Original Article | Data Sciences

Heatmap analysis for artificial intelligence explainability in diabetic retinopathy detection: illuminating the rationale of deep learning decisions

Fernando Korn Malerbi[1,2,3]^, Luis Filipe Nakayama[1], Paulo Prado[2], Fernando Yamanaka[2], Gustavo Barreto Melo[1,4], Caio Vinicius Regatieri[1], José Augusto Stuchi[2]

[1]Department of Ophthalmology and Visual Sciences, Federal University of Sao Paulo, Sao Paulo, Brazil; [2]Phelcom Technologies, Sao Carlos, Brazil; [3]Diabetes Center, Federal University of Sao Paulo, Sao Paulo, Brazil; [4]Sergipe Eye Hospital (Hospital de Olhos de Sergipe), Aracaju, Sergipe, Brazil

Contributions: (I) Conception and design: FK Malerbi, LF Nakayama, JA Stuchi; (II) Administrative support: P Prado, F Yamanaka; (III) Provision of study materials or patients: FK Malerbi, LF Nakayama, GB Melo, CV Regatieri; (IV) Collection and assembly of data: FK Malerbi, LF Nakayama, P Prado, F Yamanaka, GB Melo, CV Regatieri; (V) Data analysis and interpretation: FK Malerbi, LF Nakayama, P Prado, F Yamanaka, JA Stuchi; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Fernando Korn Malerbi, MD, PhD. Department of Ophthalmology and Visual Sciences, Federal University of Sao Paulo, Rua Botucatu, 821 CEP 04023-062, Sao Paulo, Brazil; Phelcom Technologies, Sao Carlos, Brazil; Diabetes Center, Federal University of Sao Paulo, Sao Paulo, Brazil. Email: malerbi.fernando@gmail.com.

Background: The opaqueness of artificial intelligence (AI) algorithms decision processes limit their application in healthcare. Our objective was to explore discrepancies in heatmaps originated from slightly different retinal images from the same eyes of individuals with diabetes, to gain insights into the deep learning (DL) decision process.

Methods: Pairs of retinal images from the same eyes of individuals with diabetes, composed of images obtained before and after pupil dilation, underwent automatic analysis by a convolutional neural network for the presence of diabetic retinopathy (DR), output being a score ranging from 0 to 1. Gradient-based Class Activation Maps (GradCam) allowed visualization of activated areas. Pairs of images with discordant DL scores or outputs within the pair were objectively compared to the concordant pairs, regarding the sum of activations of Class Activation Mapping (CAM), the number of activated areas, and DL score differences. Heatmaps of discordant pairs were also qualitatively assessed.

Results: Algorithmic performance for the detection of DR attained 89.8% sensitivity, 96.3% specificity and area under the receiver operating characteristic (ROC) curve of 0.95. Out of 210 comparable pairs of images, 20 eyes and 10 eyes were considered discordant according to DL score difference and regarding DL output, respectively. Comparison of concordant versus discordant groups showed statistically significant differences for all objective variables. Qualitative analysis pointed to subtle differences in image quality within discordant pairs.

Conclusions: The successfully established relationship among objective parameters extracted from heatmaps and DL output discrepancies reinforces the role of heatmaps for DL explainability, fostering acceptance of DL systems for clinical use.

Keywords: Artificial intelligence (AI); deep learning (DL); diabetic retinopathy (DR); explainability; retina

Submitted Apr 15, 2024. Accepted for publication Aug 27, 2024. Published online Oct 12, 2024.
doi: 10.21037/atm-24-73
View this article at: https://dx.doi.org/10.21037/atm-24-73

^ ORCID: 0000-0002-6523-5172.

Introduction

Artificial intelligence (AI) has recently gained public prominence with the advent of deep learning (DL) models, which correspond to computational models with multiple processing layers that learn representations of data with multiple levels of abstraction (1). Such models have shown promising performance in areas such as image and speech recognition, as well as natural language processing (2-4). In medicine, one of the areas with notable advances involving AI is the screening of diabetic retinopathy (DR), a leading cause of vision impairment and blindness worldwide. Despite the efficacy of screening programs, aimed at early DR detection, and mostly based on the analysis of fundus photographs, their widespread implementation faces several challenges; many established screening systems rely on human graders, a costly and limited resource (5-8). DL algorithms have emerged as a promising solution, offering diagnostic performance for the detection of DR comparable to experts and the potential to scale up screening programs efficiently (7,9). AI-assisted DR screening has been considered a successful and cost-effective strategy for the increasing demand brought by the growing global prevalence of diabetes (8).

However, implementation of AI in healthcare also faces significant challenges, including ethical, technical, and human-centered considerations; one of such challenges is related to explainability, or the ability to summarize the reason for an AI behavior (10,11). Transparency in DL decisions is fundamental for building trust, as it allows stakeholders to gain familiarity with AI decision processes; explainability is also an important consideration when clearing autonomous diagnostic AI products (5). However, due to the nature of DL itself, the estimated function relating inputs to outputs is not understandable at an ordinary human comprehension level, due to the large number of parameters, their complex combinations, or their nonlinear transformations, among other causes (12). Such opaqueness, also referred to as the "black box" nature of DL algorithms, limits their application in health care (12-15).

