

Artificial intelligence-based model for the interpretation and reporting of standard automated perimetry

Joacy Pedro Franco David¹ , Alexandre Antonio Marques Rosa¹ , Rafael Scherer² ,

Cláudio Eduardo Corrêa Teixeira³ , Douglas Costa² 

1. Universidade Federal do Pará, Belém, PA, Brazil.

2. Bascom Palmer Eye Institute, Miami, Miami, FL, United States of America.

3. Centro Universitário Estado do Pará, Belém, PA, Brazil.

ABSTRACT | Purpose: Standard automated perimetry has been the standard method for measuring visual field changes for several years. It can measure an individual's ability to detect a light stimulus from a uniformly illuminated background. In the management of glaucoma, the primary objective of perimetry is the identification and quantification of visual field abnormalities. It also serves as a longitudinal evaluation for the detection of disease progression. The development of artificial intelligence-based models capable of interpreting tests could combine technological development with improved access to healthcare. **Methods:** In this observational, cross-sectional, descriptive study, we used an artificial intelligence-based model [Inception V3] to interpret gray-scale crops from standard automated perimetry that were performed in an ophthalmology clinic in the Brazilian Amazon rainforest between January 2018 and December 2022. **Results:** The study included 1,519 standard automated perimetry test results that were performed using Humphrey HFA-II-i-750 (Zeiss Meditech). The Subsequently, 70%, 10%, and 20% of the dataset were used for training, validation, and testing, respectively. The model achieved 80% (68.23%–88.9%) sensitivity and 94.64% (88.8%–98%) specificity for detecting altered perimetry results. Furthermore, the area under the receiver operating characteristic curve was 0.93. **Conclusions:** The integration of artificial intelligence in the diagnosis, screening, and monitoring of pathologies represents a paradigm shift in ophthalmology, enabling significant improvements in safety, efficiency, availability, and accessibility of treatment.

Keywords: Glaucoma; Disease progression; Perimetry; Visual Fields; Visual field tests; Artificial intelligence; Neural networks, computers; Machine learning

Submitted for publication: October 9, 2024
Accepted for publication: January 24, 2025

Funding: This study received no specific financial support.

Disclosure of potential conflicts of interest: The authors declare no potential conflicts of interest.

Corresponding author: Joacy Pedro Franco David.
E-mail: joacydavid@outlook.com

Approved by the following research ethics committee: Centro Universitário Estado do Pará (# 4.418.799).

INTRODUCTION

Standard automated perimetry (SAP) has been the standard method for evaluating visual field changes for several years. SAP measures an individual's ability to detect a light stimulus emitted from a uniformly illuminated background⁽¹⁾. The main goals of SAP in glaucoma management are the identification and quantification of abnormalities in the visual field and serving as a longitudinal assessment method for the detection of disease progression⁽²⁾.

In SAP, the patient's central and peripheral vision sensitivity is quantified using algorithms that precisely determine the sensitivity threshold at each visual field location. At these locations, the light stimulus is presented at different intensities, and the patient must respond whenever they perceive the applied stimulus. Subsequently, a probability analysis is performed by comparing these responses with those of healthy individuals and individuals with the disease that are included in a database⁽²⁾.

Artificial intelligence (AI) is a branch of computer science that attempts to emulate human behavior or reasoning in computers. AI broadly encompasses several modalities, including machine learning (ML) and deep learning (DL)⁽³⁾. It was first discussed in 1956⁽⁴⁾ and was defined as a technology capable of imitating human behavior. Since then, there have been significant developments in AI and the creation of subfields such as ML and DL.

The World Economic Forum considers AI as the fourth industrial revolution in the history of humankind. AI modalities such as ML and DL have been widely used in the medical field and are generally applied in the processing of medical imaging. They have exhibited promising results in diagnosing various pathologies such as tuberculosis via radiographic examination, cutaneous

melanoma via photographs, and lymph node metastasis secondary to breast cancer on the basis of histopathological examination⁽⁵⁻⁷⁾.

