International Journal of Combinatorial Optimization Problems and Informatics, 12(3), Sep-Dec 2021, 79-86. ISSN: 2007-1558.

www.editada.org

Use of artificial intelligence to evaluate the detection of retinal alterations as a screening test in Mexican patients

Moises Argueta-Santillan[1], E. Mahuina Campos-Castolo[1], Miguel Ángel Méndez-Lucero[1], Dania N. Lima-Sánchez[1], Josué Fabricio Urbina-González[2], Orlando Cerón-Solís[1], Alejandro Alayola-Sansores[1], German Fajardo-Dolci[3]*

[1] Departamento de Informática Biomédica, Facultad de Medicina, Univ. Nacional Autónoma de México, México.

[2] Facultad de Ingeniería, Univ. Nacional Autónoma de México, México

[3] Director de la Facultad de Medicina, Univ. Nacional Autónoma de México, México

el_moi@ciencias.unam.mx, dibfm@unam.mx, mendezluceromiguelangel@gmail.com, *danianimbe@gmail.com, urbgon@gmail.com, orlandoceronsolis@gmail.com, ale.alayola@gmail.com, direccionfm@unam.mx

Abstract. In Mexico, chronic degenerative diseases are the leading cause of morbidity, which has frequent retina complications, being the leading cause of blindness in our population. Unfortunately, the detection of the pathology is usually late, resulting in more significant disability. To propose the detection of different pathologies with different artificial intelligence algorithms have been used for the images taken from the fundus of the eye. Objective. Evaluate different machine learning algorithms for the detection of retinal alterations in the Mexican population. Methodology. Evaluate two types of models to estimate artificial intelligence tools' screening capacity, one based on transfer learning and ensemble methods against one based only on convolutional networks. Results. We obtained good values to differentiate between healthy and sick but not to diagnose different pathologies. Conclusions. It is necessary to enlarge the imaging sample and to improve the screening models.

Keywords: Retinal image, Transfer learning, ocular disease classifier, ensemble methods, deep learning, convolutional networks

Article Info
Received Jan 18, 2021
Accepted April 17, 2021

1 Introduction

Vision pathologies, especially those related to the retina, are widespread among the Mexican population, primarily associated with other pathologies such as diabetes and hypertension [1]. The inside of the eye is a susceptible area, pathologies in the eye's fundus are sometimes associated with systemic diseases such as diabetes or cancer [2]. Diagnosis is not always available due to the enormous demand and shortage of specialists in the field. These pathologies are more common in the presence of comorbidity and aging, something that occurs predominantly in the Mexican population [3]. For example, in Diabetes Mellitus, it is estimated that approximately 35% will have diabetic retinopathy alterations, and 10% will have their vision threatened by this cause [4].

Early detection is a method that reduces treatment costs and, in general, improves the quality of life. Therefore, retinal images taken with a non-mydriatic camera would be a breakthrough in this regard. It is possible to take fundus images with non-mydriatic cameras, even with smartphones, with good sensitivity and specificity [5]. These images can be analyzed using artificial intelligence to detect different alterations [6]. The development of artificial intelligence (AI) algorithms to solve medical problems has been one of the most promising areas, mainly to perform patients' mass screening. Machine learning is an essential branch of AI, which can be divided into supervised or unsupervised when images are previously labelled [7]. Convolutional neural networks are deep learning models capable of detecting complex features of an image. Convolutional neural networks allow the analysis, classification, and regression of multidimensional data; without necessarily performing data preprocessing. [8], [9].

Most artificial intelligence applications have been based on a single pathology, mainly diabetic retinopathy. Lim [10] developed an algorithm to detect multiple pathologies (Glaucoma, age-related macular degeneration, diabetic retinopathy, and diabetic retinopathy with risk of vision loss). The system used eight fields of neural networks with a VGGNet architecture, with two networks for diabetic retinopathy severity classification, two networks for referable Glaucoma, and a set of two networks for the identification of age-related macular degeneration requiring referrals, obtained a sensitivity of 90% a specificity of 73.3% and an accuracy of better than 0.89.