To address the challenge of opaqueness and contribute to the explainability of DL systems, techniques such as the Gradient-based Class Activation Map (GradCam) have emerged. The GradCam is a class-discriminative localization technique which generates visual explanations for convolutional neural networks, allowing a visual display of a DL system output while unveiling the most important regions within an image that contribute to the algorithm decision-making process (8,16). By providing clinicians with visualizations which are aligned with their expertise, it is assumed that they will more likely trust DL systems and eventually adopt them (5). Since convolutional layers retain spatial information, which is lost in fully connected layers, the last convolution layers are expected to have the best compromise between high-level semantics and detailed spatial information; in that sense, values obtained on the last convolutional layer are employed in GradCam to generate a heatmap which highlights the most important discriminatory regions (8,16). Heatmaps may be visualized as color maps which may be superimposed on retinal photographs.

Recently, we collected a dataset of retinal images from individuals with diabetes from three Brazilian centers: the Department of Ophthalmology and Visual Sciences and the Diabetes Center, both at the Federal University of Sao Paulo (Sao Paulo State), and also from Hospital de Olhos de Sergipe, in Aracaju (Sergipe State); fundus images were obtained before and after pupil dilation and underwent further analysis with a DL system. In some cases, DL outputs were different for images of the same eye, obtained under slightly different conditions (before and after mydriasis). We believed that, in trying to understand why the different outcomes were produced, we had an opportunity to gain insights into the DL decision process, since a system should ordinarily generate similar outputs for similar inputs (12).

Thus, our objective was to explore qualitative and quantitative discrepancies in GradCam visualization heatmaps between pairs of retinal images from the same eyes, obtained under varying conditions, from a dataset

Highlight box

Key findings

- A successful relationship was established among objective parameters extracted from heatmaps and deep learning output discrepancies.

What is known and what is new?

- Despite high diagnostic accuracies, opaqueness limits clinical application of deep learning algorithms.
- Objective variables extracted from heatmaps helped shed light onto output discrepancies in automatic diabetic retinopathy detection.

What is the implication, and what should change now?

- Explainability is essential for the deployment of artificial intelligence in health, fostering acceptance for all stakeholders. The role of heatmaps for explainability of deep learning system should be emphasized.

composed of fundus images of individuals with diabetes, to gain insights into the DL decision process.

Methods

This retrospective study evaluated a dataset of retinal fundus photographs, consisting of images from 327 patients with diabetes. Each image was composed of 1,600 × 1,600 pixels and was captured using a portable retinal imaging device (Eyer, Phelcom LLC, Boston, MA, USA). Ethical approval for data collection was obtained from the Institutional Research Ethics Committee of the Federal University of Sao Paulo (No. 33,842,220.7.1001.5505) and informed consent was obtained from all individual participants, adhering to the tenets of the Declaration of Helsinki (as revised in 2013). All participating institutions were informed and agreed on the study.

The database had the following frequencies according to DR severity, in the patient-level classification: 44% had no retinopathy, 26.47% had non-proliferative DR, and had 29.31% proliferative DR. All images were anonymized, de-identified, and reviewed to ensure the removal of any personal health information. Written informed consent was obtained from all participants prior to data collection. The details of the dataset and of the sample are described elsewhere. Briefly, patients had a mean age of 57 years (standard deviation 16.82, range, 9–90 years) and 45.3% were men. Race distribution was as follows: 40.7% mixed, 32.7% White, 21.4% Black, 3% Asian and 1% Indigenous (17).

The present study was based on the analyses of pairs of images from the same eye. Each pair was composed of a macula-centered image obtained without pharmacological mydriasis and the respective image acquired after pupil dilation. In order for each pair to be analyzed, both non-mydriatic and mydriatic images had to be deemed gradable as per the expert evaluation (17). Moreover, we included only pairs composed of comparable images, determined through expert evaluation for coincidental framing and the presence of noise and artifacts in the images, such as shadows, over, or underexposure. Pairs with framing differences that omitted a significant part of the image or had shadows or artifacts hindering comparison were excluded.

Automated detection of DR

We employed a DL system which performs image identification into classes, the Diabetic Retinopathy Alteration Score (DRAS), which utilizes a modified version of the EfficientNetV2S convolutional neural network, with distinct input and output parameters, while maintaining the same intermediate convolutional layers. Input was modified to receive images of size 599 × 599 × 3 red, green and blue (RGB) channels. Additionally, the last three layers were dropped, and new layers of convolutional, batch normalization, activation, global average pooling, dense, and output layers were added. The DL system training process allowed its internal parameters to be progressively adjusted to obtain an output from the last layer that closely aligns with the corresponding image class. For training purposes, transfer learning was employed using the EyePACS dataset along with an internal Phelcom dataset comprising 17,330 DR images captured exclusively using the Eyer device (resolution 1,600 × 1,600 × 3 RGB). For validation purposes, 30% of these images were used to periodically evaluate the performance of the network. To add more diversity, data augmentation was applied to images, with rotation, width and height shift, zoom, and brightness values randomly applied. Ground truth data relied on DR severity level classification determined by expert reading, performed independently by two masked, certified ophthalmologists, with a third senior retinal specialist adjudicating in discordant cases (17).

