] PhD

[1]Massachusetts Institute of Technology, MIT Media Laboratory Cambridge US
[2]Prosperia Salud Mexico City MX
[3]Retina Department, Asociación para Evitar la Ceguera en México Mexico City MX
[4]Engineering Academic Division, Instituto Tecnológico Autónomo de México Mexico City MX
[5]Posgrado de Ciencias Bioquímicas, Universidad Nacional Autónoma de México Mexico City MX
[6]Sloan School of Management, Massachusetts Institute of Technology Cambridge US

Corresponding Author:
Alejandro Noriega PhD
Prosperia Salud
58D Secretaria de Marina 1206, Lomas del Chamizal
Mexico City
MX

Abstract

Background: The automated screening of patients at risk of developing diabetic retinopathy (DR) represents an opportunity to improve their mid-term outcome, and lower the public expenditure associated with direct and indirect costs of common sight-threatening complications of diabetes.

Objective: The present study, aims to develop and evaluate the performance of an automated deep learning–based system to classify retinal fundus images from international and Mexican patients, as referable and non-referable DR cases. In particular, the performance of the automated retina image analysis (ARIA) system is evaluated under an independent scheme (i.e. only ARIA screening) and two assistive schemes (i.e., hybrid ARIA + ophthalmologist screening), using a web-based platform for remote image analysis to determine and compare the sensibility and specificity of the three schemes.

Methods: A randomized controlled experiment was performed where seventeen ophthalmologists were asked to classify a series of retinal fundus images under three different conditions: 1) screening the fundus image by themselves (solo), 2) screening the fundus image after being exposed to the retina image classification of the ARIA system (ARIA answer), and 3) screening the fundus image after being exposed to the classification of the ARIA system, as well as its level of confidence and an attention map highlighting the most important areas of interest in the image according to the ARIA system (ARIA explanation). The ophthalmologists' classification in each condition and the result from the ARIA system were compared against a gold standard generated by consulting and aggregating the opinion of three retina specialists for each fundus image.

Results: The ARIA system was able to classify referable vs. non-referable cases with an area under the Receiver Operating Characteristic curve (AUROC) of 98.0% and a sensitivity and specificity of 95.1% and 91.5% respectively, for international patient-cases; and an AUROC, sensitivity, and specificity of 98.3%, 95.2%, and 90.0% respectively for Mexican patient-cases. The results achieved outperformed the average performance of the seventeen ophthalmologists enrolled in the study. Additionally, the achieved results suggest that the ARIA system can be useful as an assistive tool, as significant sensitivity improvements were observed in the experimental condition where ophthalmologists were exposed to the ARIA's system answer previous to their own classification (93.3%), compared to the sensitivity of the condition where participants assessed the images independently (87.3%).

Conclusions: These results demonstrate that both use cases of the ARIA system, independent and assistive, present a substantial opportunity for Latin American countries like Mexico, towards an efficient expansion of monitoring capacity for the early detection of diabetes-related blindness.

(JMIR Preprints 18/11/2020:25290)
DOI: https://doi.org/10.2196/preprints.25290

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Screening Diabetic Retinopathy Using an Automated Retinal Image Analysis System in Mexico: Independent and Assistive use Cases

Alejandro Noriega[a,b,*], PhD; Daniela Meizner[c], MD; Dalia Camacho[b,d], MSc; Jennifer Enciso[b,e], MSc; Hugo Quiroz-Mercado[c], MD; Virgilio Morales-Canton[c], MD; Abdullah Almaatouq[f], PhD; and Alex Pentland[b], PhD

[a] MIT Media Laboratory, Massachusetts Institute of Technology, Cambridge, USA.
[b] Prosperia Salud, Mexico City, Mexico.
[c] Retina Department, Asociación para Evitar la Ceguera en México. Mexico City, Mexico.
[d] Engineering Academic Division, ITAM, Mexico City, Mexico.
[e] Posgrado de Ciencias Bioquímicas, UNAM, Mexico City, Mexico.
[f] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, USA.

*Corresponding Author: noriega@mit.edu

Abstract

Background: The automated screening of patients at risk of developing diabetic retinopathy (DR) represents an opportunity to improve their mid-term outcome, and lower the public expenditure associated with direct and indirect costs of common sight-threatening complications of diabetes.

Objective: The present study, aims to develop and evaluate the performance of an automated deep learning–based system to classify retinal fundus images from international and Mexican patients, as referable and non-referable DR cases. In particular, the performance of the automated retina image analysis (ARIA) system is evaluated under an independent scheme (i.e. only ARIA screening) and two *assistive* schemes (i.e., hybrid ARIA + ophthalmologist screening), using a web-based platform for remote image analysis to determine and compare the sensibility and specificity of the three schemes.

Methods: A randomized controlled experiment was performed where seventeen ophthalmologists were asked to classify a series of retinal fundus images under three different conditions: 1) screening the fundus image by themselves (*solo*), 2) screening the fundus image after being exposed to the retina image classification of the ARIA system (*ARIA answer*), and 3) screening the fundus image after being exposed to the classification of the ARIA system, as well as its level of confidence and an attention map highlighting the most important areas of interest in the image according to the ARIA system (*ARIA explanation*). The ophthalmologists' classification in each condition and the result from the ARIA system were compared against a *gold standard* generated by consulting and aggregating the opinion of three retina specialists for each fundus image.

Results: The ARIA system was able to classify referable vs. non-referable cases with an area under the Receiver Operating Characteristic curve (AUROC) of 98.0% and a sensitivity and specificity of 95.1% and 91.5% respectively, for international patient-cases; and an AUROC, sensitivity, and specificity of 98.3%, 95.2%, and 90.0% respectively for Mexican patient-cases. The results achieved outperformed the average performance of the seventeen ophthalmologists enrolled in the study. Additionally, the achieved results suggest that the ARIA system can be useful as an assistive tool, as significant sensitivity improvements were observed in the experimental condition where ophthalmologists were exposed to the ARIA's system answer previous to their own classification (93.3%), compared to the sensitivity of the condition where participants assessed the images independently (87.3%).

Conclusions: These results demonstrate that both use cases of the ARIA system, *independent* and *assistive*, present a substantial opportunity for Latin American countries like Mexico, towards an efficient expansion of monitoring capacity for the early detection of diabetes-related blindness.