In ophthalmology, AI and DL have been used to analyze ophthalmological images, especially optical coherence tomography (OCT) and color fundus photography. Furthermore, AI has demonstrated promising results in the interpretation of SAP, which is widely used in the diagnosis and monitoring of glaucoma^(8,9).

In this study, we aimed to develop an AI program capable of differentiating normal SAP results from pathological SAP results and categorizing them as normal and altered results, respectively. Furthermore, we aimed to evaluate the performance of the AI model and demonstrate its effectiveness as a screening tool for visual field pathologies.

METHODS

This is a cross-sectional and descriptive study of the creation of an AI model capable of differentiating normal SAP records from pathological SAP records. At this time, the model does not aim to determine the pathology or classify the severity of a specific disease as its purpose is to serve as a modality for screening SAP records.

Data collection

The study was developed using a database that included the anonymous results of SAPs that had been performed in patients treated at two private practices in Belém do Pará between January 2018 and December 2022. The research sample consisted of 1,519 SAP records, which did not state the specific baseline pathologies, visual status, race, age, or sex. All exams were performed using a Humphrey HFA II-i 750i® perimeter (Carl Zeiss Meditec, Inc). All the SAPs were based on the Swedish interactive threshold algorithm, which is a rapid method of examination

Statistical analysis

After data collection, the SAP reports were analyzed by two glaucoma specialists and classified as normal or altered (visual field defects). Classical glaucoma changes such as arcuate defects, nasal steps, and paracentral scotomas and visual field defects emerging from other pathologies such as bitemporal hemianopsias secondary to central nervous system tumors were considered as altered visual fields.

The developed AI model utilized InceptionV3, a convolutional neural network (CNN), for transfer learning from a pre-trained model. Specifically, the model used the printout gray-map images generated by the SAP test, leveraging InceptionV3's capability to effectively learn discriminative features. The deeper layers of the network were frozen during training, using weights pre-trained on the ImageNet database. However, the final layer was unfrozen to allow the model to adapt these features for visual field classification.

The model employed the rectified linear unit as the activation function. Thereafter, regularization was achieved using a 20% dropout technique. Finally, a sigmoid activation function was utilized. The Adam optimizer was selected for its ability to adaptively balance the learning rate during training, and the loss was calculated using the cross-entropy method. The model was configured to train with 40 epochs, and callbacks were implemented to save the best performing model based on the accuracy of the validation set. Given the relative simplicity of the gray-map images, no preprocessing techniques were applied other than standardizing pixel values from 0 to 255 to values between 0 and 1.

A total of 634 low-reliability exams were excluded from the analysis. Of the remaining 885 exams, approximately 70% (n=620) were used for model training, 10% (n=90) were used for validation, and the remaining 20% (n=175) were used for testing. The data were split into one of these sets at the patient level to ensure that the same patient was not included in more than one set. This splitting methodology prevents data leakage and ensures that the model does not learn patient-specific characteristics, thereby improving its generalizability across diverse datasets. Any test result with >20% fixation losses and >15% of false positives, which are the historical and traditional cut-off points used for the SAP-related analysis, were considered low-reliability exams.

The following metrics were used to analyze the model's performance, considering $p < 0.05$ and 95% confidence interval (CI) using the bootstrapping method:

- Accuracy:** It is a general measure of the AI model's success. It is calculated from the coefficient between the number of correctly categorized exams and the total number of exams⁽¹⁰⁾.
- Sensitivity:** It refers to the model's ability to identify truly altered exams. It is obtained through the coefficient between the truly altered exams (true positives) and the sum of the true positive and false negative

results. A high sensitivity rate, which indicates a high rate of accurate identification of altered exams⁽¹¹⁾, is generally desirable in screening tests. The correct identification of altered exams would reduce the number of pathological results that may be missed by the model.