DL analysis was conducted for all images from comparable pairs, both non-mydriatic and mydriatic. The output was represented numerically, indicating the absence or presence of DR, with scores ranging from 0 (absence) to 1 (presence) for each image.

Explainability analysis

We separated all pairs of comparable images into two categories—concordant pairs and discordant pairs—according to the following parameters: DL score (ranging 0 to 1) and DL output (DR presence or absence). Regarding DL score, we arbitrarily chose a score difference of 0.1 as the threshold: pairs were considered "discordant" if the absolute score difference among images within a given pair exceeded 0.1. We named the classification according to DL score difference as "Scenario A". Regarding DL output (presence or absence of DR), pairs in which the output diverged among corresponding images—i.e., the output for non-mydriatic images was different from the respective output of the same eye for the mydriatic image—were considered "discordant". We named the classification according to DL output divergence as "Scenario B".

For objective comparison among concordant and

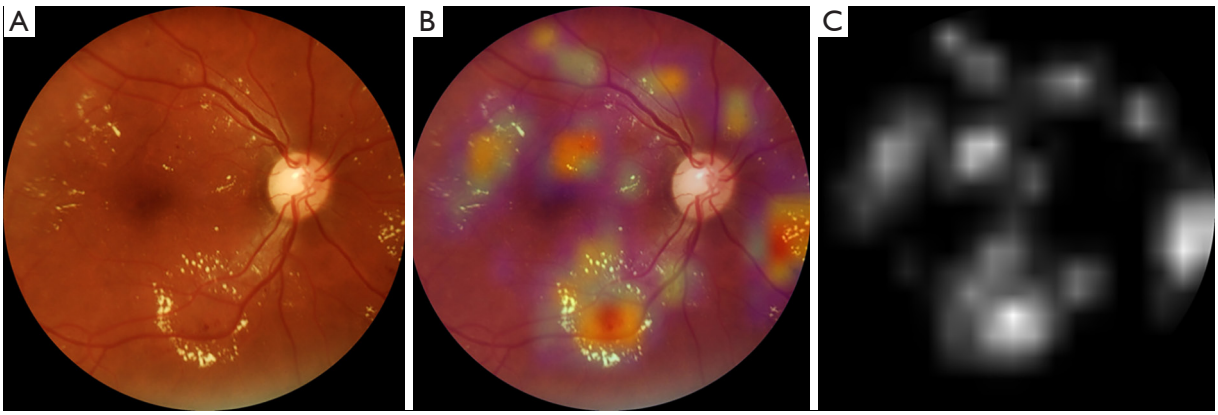

Figure 1 Fundus image showing signs of diabetic retinopathy, with the respective color heatmap and grayscale heatmap. (A) Raw image depicting microaneurysms, hard exudates and retinal hemorrhages; (B) color heatmap combined with raw image; (C) grayscale heatmap. The parameters for this mydriatic image are as follows: sum of activations =133,812,700; number of activated areas =14,653; deep learning score =1.0.

discordant pairs, grayscale heatmaps were automatically generated for each image (*Figure 1*). Heatmap images were divided into 10,000 segments (areas), each pixel within a segment receiving the average score value of the entire segment (ranging 0 to 1); the resulting images were normalized. Two variables were extracted from grayscale heatmaps for objective comparison: (I) the sum of activations of Class Activation Mapping (CAM), corresponding to the summation of every activated pixel in the grayscale heatmap (18); and (II) the number of activated areas, corresponding to the numerical sum of all areas with an average value greater than 0, corresponding to all areas of heatmap activation.

Then, we performed the subtraction of these objective variables for each pair, considering the individual values of images within the pair, in order to obtain (I) the sum of activations difference, corresponding to the mydriatic sum of activations minus the non-mydriatic sum of activations; and (II) the number of activated areas difference, corresponding to the mydriatic number of activated areas minus the non-mydriatic number of activated areas. In addition, we calculated the DL score differences among images from each pair, by performing the subtraction of the mydriatic image DL score minus the non-mydriatic image DL score, as seen in Eq. [1].

$$\text{DL score}_{\text{differences}} = \text{DL score}_{\text{mydriatic}} - \text{DL score}_{\text{non-mydriatic}} \quad [1]$$

This latter calculation was performed only for Scenario B since, by definition, concordant and discordant pairs from Scenario B should necessarily differ regarding DL score. Since our intention was to measure the absolute difference between the parameters of each image within the pairs, absolute results of the subtractions were considered for the comparisons. After obtaining the absolute differences for the above-mentioned parameters, we performed a statistical analysis to determine if concordant pairs were different from discordant pairs according to those parameters.

In addition to the objective comparisons, we also performed a qualitative sub-analysis of heatmaps differences for each image within the discordant pairs. A retina specialist examined whether heatmap-highlighted regions coincided in both images of each given pair, and also identified the heatmap-highlighted areas on the respective raw images, to ascertain if heatmaps accurately corresponded to regions with DR lesions (*Figure 1*). Due to the potential of bias related to subjective evaluation, such qualitative comparison was performed only for discordant pairs.