Keywords: Diabetic retinopathy; automated diagnosis; retina; fundus image analysis.

Introduction

Diabetes is one of the most challenging health problems in the world affecting more than 400 million people. Particularly, diabetes threatens the health care systems of low- and middle-income countries where 80% of the world's diabetic population live [1,2]. Diabetes is a multifactorial and complex disease with a strong genetic component. In this regard, it has been demonstrated that Hispanic/Latinos have a greater susceptibility to develop type II diabetes (T2D), as well as diabetes-associated complications including renal insufficiency and visual impairment [1–4].

In 2015, there were more than 41 million adults diagnosed with diabetes in Latin America and Caribbean (LAC) countries, making it one of the major causes of premature death and disability in the region [5,6]. Particularly, Mexico ranked sixth among the world´s diabetes prevalence in 2015, and the 2nd in Latinamerica, only after Brasil [7,8]. It is estimated that 26 million adults live in Mexico with diabetes and pre-diabetes, and only half of them have been diagnosed. Diabetes and its related complications are the first cause of disability and the 3rd cause of death in the country, having a great impact in productivity, life quality and economy [5].

Evolution and treatment of Diabetic Retinopathy

Diabetic retinopathy (DR) is the most common complication in advanced and/or uncontrolled diabetic patients and is the leading cause of irreversible vision loss in working-age adults [9,10]. DR is a microvascular complication that emerges in diabetic patients as a consequence of chronic hyperglycemia that contributes to blood vessels damage in the retina causing fluid leakage, swelling of the surrounding tissue, blood flow obstruction and/or abnormal neovascularization [9,10].

DR progression is slow and gradual, and reversible in its first stage, however if not treated promptly, evolves to irreversible blindness. According to the International Clinical Diabetic Retinopathy Severity Scale, the first stage of DR is classified as mild non-proliferative DR (NPDR), it is characterized by the presence of at least one microaneurysm and is highly reversible through blood pressure, cholesterol and sugar levels control. Only very rare cases which present macular edema (swelling of fluid and protein deposits on or under the macula) might require laser photocoagulation or intravitreal injections. Without adequate diabetes control, the disease advances to moderate and severe NPDR stages,which include the presence of hemorrhages, microaneurysms, hard exudates, venous beading and/or intraretinal microvascular abnormalities. At these stages, metabolic control is not sufficient to stop the disease progression and the patient will require invasive treatments like photocoagulation and intravitreal anti-vascular endothelial growth factor (VEGF) agents or corticosteroids. The most advanced stage is proliferative DR (PDR) and is characterized by neovascularization, preretinal hemorrhages, hemorrhage into the vitreous, traction retinal detachments, or macular edema (ME). PDR is treated with a more aggressive laser therapy called scatter or pan-retinal photocoagulation, intravitreal injections, and in some cases, vitreoretinal surgery to remove scar tissue and/or blood from the vitreous cavity, for laser repair of retinal detachments and treatment of macular holes. [10-13]

To increase early detection and prevent the progression of DR to advanced stages, diabetic patients are recommended to have annual or semi-annual retinal screenings beginning at the moment when they are diagnosed with diabetes. However, according to data from the Diabetic Retinopathy Barometer, 27% of people living with diabetes declared that they never discussed eye complications with their doctors before the onset of complications and only 13% of the diabetic population have visited an ophthalmologist after their diagnosis [4,14]. Through frequent, preventive screenings, 70% of the cases can be captured at the initial stages of the disease and treated with non-invasive strategies, like metabolic control or photocoagulation [15]. Unfortunately, in most developing countries there is no ophthalmological attention at primary care clinics, and it is only until diabetic patients develop vision attenuation that they are referred to second and third level hospitals to be screened, diagnosed and treated [16]. At this point, significant retinal damage has occurred and even with invasive vitreoretinal surgery or photocoagulation, vision can't be restored.

The limited access to ophthalmologists and retina specialists at primary care clinics, due to financial and staff limitations at national healthcare institutions, precludes the continuous monitoring of diabetic patients in low and middle income countries like Mexico.

Challenges of DR screening at a large scale

In Mexico, DR is a leading cause of irreversible blindness among the working-age population [4,13]. Around 30% of the patients diagnosed with diabetes develop DR, and based on the predictions of diabetes increasing prevalence, by 2045 there will be 245 million people with DR lesions and 77 millions with vision-threatening DR [17].

One of the main limitations for the establishment of a systematic eye-screening program is the limited availability of ophthalmologists and their unequal distribution around the country. Based on the 2013 registry of society-affiliated ophthalmologists from the Mexican Society of Ophthalmology, the average number of ophthalmologists per 100 000 persons is lower than the regional average (5.27 per 100,000), showing a particularly worrying distribution in rural areas where there were 2 ophthalmologists per 100,000 persons [18].

Automated retinal image analysis (ARIA) for DR screening

In recent years, the juncture between the development of advanced statistical methods, the greater availability of data, and the substantial increase in computing power, has allowed the application of advanced computational methodologies, including artificial intelligence (AI), in diverse social and medical domains. Among the use of AI for social welfare, AI applications in healthcare domains is one of the fastest-growing sectors, with a compound annual growth rate between 2014 and 2021 above 40%[19]. AI tools have been successfully applied in diagnostics, therapeutics, population health management, administration, and regulation, probing their capacity to augment societies' ability to increase access to healthcare, and improve the coverage and quality of the services provided.

Ultimately, AI applications in healthcare present opportunities to improve overall quality of life, patients' prognosis, and optimize human and financial resources [20]. In particular, automated retinal image analysis (ARIA) systems have emerged as a promising solution to massify early detection of DR at primary care clinics, particularly in resource-constrained developing countries, thereby improving health outcomes, avoiding incapacitating complications and reducing treatment costs.

ARIA systems analyze retinal fundus images by applying techniques like deep learning (DL) to classify diabetic patients in a) cases without retinal lesions associated to DR (*non-referable* output) and b) cases that need to undergo examination by an ophthalmologist to confirm diagnosis and define treatment (*referable* output) [21–25]. As of today, various analysis systems have been developed and implemented on the market in European countries, Canada and the United States. Though, very few have been tested in LAC to evaluate their performance and usability in the particular resource-constrained settings of these countries [26], including patients ethnicity, training of the healthcare personnel, community openness to new technologies, and hospital resources as players for their successful implementation .