- c) Specificity: It refers to the model's ability to identify normal exam results as negative results. It is calculated by the coefficient of the true negatives and the sum of the true negative and false positive results. High specificity, which indicates a high rate of accurate detection of normal exams⁽¹²⁾, prevents an overload of patients with normal exams being referred for further investigation.
- d) F1 score: It is the harmonic mean of the accuracy and detection of true positives. It is a balanced measure of the model that is both sensitive and specific. The score varies between 0 and 1, with higher scores indicating better performance⁽¹³⁾. F1 score is used when a clear number representing the model's performance is desirable, especially when dealing with an uneven dataset.
- e) Area under the receiver operating characteristic curve (AUC-ROC): It measures the model's ability to distinguish abnormal results and normal results at different classification thresholds⁽¹⁴⁾. A good AUC-ROC is desirable, as the cost of not detecting a false negative results is higher than that of obtaining false positive result. False negative SAPs may delay the treatment of ocular pathologies, increasing treatment costs and impairing prognosis. However, false positive results may be further assessed to strongly establish a diagnosis before initiating treatment.

RESULTS

The sensitivity and specificity of the AI model were 80% (68.23%–88.9%) and 94.64% (88.8%–98%), respectively. Using the cross-entropy method, the accuracy and loss rate of the model were 89% and 0.56, respectively (Table 1, Figure 1).

The model achieved an AUC-ROC of 0.93 (Figure 2). In general, an area under the curve (AUC) of 0.5 indicates no discriminatory ability (i.e., ability to differentiate between patients with and without the disease or condition based on the test), AUC of 0.7 to 0.8 indicates acceptable ability, AUC of 0.8 to 0.9 indicates excellent ability, and >0.9 indicates outstanding ability⁽¹⁵⁾.

In the test dataset, 65 results were altered and 112 were normal. The model achieved an accuracy of 0.90 and sensitivity of 0.80. The F1 score for the altered and normal exams were 0.85 and 0.92, respectively (Table 2).

Table 1. Statistical analysis of the AI model's performance

Statistic	Value	95% CI
Sensitivity	80.00%	68.23%–88.90%
Specificity	94.64%	88.70%–98.01%
Accuracy	89.27%	
Loss	56.94%	

*CI= confidence interval.

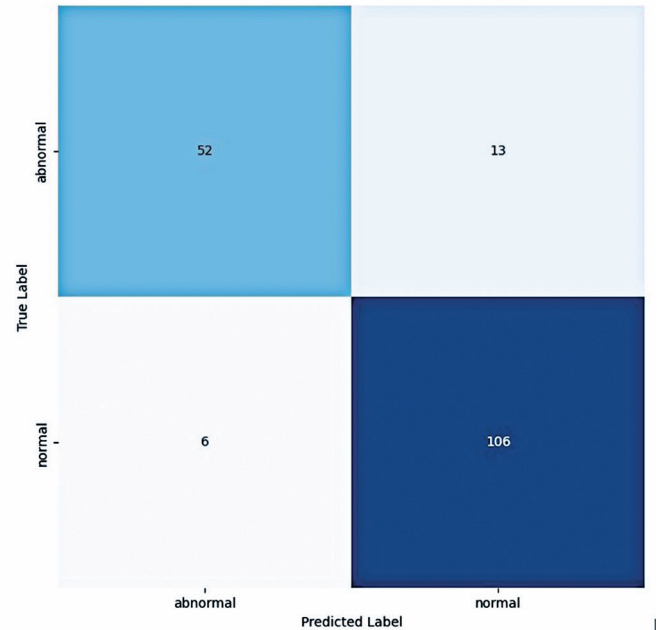


Figure 1. Confusion matrix.