The effectiveness of machine learning considers the performance of the model. A confusion matrix is used, where the proposed image will be classified and compared with the accurate classification, obtaining the values of sensitivity and specificity. Accuracy evaluates the training of the model when there are many samples. When there is an imbalance in the selection of categories, it is more reliable to use the accuracy, which evaluates the quality of the model in the classification task, and the completeness metric on how much the model can identify, i.e., what percentage can correctly select.

For value global, the F value is used, a value equal to 1 considers the model to be perfect, and 0 implies that the model does not conform to reality. Also, the receiver operating characteristic (ROC) curve is used. Ensemble methods involve the combination of multiple classifier models, producing a final classifier model. At the same time, transfer learning is a method of adapting a model trained in one domain to another domain [11]. It was decided to evaluate two models with sufficient accuracy and sensitivity for medical standards to test eye disease detection and classification. The first model used machine learning, combined with ensemble methods to evaluate different image classifications. The second model used deep learning. We made this comparison to explore which performed better on a sample of Mexican data.

2.1 First model

The first model was made from 1352 images, classified into 10 categories verified by a retina physician. The diagnoses were: retinal thinning (n = 36); drusen (n = 25); glaucoma (n = 103), macular degeneration (n = 15); mild diabetic retinopathy (n = 380); moderate diabetic retinopathy (n = 45); severe diabetic retinopathy (n = 20); hypertensive retinopathy (n = 68) and normal (n = 650). Given the small sample number, data magnification technique was performed to improve accuracy (4) using the following parameters: rotation range = 180 °, width shift range = 12%, height shift range = 12%, zoom range = 12%, shear range = 20%, horizontal flip = True, fill mode = "closest".

Two types of magnification were applied, which increased all the subgroups by 650 images considering the largest group. The second considered enlarging all groups up to 320 images for each category. Also, the images were divided between normal and diseased eyes. A pre-trained CNN was used to extract the relevant features from each image, and then we input them into a multiclass algorithm. We used the pre-trained CNN VGG16 model to extract each image's relevant features due to its relatively simple structure (16 layers) and good performance. In the 2014 ImageNet competition, it obtained accuracy values between 96% and 97%. It has 13 convolutional and three dense layers trained to solve 1000-class classification problems in ImageNet, which is a good candidate for binary classification [12]. To find the model that best fits and classifies the preprocessed data, several multiclass algorithms were implemented. First individually: Support Vector Machine (SVM), Random Forests (with a depth of eight levels), Logistic Regression, Bernoulli Naïve Bayes, and Multinomial Naïve Bayes. Bagging methods [13] were then used: using the algorithms as a baseline and random forest estimators. The fitted models and the average of their predictions were used, thus reducing their variance. To evaluate the performance of each trained model, we used the ROC curve using a confusion matrix. In addition to prediction, completeness, sensitivity, specificity, and accuracy (Figure 1).

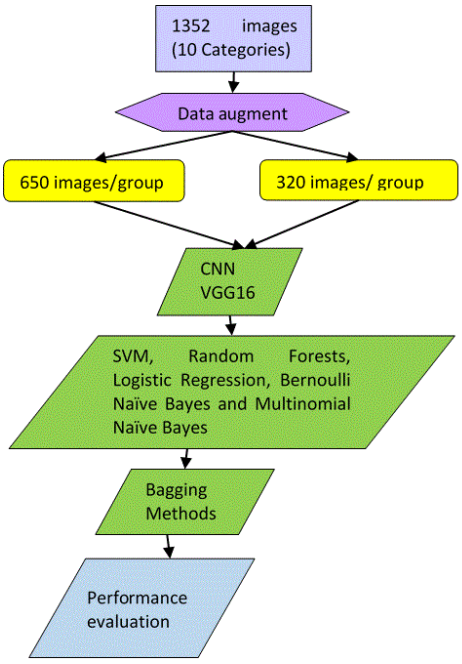

Fig. 1. Flow diagram of the first model

2.2 Second model

We used convolutional neural networks to classify the fundus images into pathologies: Glaucoma or suspicious excavation, diabetic retinopathy, hypertensive retinopathy, as well as those images that do not present any pathology. The fundus images in JPG format with dimensions 2592x1944 in color, black background, and the diagnosis. The images were divided into four main categories: 959 images of healthy patients, 959 images suggestive of fundus excavation or Glaucoma, 466 images of hypertensive retinopathy, and 888 images of diabetic retinopathy with a total sample of 3,272 images.