Aims and key findings of the study

The present work evaluates the performance of a DL-based ARIA system that classifies retinal fundus images in non-referable or referable based on the presence of DR damage, as well as the potential benefits of its use as an assistive tool for ophthalmic doctors. It also reports the results on a randomized controlled trial where the performance of the ARIA system was compared to the accuracy of seventeen ophthalmologists of one of the most reputed ophthalmic hospitals in Mexico,

"Hospital de la Ceguera" from the "Association to avoid blindness in Mexico" (APEC). In particular, the performance of ophthalmologists in three experimental conditions was assessed: one *independent* condition, where the ophthalmologists assess the images independently from the ARIA system, and two *assistive* conditions, in which ophthalmologists can observe and be influenced by the ARIA system's classification and confidence, or an ARIA system-generated attention heatmap highlighting probable DR lesions in the retina.

Key findings. The ARIA system developed using DL strategy was able to classify referable vs. non-referable cases with an area under the Receiver Operating Characteristic curve (AUROC), sensitivity, and specificity of 98.0%, 95.1% and 91.5% respectively, for international patient-cases; and an AUROC, sensitivity, and specificity of 98.3%, 95.2%, 90.0% respectively for Mexican patient-cases. The results achieved on Mexican patient-cases outperformed the average performance of the seventeen ophthalmologist participants of the study. Moreover, we find that the ARIA system can be useful as an assistive tool, as we found significant improvement in the specificity in the experimental condition where participants were able to consider the answer of the *ARIA system* as a second opinion (87.3%), compared to the specificity of the condition where participants assessed the images independently (93.3%).

Hence, the present study demonstrates a high potential value of the use of ARIA systems, in both independent and assistive schemes, towards effectively massifying the early detection of DR in developing countries like Mexico.

Methods

ARIA system design

The ARIA system consists of an image preprocessing module, and an image analysis module that returns a binary referable and non-referable DR classification, the level of confidence of that classification, and an attention map that shows, pixel-wise, the indicative features for referable DR according to the model (see Figure 1). The models constituting the ARIA system were implemented using the Keras library with the Tensorflow backend [27] in Python 3.5 [28].

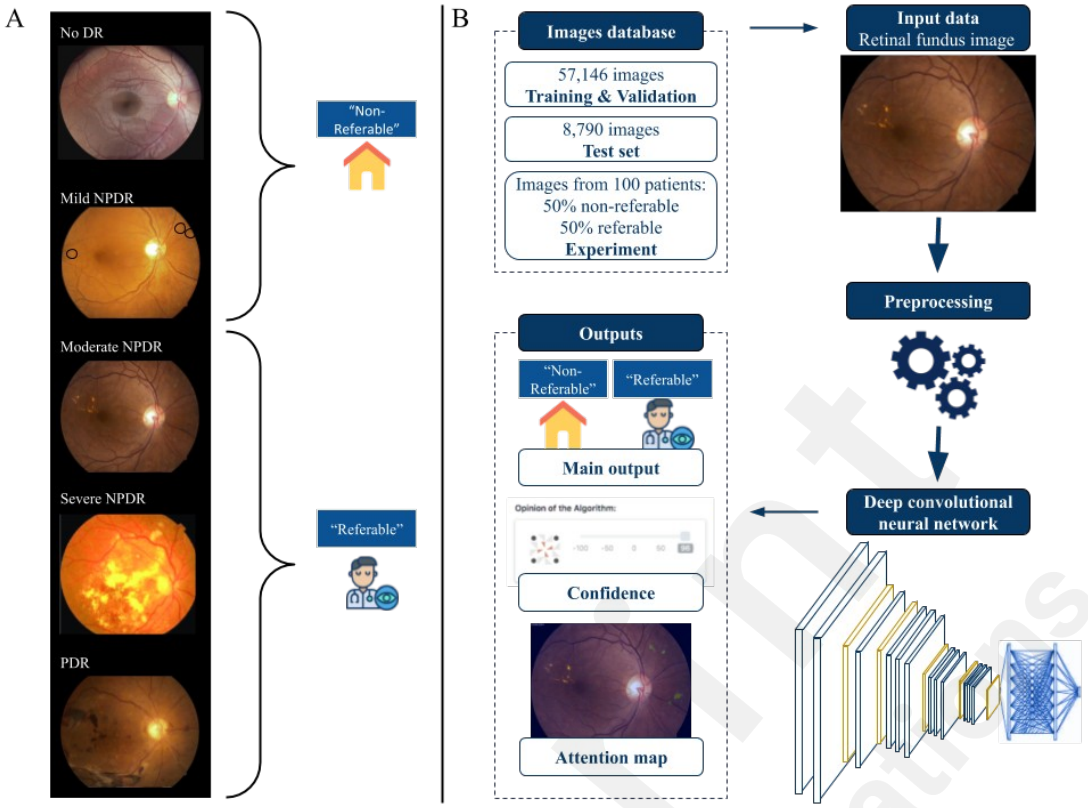

Figure 1. DL-based ARIA system. A) Example of classified retinal fundus images according to the DR severity scale used for the training data. B) Flow chart describing the design of the ARIA system; the data used for training, validation and test; and the algorithm's outputs.

Images from all datasets were annotated by ophthalmic specialists for 5-class identification according to the International Clinical Diabetic Retinopathy Severity Scales, and subsequently labeled as non-referable and referable DR [30]. Table 1 describes the classification, and Figure 1A provides a graphical example. The *gold standard* classification used for the experimental phase of the study was provided by three retina specialists, as described in the following subsections.

Table 1. International Clinical Diabetic Retinopathy Severity Scale and their classification for the ARIA system [29].

ARIA system classification	DR severity scale	Ophthalmoscopy findings
Non referable	No apparent retinopathy (No DR)	No abnormalities
	Mild non proliferative diabetic retinopathy (Mild DR)	Microaneurysms only
Referable	Moderate non proliferative diabetic retinopathy (Moderate DR)	More than just microaneurysms, but less than severe nonproliferative diabetic retinopathy
	Severe non proliferative diabetic retinopathy (Severe DR)	More than 20 intraretinal hemorrhages in each of four quadrants, definite venous beading in two quadrants, and/or prominent intraretinal microvascular abnormalities in one quadrant. No signs of proliferative retinopathy.
	Proliferative diabetic retinopathy (Proliferative DR or PDR)	Neovascularization and/or vitreous/preretinal hemorrhage.