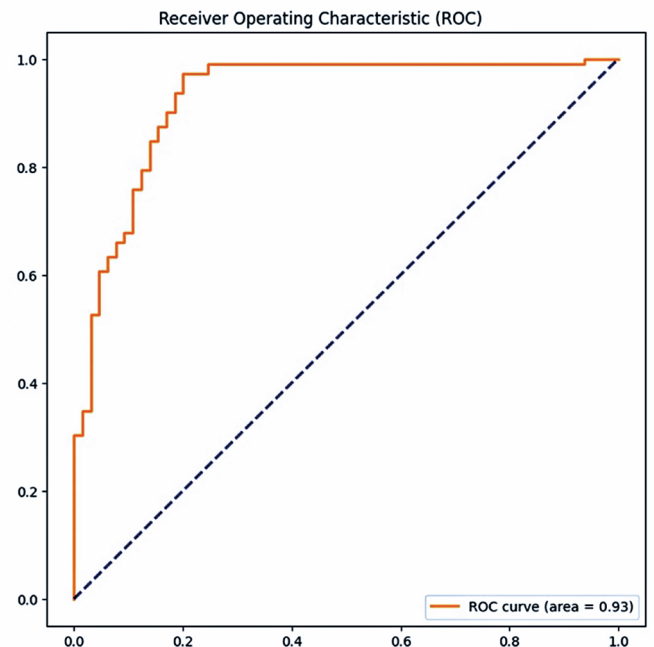


Figure 2. ROC curve of the AI model.

ROC= receiver operating characteristic; AI= artificial intelligence.

Table 2. Computational statistics of the AI model's performance

	Precision	Recall	F1- score	Sample
Altered	0.90	0.80	0.85	65
Normal	0.89	0.95	0.92	112

DISCUSSION

In recent years, there has been an increase in the consolidation and maturation of technological innovations with the potential to revolutionize different areas of human life, including healthcare. These innovations such as 5G, AI, and ML have created several possibilities and opportunities for healthcare development, which may solve the numerous problems faced by managers and public and private health services. This will ensure universal and sustainable coverage for a rapidly growing and aging population, thereby significantly increasing the demand and cost. Furthermore, these technologies could redefine the way we screen, diagnose, and monitor various pathologies. Thus, in our study, we aimed to present an AI- and ML-based model that is capable of effectively screening for ophthalmological diseases that can alter SAP.

Glaucoma is one of the diseases that alters SAP. It is the main cause of irreversible blindness in the world, affecting approximately 64.3 million people between the ages of 40 and 80. It is estimated that this number will grow to 112 million people by 2040^(16,17).

It is crucial to have an effective screening method for diseases such as glaucoma because glaucoma is asymptomatic in the early stages, which often causes a delay in diagnosis. Furthermore, by the time the patient seeks medical help for low visual acuity, their glaucoma is in its advanced stages. The cost of treating advanced glaucoma is reportedly four times higher than that of treating early-stage glaucoma. Although most irreversible vision loss can be avoided with early diagnosis and treatment, there are few effective glaucoma screening protocols. This may be attributed to the fact that glaucoma does not meet population screening criteria, with high rates of false positive results when screening individuals with early-stage glaucoma^(18,19).

There are several ways to apply AI in the diagnosis of ocular pathologies such as glaucoma. Exams such as OCT and color fundus photography have been widely used as a diagnostic and screening method due to their objectivity. SAP has also yielded interesting results as a screening method^(20,21).

SAP is a low-cost exam when compared with OCT. Furthermore, any professional or technician is capable of effectively performing the examination without requiring any additional extensive training. However, interpreting the results of SAP is challenging, even by ophthalmologists who are not specialists in the field. This challenge can be mitigated by our AI model which has been trained using reports prepared by experts in the field.

The AI model does not require a professional to interpret the results, thereby reducing the cost of the process and increasing the capacity to absorb the growing demand. Thus, patients with altered exam results can be immediately identified and quickly referred to a specialist for evaluation.

Because our model is based on visual field reports, the application can be used on a simple smartphone with a regular camera, which is available to any physician. This is crucial when dealing with highly prevalent diseases, which require the processing of numerous tests.

Our AI model could also contribute to primary care. SAP could be used for screening at less complex health services, which may aid in prioritizing the referral of patients who need it the most. In areas or cities where there are no ophthalmologists trained to evaluate SAP, the model could partially fulfill the need of an expert, which may help primary care physicians manage the situation properly.