We divided the set of images according to their purpose: the training set, the validation set, and the test set. We increased the number of the data set; for this, we rotated the images +/- 15º until we reached a rotation of 165º of all images; we also performed a horizontal and vertical mirror rotation and altering the brightness by 30%. We used an Intel i7-4790 60GHz 8-core processor and 16GB of RAM and an Nvidia GTX 1080 GPU for the training. The data augmentation process was cropping a part of the tensor and resizing, moving the image right or left, up or down, and readjusting the contrast and brightness randomly, normalizing. We applied training with backpropagation and downgradient. The validation inference is performed by receiving a tensor height x width x 3 passing through the convolution process (height x width x 3 (RGB) | pool => 50x50x5 | pool => 35x35x10| pool), then passing to full connected layers (=> FC 512 => FC 512 => FC 512 => FC 512 => FC 512 => Softmax: 4) finally the softmax layer which gives us the probabilities of each classification, taking as prediction the classification with the highest probability.

We add a simplified flowchart of the CNN training process in Figure 2.

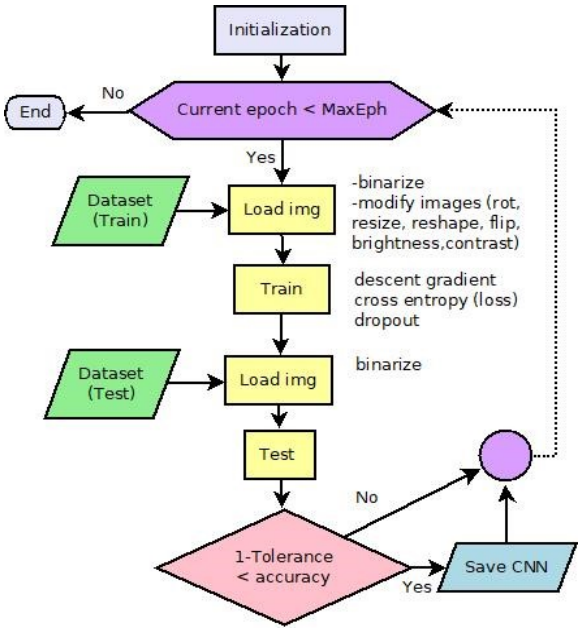

Fig. 2. Flow diagram of the second model.

3 Results

3.1. First model

The first data augmentation approach outperformed the second in almost all classifiers in the 5% to 25% range, except for Multinomial and Bernoulli Naïve Bayes. The second approach outperformed the first with 5% accuracy. This data at both the data increase to 230 images per subclass and 63. Bagging methods improve the accuracy score in the range of 2% -8%, depending on the classification algorithm.

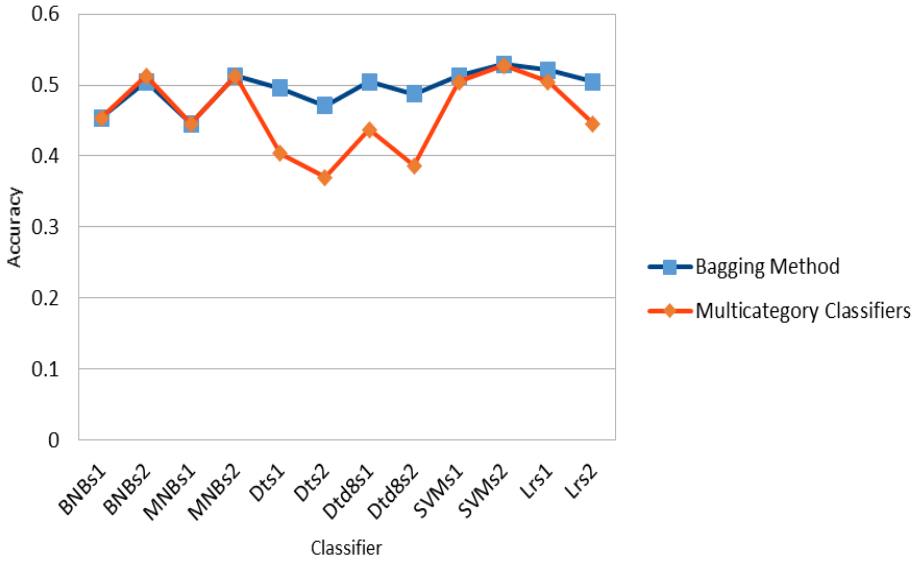

Fig. 3. Bagging method versus multi-category classifiers.