Preprocessing. Before classifying the images and training the algorithms, a preprocessing procedure was applied. The procedure consisted of cropping the background to eliminate non-informative areas, padding to guarantee consistent squared image ratios, resizing the image to 224x224 pixels, and normalizing pixel values to the range 0 to 1.

Image classification. The image classification model used was a deep convolutional neural network [30,31], trained on a dataset of 57,146 images and evaluated out of sample on a dataset of 8,790 images, all from international cases [32]. The network architecture consisted of 16 convolutional layers, a dense layer of 1,024 neurons, two dropout layers to avoid overfitting, and a binary classification layer of a single unit with sigmoid activation. Hence, the model output is a value between 0 and 1, that may be interpreted as the confidence of the model regarding a referable DR classification. Lastly, a threshold of 0.5 was used to classify non-referable and referable DR.

Attention heatmaps. Attention heatmaps were developed to show lesion areas in the image by highlighting each pixel according to their importance towards a referable DR classification according to the model. These heatmaps were obtained by applying one of the most effective methods for building saliency maps on images, the layer-wise relevance propagation (LRP) method, with an alpha-beta rule [33,34]. In essence, the LRP method redistributes the output value throughout the layers until the input layer (input image) is reached. Figure 2 shows a couple of examples of fundus images and the heatmaps generated using the methodology described.

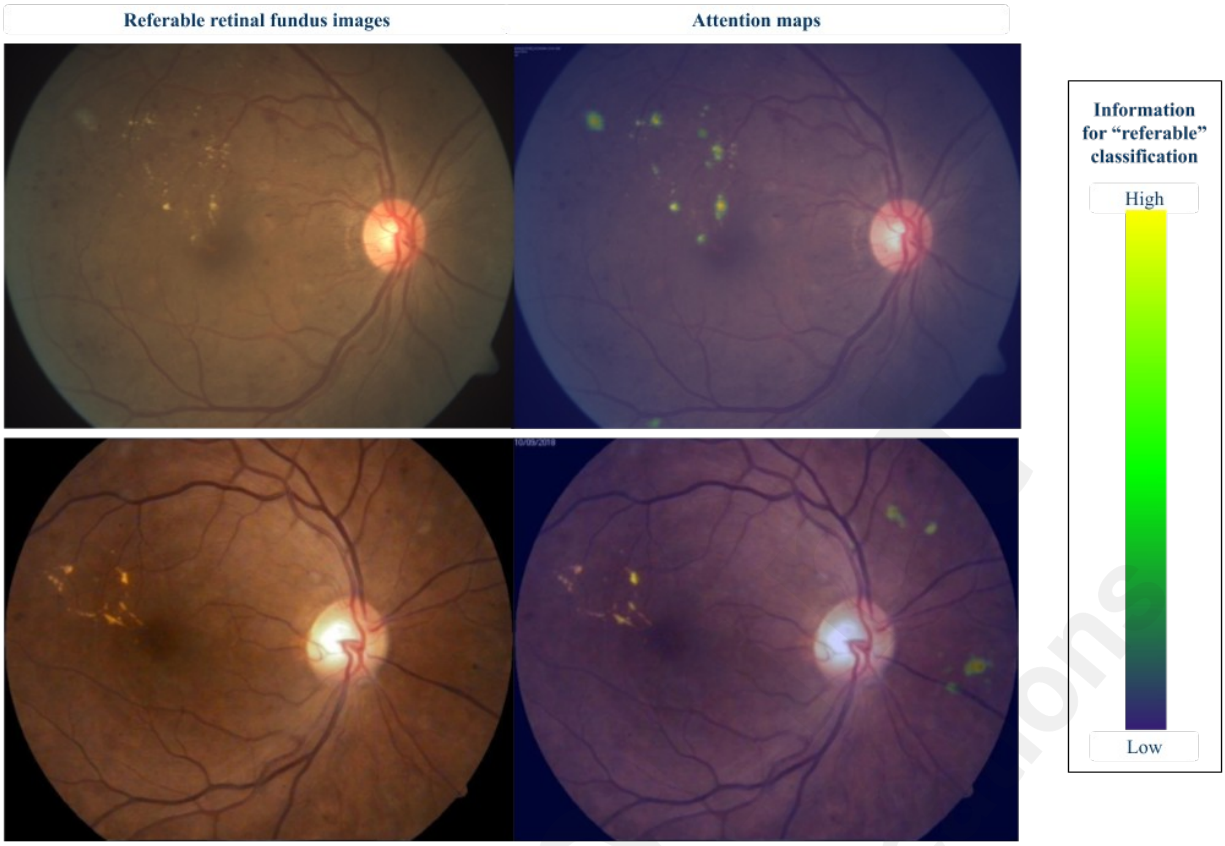

Figure 2. Attention heatmaps for two *referable* images. Green and yellow colors indicate regions in the image that provide information to the algorithm to classify the image as *referable.*

Study populations

Seventeen ophthalmologists from the Mexican ophthalmic hospital participated in the experimental study, and three retina specialists from the same institution participated in the generation of the *gold standard*. The seventeen ophthalmologists evaluated fundus images from 100 Mexican patients, where 50% had non-referable DR and 50% had referable DR levels. Each ophthalmologist evaluated 45 retinal images, in order that each image was evaluated more than once. The ophthalmologists were retina specialization resident students, with the following distribution: 3 in their second year, 13 in their third year and two in their fourth year of residency.