Our model was able to satisfactorily screen and interpret the SAPs, with a sensitivity of 80%. In 1994, a study used a dual-layer neural network to detect glaucomatous changes in visual field examinations, achieving a sensitivity of 65% and specificity of 72%⁽²²⁾. This difference in results may be attributable to the evolution and sophistication of the technology over the years and the fact that the study focused on detecting only glaucomatous changes.

In a 2003 study, neural networks were used to detect the progression of glaucomatous damage, which yielded an AUC of 0.92⁽²³⁾. Another study successfully used neural networks to detect preperimetric glaucomatous damage, achieving an AUC of 0.92⁽²⁴⁾. Although the purpose of our study was slightly different, the model achieved an AUC of 0.93. Furthermore, the previous study had been performed in an ethnically homogenous Japanese population, whereas our study evaluated a more ethnically diverse population of native South Americans, Southern Europeans, and sub-Saharan Africans.

Our AI model demonstrated an accuracy of 89% in detecting altered perimetry results regardless of the pathology. In one study, visual field defects typical of glaucoma were detected using neural networks with an accuracy of 87%, demonstrating a performance that is superior to other algorithms and even ophthalmologists⁽²⁵⁾. In another study, a CNN exhibited a strong performance in diagnosing glaucoma, achieving high accuracy in categorizing visual fields as glaucomatous and healthy⁽²⁶⁾.

AI models have demonstrated advantages in perimetry analysis in terms of process automation, efficiency, and objectivity. When compared with the traditional method of perimetry evaluation by an ophthalmologist, AI models save time and reduce subjectivity in interpreting results. Automating the interpretation of results makes the process faster and more efficient, which reduces the operational costs for managers and health services and enables greater scalability of glaucoma screening programs.

Despite the favorable results, our study has some limitations. The AI model cannot differentiate between the diseases that alter the visual field. Furthermore, the severity of glaucoma may impact the sensitivity and specificity of an AI model. Moreover, different pathologies may yield different results when analyzing them individually. A larger database could improve the model's accuracy by increasing the data available for training and validation.

In conclusion, we were able to develop an AI model capable of accurately differentiating normal SAPs from pathological SAPs. Thus, it can satisfactorily be used for screening and interpreting SAP results.

AUTHORS' CONTRIBUTIONS:

Significant contribution to conception and design: Joacy Pedro Franco David, Alexandre Antonio Marques Rosa, Rafael Scherer. **Data acquisition:** Joacy Pedro Franco David, Rafael Scherer. **Data analysis and interpretation:** Joacy Pedro Franco David, Rafael Scherer, Cláudio Eduardo Corrêa Teixeira, Douglas Rodrigues Costa. **Manuscript drafting:** Joacy Pedro Franco David, Rafael Scherer, Cláudio Eduardo Corrêa Teixeira. **Significant intellectual content revision of the manuscript:** Joacy Pedro Franco David, Alexandre Antonio Marques Rosa, Rafael Scherer, Cláudio Eduardo Corrêa Teixeira. **Dinal approval of the submitted manuscript:** Joacy Pedro Franco David, Alexandre

Antonio Marques Rosa, Rafael Scherer, Cláudio Eduardo Corrêa Teixeira, Douglas Rodrigues Costa. **Statistical analysis:** Rafael Scherer, Cláudio Eduardo Corrêa Teixeira, Douglas Rodrigues Costa. **Obtaining funding:** not applicable. **Supervision of administrative, technical, or material support:** not applicable. **Research group leadership:** Alexandre Antonio Marques Rosa.