The bagging models in this table were built using 10-base estimators and allowing for an initializer. The multi-category classifiers used in this table are Support Vector Machine (SVM), Decision tree (Dt and Dtd8, where d8 represents the maximum depth used in the decision tree), Logistic Regression (LR), Naïve Bernoulli Bayes (BNB), Multinomial Naïve Bayes (MNB). The suffix s1 and s2 represent the subdivision of the category used for training (Fig.3).

Table 1. Performance results of the multicategory deep learning model for each class subdivision (model 1).

Number of categories	Accuracy	Sensitivity	Precision	F1 Score
10 Classes	0.4308	0.4332	0.2012	0.1773
8 Classes	0.5294	0.5629	0.2503	0.2383
6 Classes	0.5855	0.6003	0.3380	0.3303
2 Classes	0.8884	0.8973	0.8915	0.8943

Table 1 shows the model with the best performance in each subdivision. We found an overall increase in subdivision two over subdivision one between 4% and 16%. The results shown here were obtained from the model trained by the second subdivision.

Figure 4 shows the value of the best performance obtained in the subdivision of 6 classes with the MNB2 model (sensitivity of 61.38%, accuracy of 56.75%, and F1 Score of 35.85%). The basic estimators used in this graph are the support vector machine (SVM), the decision tree (Dt and Dtd8 where d8 represents the maximum depth used in the decision tree), logistic regression (LR), Bernoulli Naïve Bayes (BNB), Multinomial Naïve Bayes (MNB) and Random Forest (RF). The suffix s1 represents models that used all levels of diabetic retinopathy to train the diabetic retinopathy class. In contrast, s2 represents models that used only severe diabetic retinopathy to train the diabetic retinopathy class.

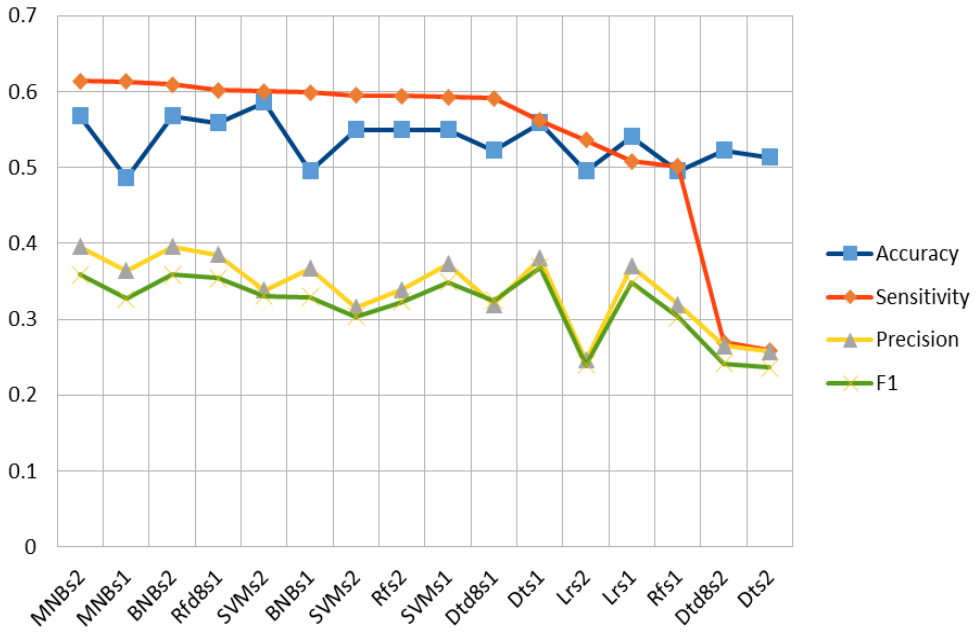

Fig. 4. Performance Measurements of Six-Class Bagging Methods

Several algorithms were used to train the models: SVM, Bernoulli Naïve Bayes, multinomial Naïve Bayes (MBN), decision trees and logistic regression. The SVM algorithm outperformed the latter in almost all subclassifications (between 5% and 25%), except those based on Bayes, which could classify the proposed second-class subclassification with 5% better accuracy. The ensemble methods improve the accuracy score in the range of 2% to 8%, depending on the classification algorithm. The models were trained using features extracted from a background image dataset with VGG-16.