Experimental design

We conducted a randomized controlled experiment to assess the performance of the ARIA system in comparison with ophthalmic doctors from the Mexican ophthalmic hospital, and to evaluate the potential benefits of using the system as an assistive tool for doctors. To achieve this a web-based experiment platform was developed where ophthalmologists evaluated fundus retinal images under three different conditions—*solo*, *ARIA answer*, and *ARIA explanation*—described below. Figure 3 displays the main screens of the web platform used in this experiment

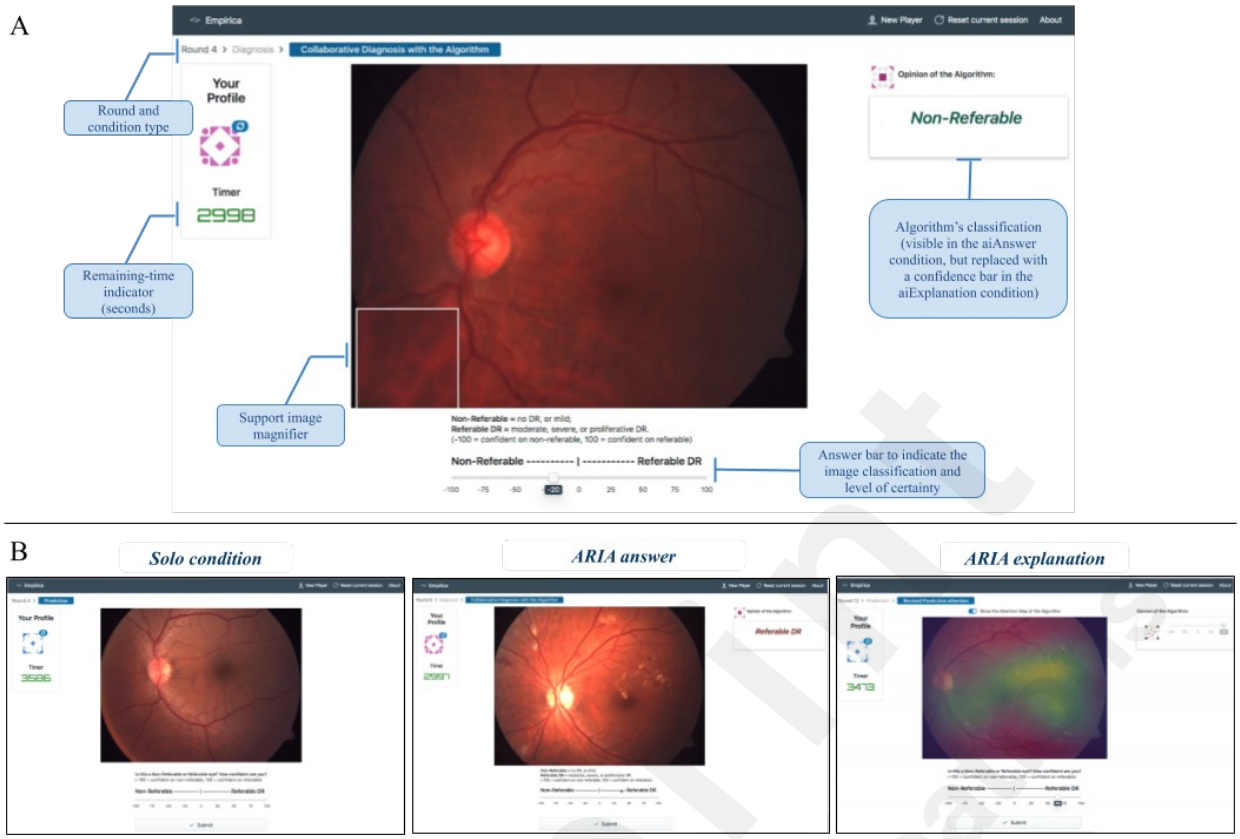

Figure 3. Web-platform design for patient-case classification. A) Visual indicators and components of the classification window, B) visualization of the three experimental conditions.

Gold standard and image quality. In order to generate a *gold standard* the fundus images of all patient-cases used in the experiment were graded by three retina specialists of the ophthalmic hospital, and a majority rule was used, i.e. if there was a disagreement in the non-referable/referable label, the label selected by two out of three experts was considered the *gold standard*. Image grading was done through the same web-based platform described in figure 3. The retina specialists also graded the image quality, and images graded as having bad quality were not considered for the experiment. From the remaining, images from 50 patients with referable DR, and 50 with non-referable DR, were selected at random to be used for the study.

Experimental conditions. The experiment followed a within-subjects design, where each ophthalmologist evaluated 45 randomly selected fundus images (from 45 different patients), 15 for each of the three treatment conditions: *solo*, *ARIA answer* and *ARIA explanation*. The ophthalmologists were first asked to evaluate 15 fundus retinal images in the *solo* condition, followed by 30 images that randomly alternate between the *ARIA answer* and the *ARIA explanation* conditions. In the *solo* condition, participants responded to the task in isolation, without any exposure to the ARIA system. In contrast, in the *ARIA answer* condition, participants were exposed to the binary answer of the ARIA system (i.e., non-referable or referable), as a second opinion, and then asked to submit their post-exposure answer. The *ARIA explanation* condition was identical to the *ARIA answer* condition, with the exception that participants were shown not only the binary answer of the ARIA system, but also its level of confidence and attention heatmap.

Finally, after completing all the classification tasks, the ophthalmologists were asked to submit an optional feedback survey about their experience.

The study was reviewed and approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology, and all participants provided explicit consent prior to their participation.

Results

ARIA performance

The ARIA system was first tested in the large dataset of international cases. It there achieved an out-of-sample area under the Receiver Operating Characteristic curve (AUROC) of 98.0%. In particular, using a given acceptance threshold, the ARIA system achieved a sensitivity of 95.1% and a specificity of 91.5%. Most importantly, the ARIA system also displayed high accuracy classifying images from patients from the Mexican ophthalmic hospital, where it had an AUROC of 98.3%, a sensitivity of 95.2%, and specificity of 90.0% (see Figure 4).

Participants' performance

Figure 4 shows the sensitivity and false positive rate (*1 - specificity*) for each condition—*solo*, *ARIA answer*, and *ARIA explanation*—and compares them with the ROC curve of the ARIA system. The average sensitivity in the *solo* condition across the seventeen participants was 87.3%, and the average specificity was 86.8%. In comparison, the average sensitivity and specificity across the seventeen participants for the *ARIA Answer* condition were 93.3% and 89.3%, and the average sensitivity and specificity across participants for the *ARIA Explanation* condition were 91.5% and 79.0%. The joint analysis of the ARIA system performance on mexican patients vs. the three experimental conditions involving ophthalmologists assessment, show that the ARIA system is more accurate than the average of participants under any of the exposure conditions. In particular, the ARIA system increased sensitivity from 87.3% to 93.3% (vertical movement between the dark blue dot and the green line in Figure 4), while maintaining participants' specificity of 86.8% constant; or increase specificity to 100% while maintaining participants' average sensitivity of 87.3% constant (horizontal movement from the dark blue dot leftwards to the green line in Figure 4) against the *solo* condition.