Statistical analysis

Data were compiled in MS Excel 2010 files (Microsoft Corporation, Redmond, WA, USA). Statistical analyses were performed using IBM SPSS Statistics for Windows, version 29 (IBM Corp., Armonk, NY, USA). The Kolmogorov-Smirnov and Shapiro-Wilk tests were used to check for normal distribution. The Mann-Whitney test was employed for comparison of variables among both groups. A significance level of 5% was applied for all analyses.

Results

After excluding ungradable images, we obtained 351 pairs of

Table 1 Discordant pairs in Scenario A (absolute values)

Patient identification	Laterality	Difference of sums of activations	Difference of number of activated areas	Score difference
20	RE	11,546,300	3,398	0.43
20	LE	2,778,800	305	0.29
40	RE	6,366,200	1,370	0.21
67	LE	5,556,200	1,094	0.12
70	LE	4,607,800	64	0.62
75	RE	15,403,500	2,368	0.41
111	RE	8,998,400	977	0.37
132	RE	5,973,200	1,055	0.5
136	RE	1,321,400	1,087	0.19
168	RE	10,836,200	2,077	0.65
209	RE	13,947,400	5,074	0.34
224	LE	23,224,300	4,108	0.81
226	RE	9,814,400	1,247	0.37
226	LE	11,049,800	4,094	0.88
229	LE	13,937,200	2,142	0.43
282	RE	13,572,000	1,891	0.67
298	LE	10,181,900	1,627	0.87
300	LE	6,852,400	667	0.12
305	RE	15,929,400	2,332	0.92
321	RE	11,916,200	3,968	0.85

RE, right eye; LE, left eye.

images (351 eyes); algorithmic performance for DR detection at the image level was as follows: sensitivity 89.78%, specificity 96.26%, area under the receiver operating characteristic (ROC) curve 0.952, for mydriatic images.

Out of those 351 pairs of images, 210 pairs (420 images) were considered comparable, after qualitative assessment by a retinal specialist (F.K.M.) who determined which pairs had images with equivalent framing and/or coincident areas of relevance, not compromised by shadows or artifacts that would preclude comparison with the respective image within the pair.

Discordant pairs analysis

In Scenario A, a total of 20 (out of 210) eyes showed a DL score difference greater than 0.1 among scores of non-mydriatic and mydriatic images. The comparison of those 20 discordant pairs versus the remaining 190 pairs was performed regarding the sum of activations and the number of activated areas; averages were compared among groups. For both variables tested, differences were statistically significant: the sum of activations ($P<0.001$) and the number of activated areas ($P<0.001$). The differences in the analyzed variables for discordant cases in Scenario A are displayed in *Table 1*; no linear relationship was found among the values of different parameters for each case.

Table 2 Discordant pairs in Scenario B (absolute values)

Patient identification	Laterality	Difference of sums of activations	Difference of number of activated areas	Score difference
74	RE	8,064,500	237	0.01
75	RE	15,403,500	2,368	0.41
111	RE	8,998,400	977	0.37
209	RE	13,947,400	5,074	0.34
224	RE	1,622,100	539	0.04
224	LE	23,224,300	4,108	0.81
229	LE	13,937,200	2,142	0.43
300	RE	16,917,600	2,069	0.08
300	LE	6,852,400	667	0.12
321	RE	11,916,200	3,968	0.85

RE, right eye; LE, left eye.

In Scenario B, a total of 10 (out of 210) eyes showed different DL outputs considering non-mydriatic and mydriatic images within the same pair. The comparison of those 10 discordant pairs versus the remaining 200 pairs was performed regarding the sum of activations, the number of activated areas and the DL score; averages were compared between groups. For all variables tested, groups were significantly different: the DL score ($P<0.001$), the sum of activations ($P<0.001$) and the number of activated areas ($P=0.005$). The differences in the analyzed variables for discordant cases in Scenario B are displayed in *Table 2*; no linear relationship was found among the values of different parameters for each case; DL output difference occurred even in cases where DL score differences were smaller than 0.1.