The data set was divided into 90% for training and 10% for testing. Data augmentation techniques were applied to the training data set to improve the sensitivity of the models. For the screening test algorithm, a slight modification of the data was performed to measure performance. From the initial data set, we had 650 belonging to the normal class and 702 in total for the pathologies. We divided the data set into 90% for training and 10% for testing and then applied data augmentation to double the amount of data for training. The performance results are described in the ROC curve shown in Figure 3. SVM is the algorithm that stands out from the rest of the machine learning techniques, showing a sensitivity of 89.7%, followed by RF and LR. The performance results are described in the ROC curve shown in Figure 5.

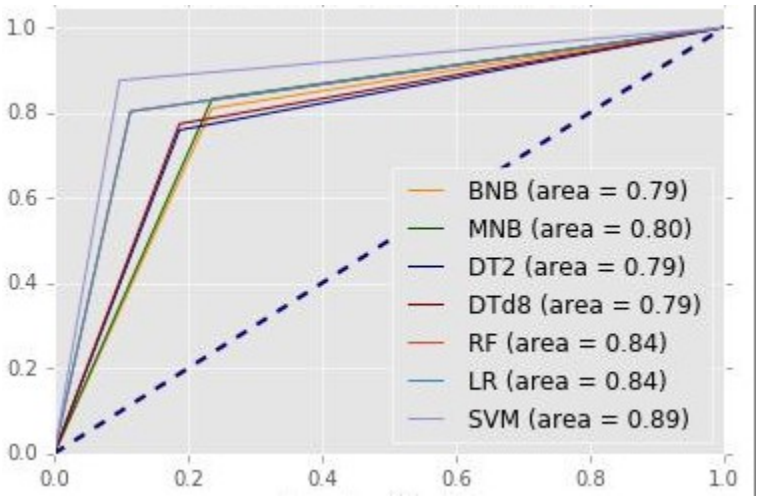

Fig. 3. Receiver operating characteristic curves of various trained models**.**

3.2 Second model

The confusion matrix shows the values with the Algorithm diagnosis and the Expert Diagnosis (Table 2); with these data, the values of the diagnostic test evaluation were obtained: sensitivity, specificity, precision, negative predictive value (NPV), positive likelihood ratio (LR +), negative likelihood ratio (LR-), accuracy and F1 Score shown in Table 3.

Table 2. Confusion Matrix of the Algorithm Diagnosis versus actual diagnosis.

	Real diagnosis				
Algorithm Diagnosis	Healthy	Glaucoma	Hypertensive	Diabetic	Total
Healthy	459	105	2	111	677
Glaucoma	50	595	0	84	729
Hypertensive	39	5	442	101	587
Diabetic	411	254	22	592	127
Total	959	959	466	888	327

Table 3. Evaluation of diagnostic test for each of the parameters (overall accuracy : 63.81%)

Diagnostic	Sensitivity	Specificity	Precision	Negative Predictive Value	Positive Likelihood Ratio	Negative Likelihood Ratio	Accuracy	F1 Score
Healthy	0.48	0.91	0.68	0.82	0.332	0.52	0.794	0.56
Glaucoma	0.62	0.94	0.82	0.86	10.33	0.4	0.816	0.70
Hypertensive	0.95	0.95	0.75	0.99	19	0.05	0.752	0.84
Diabetic	0.64	0.85	0.64	0.85	4.27	0.42	0.432	0.55

We performed the tests obtaining the following ROC Curve values for each of the classifications: Healthy, Hypertensive Retinopathy, Diabetic Retinopathy and Suspicious Glaucoma / Excavation, respectively (table 4).

Table 4. The area under the curve of each of the Categories

Diagnostic	Area	Standard error	Sig	IC 95% Upper limit	Lower limit
Healthy	.572	.019	.000	.536	.609
Hypertensive	.866	.021	.000	.824	.907
Diabetic	.651	.020	.000	.612	.690
Glaucoma	.851	.012	.000	.827	.874
Global	.750	.008	.000	.734	.767

In addition, we evaluated the coefficients of agreement between the algorithm and the expert's diagnosis. We had a kappa coefficient of 0.513, which indicates an acceptable performance.