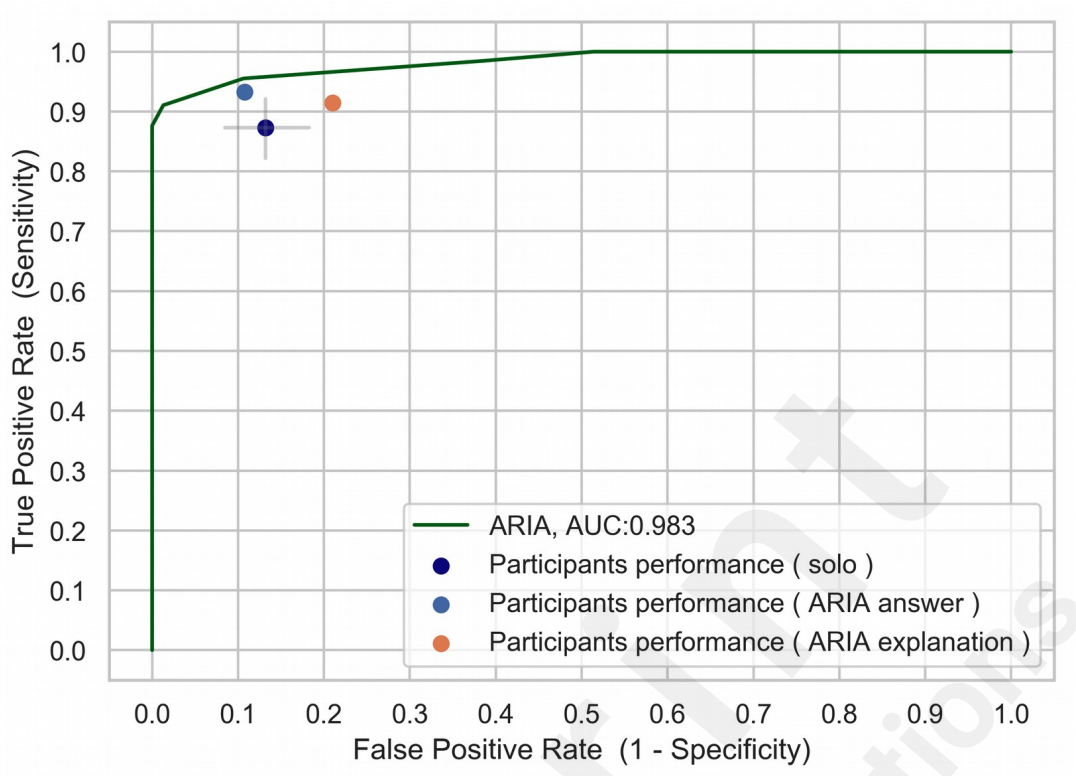

Figure 4. ROC curve of the ARIA system compared to the ophthalmologist's accuracy under the three experimental conditions (*solo*, *ARIA answer* and *ARIA explanation*). Grey lines indicate 95% confidence intervals for the *solo* condition.

Most interestingly, Figure 4 shows that exposure to the ARIA system was able to improve the performance of human experts, particularly in the *ARIA answer* condition, which significantly improved the sensitivity and specificity compared to the *solo* condition (distance between dark blue and light blue dots in Figure 4). However, performance in the *ARIA explanation* condition had mixed results, improving sensitivity, but lowering the specificity (distance between dark blue and orange dots in Figure 4).

Figure 5 provides more detail on the effect that exposure to information of the ARIA system had on the performance of ophthalmologists. In particular it shows the accuracy (% of correct answers) of the seventeen experts consistently improved in the *ARIA answer* condition, shifting the distribution upwards, and decreasing the variance across participants. For example, while only two participants had a perfect score in the *solo* condition, up to 6 participants had a perfect score in the *ARIA answer* condition. However, the *ARIA explanation* condition had mixed beneficial and detrimental effects on participants' accuracy, and increased the variance of performance across participants compared to the *solo* condition.

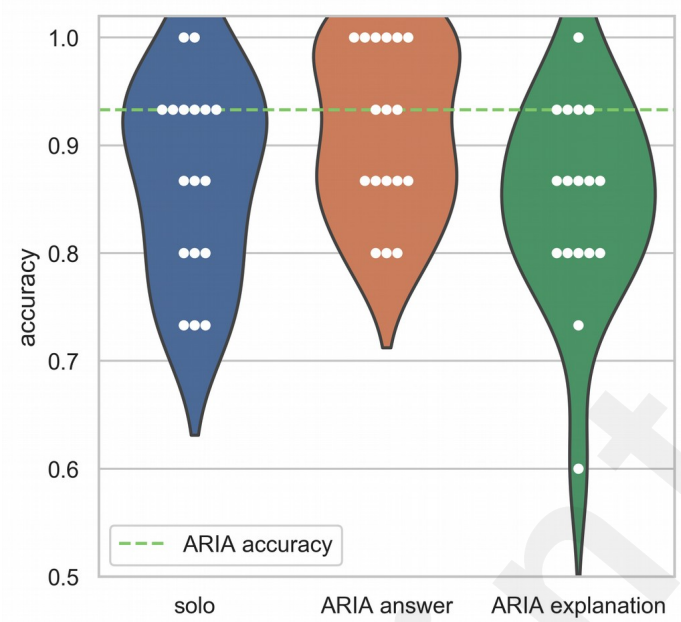

Figure 5. Influence of the ARIA system on the ophthalmologists decision. A) Ophthalmologists performance and B) opinion shift after exposure to the AI outputs.

Discussion

The number of people living with diabetes is projected to rise by 50% in 2045, reaching a 700 million people worldwide [7,35]. Considering the current prevalence of DR in diabetic population (36%), this translates into 230 million patients developing DR which will require routinary eye screenings to be treated on time and prevent vision loss. The development or ARIA systems represent a possible solution to the increasing demand of eye screenings, offering a scalable and cost-effective alternative to healthcare systems, which will be highly valuable particularly in limited-resource settings. Just in Mexico, the prevention of DR in patients living with diabetes would implicate savings for up to $10 million dollars to the three main public institutions of the national health care system [36].