REFERENCES

1. Andersson S, Heijl A, Bizios D, Bengtsson B. Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field assessment for the diagnosis of glaucoma. *Acta Ophthalmol.* 2013;91(5):413-7.
2. Advanced Glaucoma Intervention Study 2. Visual field test scoring and reliability. *Ophthalmology.* 1994;101:1445-55.
3. Ittoop SM, Jaccard N, Lanouette G, Kahook MY. The Role of Artificial Intelligence in the Diagnosis and Management of Glaucoma. *J Glaucoma.* 2022;31(3):137-146.
4. McCarthy J, Minsky ML, Rocheste, N, Shannon CE. A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Mag.* 2006;27(4):12.
5. Lakhani P, Sundaram B. Deep Learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology.* 2017;284(2):574-582.
6. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol.* 2017;135(11):1170-6.
7. Bejnordi BE, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JA; the CAMELYON16 Consortium; Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MC, Bult P, Beca F, Beck AH, Wang D, Khosla A, Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin HJ, Heng PA, Haß C, Bruni E, Wong Q, Halici U, Öner MÜ, Cetin-Atalay R, Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang YW, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvaari P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venâncio R. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* 2017;318(22):2199-210.
8. Ting DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318(22):2211-2223.
9. Heijl A, Lindgren G, Olsson J. A package for the statistical analysis of visual fields. In: Seventh International Visual Field Symposium, Amsterdam, September 1986. Dordrecht: Springer Netherlands; 1986. p. 153-68.
10. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging.* 2018;9(4):611-29.
11. Chen L, Li S, Bai Q, Yang J, Jiang S, Miao Y. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing.* 2021;13(22):4712.

12. Dhaka VS, Meena SV, Rani G, Sinwar D, Kavita, Ijaz MF, et al. A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. *Sensors (Basel)*. 2021;21(14):4749.
13. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 2021;8(1):53.
14. Raghavendra U, Fujita H, Bhandary SV, Gudigar A, Tan JH, Acharya UR. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. *Inform Sci*. 2018;441:41-9.
15. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health*. 1991;81(12):1630-5.
16. Stevens GA, White RA, Flaxman SR, Price H, Jonas JB, Keeffe J, Leasher J, Naidoo K, Pesudovs K, Resnikoff S, Taylor H, Bourne RRA; Vision Loss Expert Group. Global prevalence of vision impairment and blindness: magnitude and temporal trends, 1990-2010. *Ophthalmology*. 2013;120(12):2377-84.
17. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081-90.
18. Lee PP, Walt JG, Doyle JJ, Kotak SV, Evans SJ, Budenz DL, et al. A multicenter, retrospective pilot study of resource use and costs associated with severity of disease in glaucoma. *Arch Ophthalmol*. 2006;124(1):12-9.
19. Sample PA, Dannheim F, Artes PH, Dietzsch J, Henson D, Johnson CA, Ng M, Schiefer U, Wall M; IPS Standards Group. Imaging and Perimetry Society standards and guidelines. *Optom Vis Sci*. 2011;88(1):4-7.
20. Pizzi LT, Waisbourd M, Hark L, Sembhi H, Lee P, Crews JE, et al. Costs of a community-based glaucoma detection programme: analysis of the Philadelphia Glaucoma Detection and Treatment Project. *Br J Ophthalmol*. 2018;102(2):225-32.
21. Haleem MS, Han L, van Hemert J, Li B. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review. *Comput Med Imaging Graph*. 2013;37(7-8):581-96.
22. Goldbaum MH, Sample PA, White H, Côté B, Raphaelian P, Fechtner RD, et al. Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci*. 1994;35(9):3362-73.
23. Lin A, Hoffman D, Gaasterland DE, Caprioli J. Neural networks to identify glaucomatous visual field progression. *Am J Ophthalmol*. 2003;135(1):49-54.
24. Asaoka R, Murata H, Iwase A, Araie M. Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. *Ophthalmology*. 2016;123(9):1974-80.
25. Li F, Wang Z, Qu G, Song D, Yuan Y, Xu Y, et al. Automatic differentiation of Glaucoma visual field from non-glaucoma visual field using deep convolutional neural network. *BMC Med Imaging*. 2018;18(1):35. Erratum in: *BMC Med Imaging*. 2019;19(1):40.
26. Scherer R, da Costa DR, Jammal A, et al. Detecting glaucomatous visual field loss on standard automated perimetry tests with a hybrid convolutional neural network and parameter-encoded color images. Poster presented at: American Glaucoma Society Annual Meeting; February 29-March 1, 2024.