Also, the coefficients of agreement between the algorithm and the expert's diagnosis were evaluated. A kappa coefficient of 0.513 was obtained, indicating acceptable performance. The diagnosis obtained between healthy and diseased patients are lower than the values reported in international literature for retinopathy, reported in accuracy values of 0.99-0.88, sensitivity 100-91, Specificity 98-76. In Glaucoma, the accuracy of 0.92 to 0.98, sensitivity 84 to 96, and specificity 88-94, age-related macular degeneration, accuracy is 0.93-0.94, a sensitivity value ranging from 84.2 to 96.4, and a specificity of 88.7 to 94.3 [14]. In Mexican studies, Vega [15] evaluates abnormalities in blood vessels to make the diagnosis of different diseases, obtaining accuracy values of 0.94, specificity of 0.96, and sensitivity 0.74; we should note that to test their methodology, they used the STARE dataset, which is a sample of international images. These studies have been carried out in the Latin American population; in Chile, they have obtained sensitivity values of 94.6 for retinopathy screening, specificity of 74.3 [16]. In studies that evaluate multiple abnormalities, Son and cols (2020) evaluated 12 different findings, Glaucoma with an area under the curve of 98.2. The other findings mentioned are not comparable with our results when referring to alterations such as haemorrhages, exudates, drusen, or cottony patches, among others [17].

4 Conclusions

The use of machine learning may be promising as a screening tool for detecting fundus pathologies in Mexican patients. However, we face several challenges: the main one is associated with the lack of a complete image bank of Mexican patients, since the health sector is divided, and this information is not standardized in digital form.

In the first model, it is observed that (except for the models trained with MNB) the results showed that it performed better when only severe retinopathy was used. The main reason is probably the well-defined features shown in severe retinopathy versus the other retinopathy levels. To eliminate these biases, it will be desirable for future studies to consider these subdivisions of wet and dry disease of age-related macular degeneration, as the significant difference in clinical presentation may explain the low values in screening [7, 17].

The results suggest that the 6-class classification may be promising to perform evaluations with a larger sample of images and improve its sensitivity and specificity values. The results are shown in both the screening test algorithm and the multiple classification algorithms. It allows us to develop new models, which with a more extensive data set could be used by healthcare and teaching staff and perform the classification of other types of images.

The second model found that the hypertensive retinopathy images seem to perform better than the rest. It may be that, although the training, validation, and test images are different, given the small number of images, there is probable overfitting, which needs to be verified by enlarging the image sample.

The data indicate that the highest incidence of false positives occurs with diabetic retinopathy diagnosed as absent pathology. It was observed in the confusion matrix and the ROC curve, with difficulty detecting it and being confused mainly with the category of healthy eyes. This may be due to two main reasons: The patients in the database from which the images were derived could have lesions indicative of some variety of diabetic retinopathy in one eye, but not in the other, despite which both eye images fell into this category. Furthermore, even if the ophthalmologist classifies it as healthy, it must be ruled out that the algorithm is not detecting an alteration that has not been clinically reported. Also, it is possible that, given the wide variety of disease presentations, the network has not been refined enough to detect the more subtle features as in other categories that would have more specific features. Leaving aside hypertensive retinopathy, the confusion matrix and kappa coefficient of 0.513 indicate acceptable or good performance. However, it does not give sufficient certainty desired for medical or educational use [6], so further training of the algorithm is required.

Adding a model that allows differentiation of the blood vessel pixels from the rest of the eye may help diagnose hypertensive retinopathy. In this regard, the LNNDP model proposed in [14] may be helpful.

Several limitations and difficulties were encountered during this work; the main one was the compilation of the data set, as well as the fact that the different disease categories had different sample values, in addition to the fact that several patients had

multiple pathologies, which is quite common in the medical field. This problem generated a bias in our data set that may explain the results when classifying multiple pathologies.