Principal results

The DL-based ARIA system presented in this work was evaluated with a subset of retinal images from international patient-cases and an image set of patients from a Mexican ophthalmic hospital. In both datasets, the ARIA system outperformed the average sensitivity and specificity of seventeen ophthalmology residents of retina specialty. The reached sensitivities (95.1% and 95.2% for the international and mexican datasets, respectively) are comparable to those reported for other seven automated DR screening systems reported in a systematic review, whose sensitivity values were between 87% and 95% [36]. On the other hand, the specificities reached by our ARIA system (91.5% and 90% for the international and mexican datasets, respectively) were higher than the reported average in [36], whose specificity values were between 49% and 69%.

Nowadays, many efforts demonstrating good performance of AI-based ARIA systems for the detection of DR have been reported, however sensitivity and specificity are not the only parameters that should be considered to guarantee a successful real-world implementation [37,38]. Therefore, on the presented work, the ARIA system evaluation included two assistive or human-AI hybrid decision schemes. This experimental design was developed to reflect that in real-world applications, results of

an automated system are reviewed and confirmed by healthcare professionals to choose the most adequate therapeutic protocol for each patient. In these assistive evaluations, the existence of significant synergies derived from the interaction of the human and AI dyads was confirmed. AI's output exerted a strong influence on the opinion of human participants, however its effect on ophthalmologist's overall precision depended on the format of ARIA system's output. A simplified output (i.e. non-referable or referable classification) resulted in the most positive input for humans' sensitivity and specificity. On the other hand a more complex output (i.e. confidence bar and attention map) partially improved human's decisions, increasing their sensitivity but also increased the incidence of false positive classifications. These results are coherent with some of the ophthalmologist's feedback submitted after the classification tasks, where some of them expressed that even when attention heatmaps were useful, the bar showing the confidence of the ARIA system was confusing.

Limitations

Further pilot studies with a larger number of patients and ophthalmologists will be useful to confirm the ARIA system accuracy. Also, further studies might include direct ophthalmoscopy by retina specialists as the *gold standard*, in order to avoid error related to image quality.

Additional experiments with alternative platform designs might be useful to generate a suitable screening tool that optimizes patient evaluation and referral in three stages. In the first stage, an ARIA system might be useful to identify patients with a higher probability of developing DR. In a second filter, ophthalmologists would be able to evaluate the retinal images of high-risk patients, in combination with the ARIA system output to make a first decision about the disease stage, treatment and finally, refer only patients with an advanced disease to retina specialists.

Conclusions

The results of the present study demonstrates a substantial opportunity for Latin American countries, like Mexico towards an efficient massification of monitoring systems for early detection of diabetes-related blindness, considering the short supply of ophthalmologists in their public healthcare system.

The web-based platform developed for this study was designed for the implementation of the ARIA system as an automatic screening tool and as a telemedicine platform useful to confirm or reject the ARIA system's output with assessment of an ophthalmologist or retina specialist. The platform was useful for the present study, and can be easily adapted for further studies that include the recopilation of additional information about other eye diseases detectable by image analysis (i.e. glaucoma, age-related macular degeneration or coat disease).

The conclusion of these results suggest the proposed ARIA system is valuable in an independent and assistive condition, and can be useful to massify and improve DR diagnosis, as well as other ophthalmic diseases in the long-term. However special attention in the design of a careful and explanatory platform, is required for a successful deployment.

Acknowledgments

The authors gratefully thank the retina specialists and the ophthalmologists from the APEC hospital involved in this study for their evaluations of the retina fundus images.

Authors' Contributions

AN conceived and designed the experiments, analyzed data and contributed to the discussion and review of the paper. DC was responsible for the models' training, performed the experiments, analyzed data and contributed to the discussion of the paper. DM contributed with the experimental design, image classification and discussion of the paper. JE contributed with data analysis, paper writing and discussion. HQM, VMC, AA and AP contributed to various aspects of the paper including experimental design, machine learning strategies, medical feedback, image evaluations and paper discussion.

Funding Statement

This project was carried out thanks to the fellowships received by individual members of the team, including fellowships of the Massachusetts Institute of Technology and the National Council of Science and Technology (CONACYT).

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

AI: artificial intelligence
ARIA: automated retinal image analysis
AUROC: area under the Receiver Operating Characteristic curve
DL: deep learning
DR: diabetic retinopathy
FPR: false positive rate
LRP: layer-wise relevance propagation
ROC: receiver operating characteristic
TPR: true positive rate