Qualitative sub-analysis

In all discordant cases, qualitative expert analysis pointed

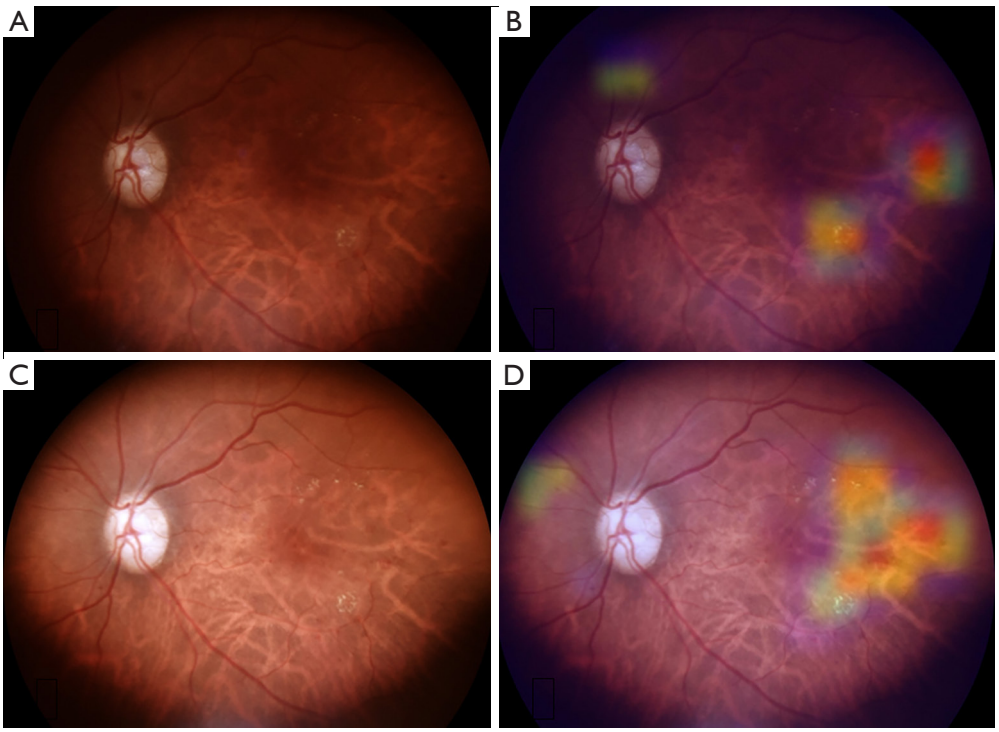

Figure 2 Example of a discordant pair of retinal images. (A) Non-mydriatic fundus image of the left eye of a patient with diabetic retinopathy. (B) Color heatmap combined with non-mydriatic image (deep learning score =0). (C) Mydriatic fundus image of the same eye. (D) Color heatmap combined with mydriatic image (deep learning score =0.43). This example displays a larger area of activation in the mydriatic image. In this case, the absolute numerical differences of parameters between non-mydriatic and mydriatic images were as follows: difference of sums of activations =13,937,200; difference of number of activated areas =2,142; score difference =0.43. Raw images depict microaneurysms, hard exudates and retinal hemorrhages.

to subtle differences in image quality within each pair, related to focus and image sharpness. Of note, all images were considered gradable as per the inclusion criteria. In addition, not every AI-highlighted area corresponded to a DR lesion in some images; that was the case in 15 images from Scenario A (out of a total of 40 images) and also in 10 images from Scenario B (out of a total of 20 images). In some cases, there was no coincidence of highlighted areas in the respective images within a pair; that was observed in 3 cases from Scenario A and 5 cases from Scenario B. Two discordant cases are displayed in *Figures 2,3*.

Discussion

Comparison among discordant and concordant groups, using objective variables extracted from heatmaps, pointed to significant differences among groups for both the sum of activations and the number of activated areas. Our findings point to the consistent role of GradCam heatmaps in explaining DL outputs, thus helping to shed light on the algorithmic decision process. Of note, the selected pairs corresponded to exceptions in a system with a considerably high performance: out of a sample of 210 eyes, in Scenario A we found only 20 discordant cases, while in Scenario B we found only 10 discordant cases. *Figure 4* depicts an example of a concordant pair. We believe that subtle differences in image quality, likely related to image focus, sharpness, or illumination conditions, and sometimes corresponding to a few pixels, may have occurred in images obtained under slightly different conditions, resulting in different DL outputs in a small fraction of the studied dataset.

Understanding DL systems' decisions is important for clinical decision support (14), as trust is often grounded in the system's ability to generate consistent results (13). Explainability also fosters accountability and enhances physicians' autonomy, enabling them to evaluate the decisions