One challenge to clinical adoption in deep learning algorithms is a shift in mindset where clinicians rely on machine diagnostics, mainly when the "black box" problem arises, to incorporate it into existing screening systems, so it could first be "diagnosed" by deep learning, and these images subsequently reviewed by a clinician. It is necessary to have high-quality "labeled" input data with multiple medical specialists to learn and classify images from large amounts of data. It often requires millions of observations to reach acceptable performance levels and requires having a heterogeneous population with different devices. These procedures also involve ethical questions regarding liability in case of screening and non-detection causing harm to the patient, so it is necessary to expand the evaluation sample to improve the evidence of use, acceptability, safety, validity, reproducibility, and reliability on the part of the patient and the physician.

References

1. González, M., Arredondo, B., González, C.: Diabetic retinopathy in Mexico. Prevalence and clinical characteristics. Arch Med Res. 25 (3), 355--360 (1994)
2. Moreno, D., Rayón-Rodríguez, M., García-Leonardo, J., Hernández-Solís, A., Landa- Alvarado, P.: Retinochoroidal findings and lung cancer: first report in mexican population. Arc Soc Oftal Esp. 94 (3), 125--129 (2019)
3. Ham-Chande, R., Rojas-Huerta, A. V., Gudiño, M. R.: Envejecimiento por cohortes de la población mexicana de 60 años y más. Rev Int Estds y Geo, 6 (2) 64-72. (2015)
4. Yau, J.W.Y., Rogers, S.L., Kawasaki, R., Lamoureux, E.L., Kowalski, J.W., Bek, T., et al.:Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care , 35, 556—64 (2013)
5. Vijayaraghavan, P. et al.: Accuracy of the smartphone-based nonmydriatic retinal camera in the detection of sight-threatening diabetic retinopathy. Ind J Ophthalmol, 68 (Suppl 1), S42--S46 (2020). doi:10.4103/ijo.IJO_1937_19
6. Cheung, C.Y., Tang, F., Ting, D.S.W., Tan, G.S.W., Wong, T.Y.: Artificial Intelligence in Diabetic Eye Disease Screening. Asia Pac J Ophthalmol (Phila), 10, 22608 (2019) doi:10.22608/APO.201976
7. Lu, D., Heisler, M., Lee, S., Ding, G.W, Navajas, E., Sarunic, M, Beg, M.: Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. Medical image analysis, 54, 100—110 (2019)
8. He, Y., Guo, J., Ding, X., van Ooijen, P. M., Zhang, Y., Chen, A., & Xie, X.: Convolutional neural network to predict the local recurrence of giant cell tumor of bone after curettage based on pre-surgery magnetic resonance images". Euro Soc Radiology, 29 (10), 5441--5451 (2019)
9. Bengio, Y., Courville, A., Vicent, P.: Representation learning: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1798-1828. (2013)
10. Lim, G., Bellemo, V., Xie, Y., Lee, X.Q., Yip, M.Y.T., Ting, D.S.W.: Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review. Eye Vis (Lond), 14-21 (2020)
11. Asperti, A. & Mastronardo, C.: The Effectiveness of Data Augmentation for Detection of Gastrointestinal Diseases from Endoscopical Images. In S. Wiebe, H. Gamboa, A. Fred, & S. Bermúdez i Badia (Eds.), Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 2: KALSIMIS. Madeira Portugal, pp. 199–205 (2018)
12. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In Advances in neural information processing systems, 3320-3328 (2014)
13. Dietterich, T.: Ensemble Methods in Machine Learning. In International workshop on multiple classifier systems, pp. 1-15 (2000)
14. Ting, D., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G., Schmetterer, L., Keane, P. A., & Wong, T. Y.: Artificial intelligence and deep learning in ophthalmology. The British journal of ophthalmology, 103(2), 167-175 (2019)
15. Vega, R., Sanchez-Ante, G., Falcon-Morales, L. E., Sossa, H., & Guevara, E. : Retinal vessel extraction using Lattice Neural Networks with Dendritic Processing. Computers in biology and medicine, 58, 20-30 (2015)
16. Arenas-Cavalli, J. T., Abarca, I., Rojas-Contreras, M., Bernuy, F., & Donoso, R. Clinical validation of an artificial intelligence-based diabetic retinopathy screening tool for a national health system. Eye (London, England), 10.1038/s41433-020-01366-0. (2021)
17. Son, J., Shin, J. Y., Kim, H. D., Jung, K. H., Park, K. H., & Park, S. J. Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images. Ophthalmology, 127(1), 85-94 (2020)