References

1. Zhang X, Saaddine JB, Chou C-F, Cotch MF, Cheng YJ, Geiss LS, et al. Prevalence of Diabetic Retinopathy in the United States, 2005-2008. JAMA 2010;304(6):649. DOI: 10.1001/jama.2010.1111
2. Williams A, Jacobs S, Moreno-Macías H, Huerta-Chagoya A, Churchhouse C, Márquez-Luna C, et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature 2014;506(7486):97–101. DOI: 10.1038/nature12828
3. Caballero AE. Understanding the Hispanic/Latino Patient. Am J Med 2011;124(10):S10–5. DOI: 10.1016/j.amjmed.2011.07.018
4. National Institute of Public Health. 2016. Mexico National Survey of Health and Nutrition Midway 2016 (ENSANUT MC 2016) Final Report. https://www.gob.mx/cms/uploads/attachment/file/209093/ENSANUT.pdf
5. (IHME) Institute for Health Metrics and Evaluation. 2017. Country profiles. Technical report. http://www.healthdata.org/results/country-profiles
6. Barcelo A, Arredondo A, Gordillo-Tobar A, Segovia J, Qiang A. The cost of diabetes in Latin America and the Caribbean in 2015: Evidence for decision and policy makers. J Glob Health 2017;7(2):020410. DOI: 10.7189/jogh.07.020410
7. International Diabetes Federation. 2019. IDF DIabetes Atlas Ninth edition 2019. https://diabetesatlas.org/upload/resources/material/20200302_133352_2406-IDF-ATLAS-SPAN-BOOK.pdf
8. Gómez, E. Political party ambitions and type-2 diabetes policy in Brazil and Mexico. Health Econ Policy Law 2020;15(2):261-276. DOI:10.1017/S1744133118000415
9. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. Phys Ther 2008;88(11):1254-64. DOI: 10.2522/ptj.20080020.
10. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. Lancet. 2010;376(9735):124–36. doi: 10.1016/S0140-6736(09)62124-3.
11. Secretaría de Salud. 2015. Guía de Práctica Clínica, Diagnóstico y Tratamiento de Retinopatía Diabética. http://www.cenetec.salud.gob.mx/descargas/gpc/CatalogoMaestro/171_GPC_RETINOPATIA_DIABETICA/Imss_171RR.pdf
12. Claramunt J. Diabetic retinopathy from prevention. Embedding screening into diabetes centres. RMCLC 2016;27(2):195-203. DOI: 10.1016/j.rmclc.2016.04.009
13. Barría von-Bischhoffshausen F, Martínez Castro F. 2011. Clinical Practice Guide for Diabetic Retinopathy for Latin America for Ophthalmologists and Healthcare Professionals. https://www.iapb.org/wp-content/uploads/2011-Clinical-Practice-Guide-for-DR-for-Latin-America.pdf
14. Cavan D, Makaroff L, da Rocha Fernandes J, Sylvanowicz M, Ackland P, Conlon J, et al. The Diabetic Retinopathy Barometer Study: Global perspectives on access to and experiences of diabetic retinopathy screening and treatment. Diabetes Res Clin Pract 2017;129:16-24. DOI:10.1016/j.diabres.2017.03.023
15. Jimenez-Baez MV, Marquez-Gonzalez H, Barcenas-Contreras R, Morales Montoya C, Espinosa-Garcia LF. Early diagnosis of diabetic retinopathy in primary care. Colomb medica 2015;46(1):14–8. DOI: https://doi.org/10.25100/cm.v46i1.1681
16. Carrillo-Alarcón LC, Ávila-Pozos R, López López E, Cruz-Castillo R, Ocampo-Torres Moisés, Alcalde-Rabanal JE. Projection of Diabetic Patients Retinopathy in Hidalgo State-México, through 2030. EC Ophthalmology 2017;5(2): 73-80
17. Thomas RL, Halim S, Gurudas S, Sivaprasad S, Owens DR. IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy

between 2015 and 2018. Diabetes Res Clin Pract. 2019;157:107840. doi:10.1016/j.diabres.2019.107840
18. Hong H, Mújica OJ, Anaya J, et al. The Challenge of Universal Eye Health in Latin America: distributive inequality of ophthalmologists in 14 countries. BMJ Open 2016;18,6(11):e012819. DOI: 10.1136/bmjopen-2016-012819. Erratum in: BMJ Open. 2016; 30,6(12):12819corr1.
19. Frost & Sullivan. 2016. From $600 M to $6 Billion, Artificial Intelligence Systems Poised for Dramatic Market Expansion in Healthcare. https://ww2.frost.com/news/press-releases/600-m-6-billion-artificial-intelligence-systems-poised-dramatic-market-expansion-healthcare/
20. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25(1):30. DOI:10.1038/s41591-018-0307-0.
21. Arenas-Cavalli JT, Ríos SA, Pola M, Donoso R. A Web-based Platform for Automated Diabetic Retinopathy Screening. Procedia Comput Sci 2015;60:557–63. DOI: https://doi.org/10.1016/j.procs.2015.08.179
22. Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins D, Duchesne S, editors. Lesion detection and Grading of Diabetic Retinopathy via Two-stages Deep Convolutional Neural Networks. Proceedings of the 20th Medical Image Computing and Computer Assisted Intervention; 2017 September 11-13. Quebec City, Canada. Switzerland: Springer Nature; 2017.
23. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda S, et al. Automated Diabetic Retinopathy Screening and Monitoring Using Retinal Fundus Image Analysis. J Diabetes Sci Technol 2016;10(2):254–61. DOI: 10.1177/1932296816628546
24. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316(22):2402. DOI:10.1001/jama.2016.17216
25. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, Owen CG, et al. Automated Diabetic Retinopathy Image Assessment Software. Ophthalmology 2017;124(3):343–51. DOI:10.1016/j.ophtha.2016.11.014
26. Dutz MA, Almeida RK, Packard TG. The Jobs of Tomorrow: Technology, Productivity and Prosperity in Latin America and the Caribbean. Washington DC, World Bank Group; 2018. ISBN: 978-1-4648-1222-4
27. Chollet F, and others. 2015. Keras. https://keras.io
28. Van Rossum G, Drake Jr FL. 2001. Python Tutorial Release 2.0.1. PythonLabs. https://docs.python.org/2.0/tut/tut.html
29. Wilkinson C, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110(9):1677–82. DOI: 10.1016/S0161-6420(03)00475-5
30. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Back-propagation applied to handwritten zip code recognition. Neural Computation 1989;1(4):541–551.
31. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv 2015:1409.1556.
32. Voets M, Møllersen K, Bongo LA. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PLOS ONE 2019;14(6):e0217541. https://doi.org/10.1371/journal.pone.0217541
33. Wojciech Samek, Gregoire Montavon, Alexander Binder, Sebastian La-puschkin, and Klaus-Robert Muller. Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation. ArXiv 2016:1611.08191.
34. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS One 2015;10(7):e0130140. DOI:10.1371/journal.pone.0130140

35. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwinet N, al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Res Clin Pract 2019;157:107843. DOI:10.1016/j.diabres.2019.107843
36. Barquera S, Campos-Nonato I, Aguilar-Salinas C, Lopez-Ridaura R, Arredondo A, Rivera-Dommarco J. Diabetes in Mexico: cost and management of diabetes and its complications and challenges for health policy. Global Health 2013;9:3. DOI:10.1186/1744-8603-9-3
37. Nørgaard MF, Grauslund J. Automated Screening for Diabetic Retinopathy - A Systematic Review. Ophthalmic Res. 2018;60(1):9-17. DOI:10.1159/000486284
38. Gulshan V, Rajan R, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol 2019;137(9):987-993. DOI:10.1001/jamaophthalmol.2019.2004