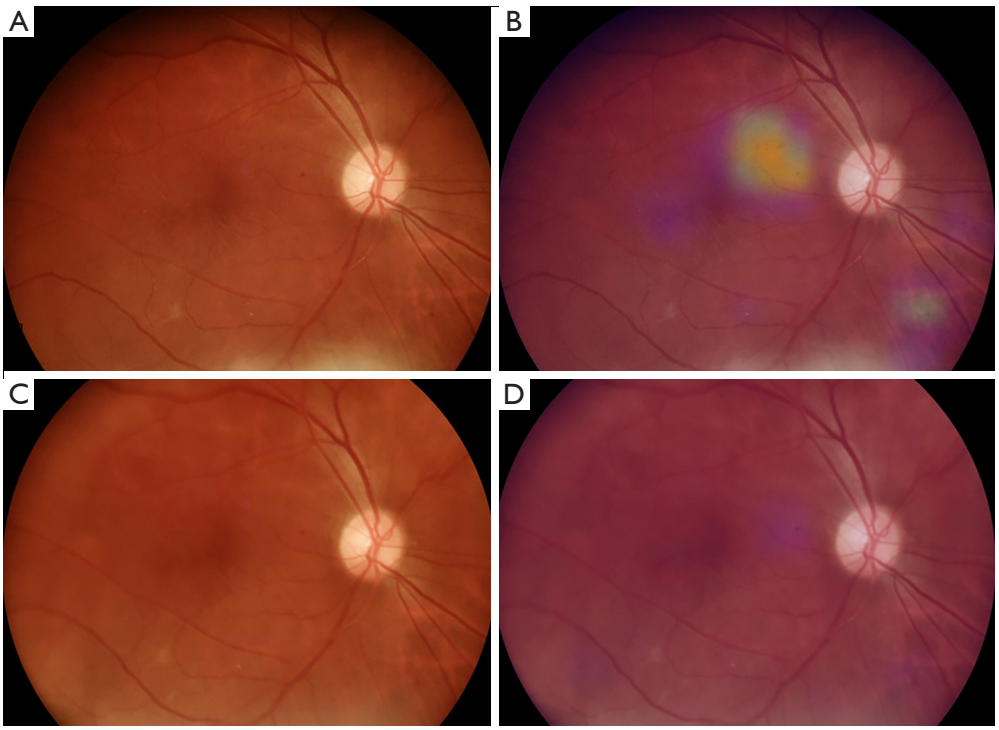

Figure 3 Example of a discordant pair of retinal images. (A) Non-mydriatic fundus image of the right eye of a patient with diabetic retinopathy. (B) Color heatmap combined with non-mydriatic image (deep learning score =0.85). (C) Mydriatic fundus image of the same eye. (D) Color heatmap combined with mydriatic image (deep learning score =0). This example displays a larger area of activation in the non-mydriatic image. In this case, the absolute numerical differences of parameters between non-mydriatic and mydriatic images were as follows: difference of sums of activations =11,916,200; difference of number of activated areas =3,968; score difference =0.85. Raw images depict microaneurysms, hard exudates and retinal hemorrhages.

merits (13,19). On the other hand, opaqueness of clinical support systems may limit capabilities of clinicians (19). Explainability also helps build trust, enabling the resolution of disagreements between AI systems and human experts, no matter on whose side the error in judgment is situated. Moreover, explainability allows clinicians to verify whether the system's parameters are aligned with clinical perspectives, enhancing the adherence to medical standards (19). From the patient's perspective, explainability may contribute as educational tools, promoting more accurate risk perceptions and motivating their engagement in shared decision-making (19). Explainability may also assure developers about the fairness of AI models, allowing them to perceive if performance is based on meta-data rather than the data itself, and possibly allowing developers to identify such types of errors before the systems go into validation and certification processes, saving time and development costs (19).

DL systems are not programmed into models that reflect the causal structure of the problem to be solved; instead, these systems learn from a large set of data: neurons activate when certain features are detected, and the system's output assigns a probability of the outcome (13). However, the relationship between features and output can be indirect and, sometimes, fragile; saliency maps may point to which pixels or regions seem to play a role in the decision process without telling us precisely how (5,13). Even though proposed methods such as GradCam can help humans understand the decision process of DL systems, such methods will hardly render automated systems totally transparent, due to the very nature of DL systems: sometimes, even small permutations in seemingly unrelated aspects of data can lead to a significantly different weighting of features (13).

Although explainability in computer vision has largely relied on the identification of the regions of importance, techniques based on saliency maps are not the only methods

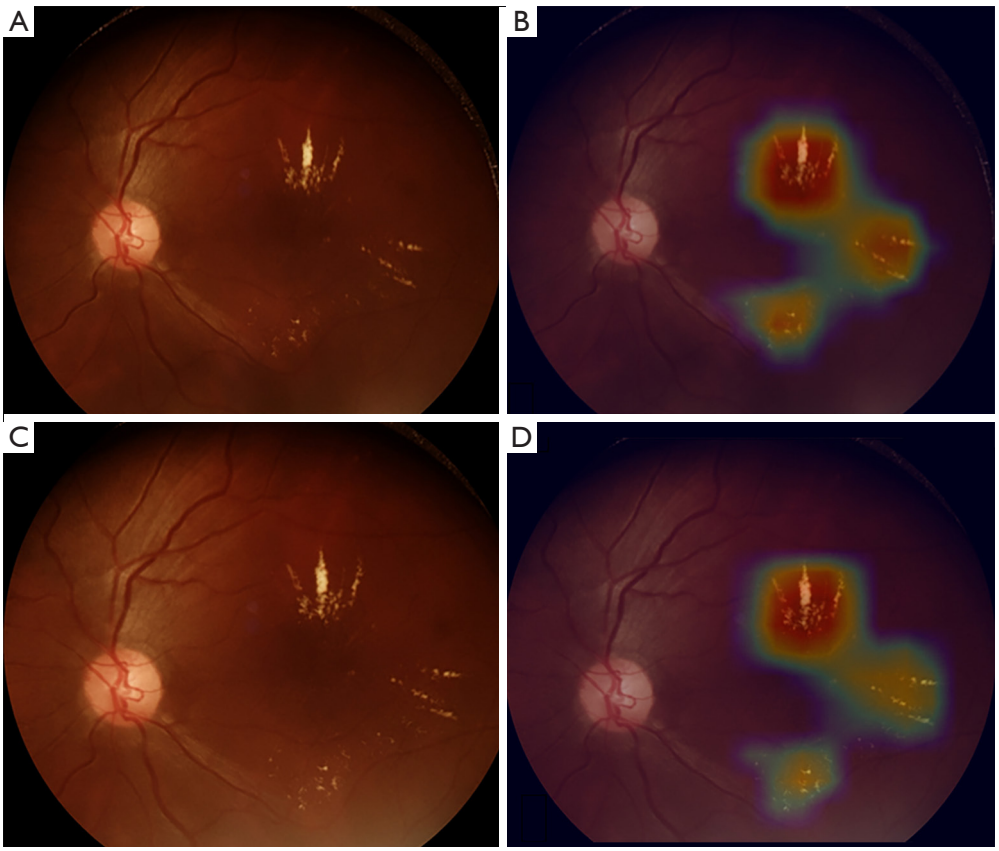

Figure 4 Example of a concordant pair of retinal images. (A) Non-mydriatic fundus image of the right eye of a patient with diabetic retinopathy. (B) Color heatmap combined with non-mydriatic image. (C) Mydriatic fundus image of the same eye. (D) Color heatmap combined with mydriatic image. In this example, the differences in the areas of activation are minimal, as they highlight hard exudates, which are surrogate markers of diabetic macular edema, both in the non-mydriatic and mydriatic images. The deep learning score difference for this case was zero.

developed for gaining insight into the decision processes of such systems; GradCam methods even bear some criticism, with previous studies having found considerable disagreement between heatmap-highlighted regions and expert annotations (5,6,8,14,20). Since feature extraction occurs at deep layers of a DL system, where the image resolution is smaller than the original input resolution, it is possible that, when the information from a deep layer is projected back onto the input image, the granularity is decreased, resulting in coarse heatmaps (6). Other approaches for explainability include linear proxy models, decision trees, automatic rule-extraction, adversarial examples techniques, presegmentation, segmentation, categorization of lesions, and pixel-level classification (5,6,11,14).

Recently, several publications have reported the performance of AI systems for DR evaluation using portable retinal cameras and yielding variable outcomes, including the detection of any DR, referable DR and sight-threatening DR (21-24): Lupidi and colleagues have reported a 96.8% sensitivity and 96.8% specificity for the detection of any DR, using the Optomed Aurora™ (Oulu, Finland) camera and the Selena+™ (Singapore) system; their sample consisted of 256 patients with diabetes, half of whom had DR (21). Ruan and colleagues have reported an 88.2% sensitivity and a 40.7% specificity in identifying referable forms of DR, using the Optomed Aurora™ camera and the Phoebus™ (Shanghai, China) AI system; their sample consisted of 315 patients with diabetes, and the sample composition regarding DR classification was not informed (22). In the study by Rajalakshmi and colleagues, performed with the Remidio™ (Bangalore, India) camera and the EyeArt™ (Woodland Hills, USA) system, the AI software showed a

95.8% sensitivity and 80.2% specificity for detecting any DR, as well as a 99.1% sensitivity and 80.4% specificity in detecting sight-threatening DR on a sample of 296 patients with diabetes, 65% of whom presenting DR (23). A prospective, multicenter study was conducted in a real-world community DR screening in India and obtained a large dataset: from a pool of 60,633 retinal fundus images, a total of 29,656 images from 11,199 patients were eligible for the study authored by Nunez do Rio and colleagues (24). The images were captured with the Zeiss Visuscout™ (Jena, Germany) camera and analysed with the Zeiss VISUHEALTH-AI DR™ (Singapore) system for the detection of referable DR; a 72.08% sensitivity and 85.65% specificity were reached; the vast majority of patients (80.2%) was classified as non-referable, with only 3.8% referable and an ungradable rate of 16.0% (24). Possible reasons for the heterogeneity of performances are individual cameras' and AI systems' characteristics, different study designs, uneven sample sizes, and variable datasets composition.

Among the present study's strengths are the high performance of the DL system, which led to the rare discordant outputs; the robust dataset; and the established relationship between objective variables extracted from GradCam heatmaps and the differences among discordant and concordant groups, a finding which helped shed light onto the decision process of the DL system. As for the study limitations, since it was based only on the AI modality of image recognition, its results and conclusions are not applicable to other modalities of AI in healthcare, such as speech recognition and natural language processing. In addition, even though objective, the criterion for Scenario A was arbitrary. Another limitation is related to the qualitative sub-analysis, which may have been intrinsically biased due to subjectivity.

Finally, we believe the very idea of explainability of DL systems comes with an intrinsic limitation: as good as any explainability method may be, we should always consider that human reasoning is different from artificial systems, and this reality is also reflected in the essential differences among human and computer visions. As we have seen in published studies, sometimes computer vision sees what no human can see, as in the example of DL systems that predicted the sex of the individual from retinal fundus photographs (25,26).

Conclusions

In conclusion, the analytical process involving objective variables extracted from heatmap analysis has provided valuable insights on the reasons for different outputs in discordant pairs, helping to provide glimpses into DL decisions. The successfully established relationship among those parameters and the output discrepancies reinforces the role of heatmaps in contributing to the explainability of DL systems. Future research should address other challenges for the deployment of AI in healthcare, in order to harness the full benefits of AI in health, as well as techniques to further increase the performance of image classification, such as test-time augmentation (27). Future research should also evaluate explainability of other AI modalities besides computer vision, such as speech recognition and natural language processing.

Acknowledgments

Funding: None.

Footnote

Data Sharing Statement: Available at https://atm.amegroups.com/article/view/10.21037/atm-24-73/dss

Peer Review File: Available at https://atm.amegroups.com/article/view/10.21037/atm-24-73/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-24-73/coif). F.K.M. is a consultant for Phelcom Technologies, and has received consulting fees for medical advice and travel support for attending meetings from Phelcom Technologies. P.P. and F.Y. are employees at Phelcom Technologies. G.B.M. has received consulting fees and travel expenses from SJJ Solutions, and has been part of advisory boards for Bayer, Roche, and West Pharmaceuticals. J.A.S. is a founding partner and CEO of Phelcom Technologies, and owns stock or stock options of Phelcom Technologies. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Research Ethics Committee of the Federal

University of Sao Paulo (No. 33,842,220.7.1001.5505) and informed consent was obtained from all individual participants. All participating institutions were informed and agreed on the study.

References

1. Beam AL, Drazen JM, Kohane IS, et al. Artificial Intelligence in Medicine. N Engl J Med 2023;388:1220-1.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
3. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-31.
4. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24-9.
5. Quellec G, Al Hajj H, Lamard M, et al. ExplAIn: Explanatory artificial intelligence for diabetic retinopathy diagnosis. Med Image Anal 2021;72:102118.
6. Andersen JKH, Hubel MS, Rasmussen ML, et al. Automatic Detection of Abnormalities and Grading of Diabetic Retinopathy in 6-Field Retinal Images: Integration of Segmentation Into Classification. Transl Vis Sci Technol 2022;11:19.
7. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye (Lond) 2020;34:451-60.
8. Malerbi FK, Andrade RE, Morales PH, et al. Diabetic Retinopathy Screening Using Artificial Intelligence and Handheld Smartphone-Based Retinal Camera. J Diabetes Sci Technol 2022;16:716-23.
9. Lim JI, Regillo CD, Sadda SR, et al. Artificial Intelligence Detection of Diabetic Retinopathy: Subgroup Comparison of the EyeArt System with Ophthalmologists' Dilated Examinations. Ophthalmol Sci 2023;3:100228.
10. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. Nat Med 2022;28:31-8.
11. Gilpin LH, Bau D, Yuan BZ, et al. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy; 2018:80-9.
12. Babic B, Gerke S, Evgeniou T, et al. Beware explanations from AI in health care. Science 2021;373:284-6.
13. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep 2019;49:15-21.
14. Chang J, Lee J, Ha A, et al. Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples. Ophthalmology 2021;128:78-88.
15. Yousefi S, Pasquale LR, Boland MV. Artificial Intelligence and Glaucoma: Illuminating the Black Box. Ophthalmol Glaucoma 2020;3:311-3.
16. Selvaraju RR, Cogswell M, Das A, et al. Grad-Cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017:618-26. Available online: https://arxiv.org/abs/1610.02391. Accessed February 2, 2024.
17. de Oliveira JAE, Nakayama LF, Zago Ribeiro L, et al. Clinical validation of a smartphone-based retinal camera for diabetic retinopathy screening. Acta Diabetol 2023;60:1075-81.
18. Zhou B, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization. Available online: https://doi.org/10.48550/arXiv.1512.04150. Accessed April 05, 2024.
19. Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20:310.
20. Van Craenendonck T, Elen B, Gerrits N, et al. Systematic Comparison of Heatmapping Techniques in Deep Learning in the Context of Diabetic Retinopathy Lesion Detection. Transl Vis Sci Technol 2020;9:64.
21. Lupidi M, Danieli L, Fruttini D, et al. Artificial intelligence in diabetic retinopathy screening: clinical assessment using handheld fundus camera in a real-life setting. Acta Diabetol 2023;60:1083-8.
22. Ruan S, Liu Y, Hu WT, et al. A new handheld fundus camera combined with visual artificial intelligence facilitates diabetic retinopathy screening. Int J Ophthalmol 2022;15:620-7.
23. Rajalakshmi R, Subashini R, Anjana RM, et al. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye (Lond) 2018;32:1138-44.
24. Nunez do Rio JM, Nderitu P, Bergeles C, et al. Evaluating

a Deep Learning Diabetic Retinopathy Grading System Developed on Mydriatic Retinal Images When Applied to Non-Mydriatic Community Screening. J Clin Med 2022;11:614.
25. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. Sci Rep 2021;11:10286.
26. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158-64.
27. Seth P, Khan A, Gupta A, et al. UATTA-ENS: Uncertainty Aware Test Time Augmented Ensemble for PIRC Diabetic Retinopathy Detection. 2022. arXiv:2211.03148v2. Available online: https://doi.org/10.48550/arXiv.2211.03148

Cite this article as: Malerbi FK, Nakayama LF, Prado P, Yamanaka F, Melo GB, Regatieri CV, Stuchi JA. Heatmap analysis for artificial intelligence explainability in diabetic retinopathy detection: illuminating the rationale of deep learning decisions. Ann Transl Med 2024;12(5):89. doi: 10.21037/atm-24-73