Peer Review File

Article information: https://dx.doi.org/10.21037/atm-24-73

Reviewer A

The paper offers a valuable contribution by aiding the interpretation of diabetic retinopathy (DR) through heatmap analysis in deep learning. However, there are several aspects where improvements could be made to increase clarity and comprehension.

The paper makes an interesting contribution by supporting the interpretation of diabetic retinopathy (DR) using heatmap analysis in deep learning. However, there are several areas where the paper can be improved to enhance clarity and understanding:

Comment 1-Binary Classification and Explainable Images: The binary classification for DR detection is impressive. Nevertheless, the authors show only a few examples of explainable images with specific deep learning (DL) scores. Please consider adding the DL scores to Figures 1 to 3 for better illustration.

Reply 1: Thank you for your comment. We have added numerical parameters to the captions of Figures 1-3. We have also included a new Figure (4) depicting an example of a concordant pair, for better illustration.

Changes in the text: We have included the requested information at Figures 1-3 captions:

Figure 1 Captions - 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 = 133812700; number of activated areas = 14653; Deep Learning score = 1.0.

Figure 2 Captions - 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 = 13937200; difference of number of activated areas = 2142; score difference = 0.43. Raw images depict microaneurysms, hard exudates and retinal hemorrhages.

Figure 3 Captions - 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 =11916200; difference of number of activated areas =3968; score difference = 0.85. Raw images depict microaneurysms, hard exudates and retinal hemorrhages.

We have also included a new Figure (4), along with the respective caption (please see our answer to Comment 3 from Reviewer B, below).

Comment 2: Explanation of Results Grouping in Section 2.2: For researchers who may not have a background in DR, it would be helpful if the authors explain the benefits of grouping results into two scenarios in Section 2.2, both from a medical and technical perspective.

Reply 2: Thank you for the opportunity to better clarify this issue. Our intention was to perform experiments which could evaluate whether heatmaps were an adequate tool for the explainability of the DL decision-making process. In that sense, we have separated the discordant pairs according to two different scenarios, as stated in the Methods section. Both scenarios designed for this study corresponded to situations in which each pair of correspondent images had discordant DL results: a DL score greater than 0.1 in Scenario A and different DL outputs in Scenario B. The benefit of grouping results as such was that the separation allowed for discordance analysis, which helped answer the main research question of this article. For those who are not familiar with DR, the DL output was the presence (DL output: positive) or absence (DL output: negative) of diabetic retinopathy, after automated evaluation of fundus photographs.

Changes in the text: We have modified our text in section 2.2, in order to include this information **(see page 7, lines 205 - 208):**

Regarding DL output (presence or absence of DR), pairs in which the output diverged among corresponding images – i.e. the ouput for non-mydriatic images was different from the respective output of the same eye for the mydriatic image - were considered "discordant".

Comment 3: Support for Arbitrary Score Difference: Please provide strong justification for choosing an arbitrary score difference of 0.1.

Reply 3: Since our objective was to evaluate discordant pairs, we noticed that, if only pairs with discordant outputs were to be compared, we would have analyzed only 10 eyes. In order to enhance our analyzed sample, we decided to include also pairs of images with DL score differences (Scenario A), in addition to pairs of images with

discordant DL outputs (Scenario B). Our choice of a 0.1 threshold for DL score difference provided us with twice that amount of eyes (20 eyes), in comparison to Scenario B, amounting to a reasonable amount of eyes to analyze.

Since the vast majority of pairs of images from the same eyes had no DL score difference (as would be expected, since images were obtained from the very same tissues), we arbitrarily chose a threshold to further separate the pairs of images, hence obtaining a larger number of pairs of images to conduct our experiments.

Considering the 210 eyes with comparable images, 169 eyes had a DL score difference equal to zero. As for the 41 eyes with greater-than-zero absolute DL score differences among the respective images, by applying the chosen criteria of a DL score difference greater than 0.1, and considering that scores with differences smaller than 0.1 would be arbitrarily close to each other, we have reached a sample of 20 eyes, as stated in our results section. Even though the choice for such score difference value was arbitrary, as would have happened with any other figure for that matter, we chose to employ objective parameters for separating comparable eyes into both Scenarios: namely, DL score differences and discordant DL outputs - in order to avoid subjectivity bias. Of note, the criteria set for Scenario B (discordant DL outputs) was not arbitrary: the DL outputs indicate the very results of the DL system, with practical implications.

Thus, the reason for creating an arbitrary threshold was that it allowed us to expand the sample studied in our experiments. We have included this point as a limitation in the Discussion section.

Changes in the text: "we have modified our text as advised **(see Page 13, line 382)**".

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.

Comment 4: Equation for Calculating Score Differences: In Section 2.2, consider providing an equation to calculate the score differences for better clarity.

Reply 4: Thank you for your suggestion. We have provided the equation as suggested.

Changes in the text: We have modified our text as advised **(see Page 8, lines 228-229)**.

Modified text in the revised version:

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 (1).

$$\text{DL score}_{\text{mydriatic}} - \text{DL score}_{\text{non-mydriatic}}. \quad (1)$$

Comment 5: Recheck P-value in Line 207: Please recheck the p-value reported in Line 207 for accuracy.

Reply 5: The results of the Mann-Whitney tests for both comparisons performed for Scenario A corresponded to a p value of 0.000. We have corrected the presentation of the value by replacing a comma with a period.

Changes in the text: We have modified our text as advised **(see Page 9, line 269)**:

For both variables tested, differences were statistically significantly: the sum of activations (p=0.000) and the number of activated areas (p=0.000).

Comment 6: Comparison in Table 1: In Table 1, patient ID 70 has a small sum of activations and activated area with a score difference of 0.62, while patient ID 282 has a large sum of activations and activated area with a score difference of 0.67. The authors should consider explaining the comparison between these two patients in greater detail.

Reply 6: Thank you for raising this point. Table 1 displays the results of our experiments regarding Scenario A, and no linear relationship was found among the values of different parameters for each case. In that sense, the comparison among such patients based solely on those values would not provide clinically meaningful information.

We have added in the results section a sentence explaining that no linear relation was found among the values of those objective parameters, as displayed in Table 1.

Changes in the text: We have added the following line in the Results section **(see Page 9, lines 270-271)**:

"no linear relationship was found among the values of different parameters for each case."

Comment 7: Comparison in Table 2: Similarly, in Table 2, patient ID 74 and patient ID 300 both have a score difference of less than 0.1. Please provide a detailed comparison between these two patients.

Reply 7: The criteria for Scenario B corresponded to a different DL output among images from the same pair, regardless of the DL score difference. That is the reason why, differently from Scenario A, some patients with DL score differences smaller than 0.1 were included. Similarly, as we pointed out in the response to comment 8, no linear relationship was found among the values of different parameters for each case in Scenario B, either. In that sense, the comparison among such patients based solely on those values would not provide clinically meaningful information.

We have also added in the results section a sentence explaining that no linear relation was found among the values of those objective parameters, as displayed in Table 2, and that DL output difference occurred even in cases where DL score differences were smaller than 0.1 (cases 74 RE and 300 RE, which presented a positive DL output for the non-mydriatic images while attaining negative DL outputs for mydriatic images). Additionally, case 224 RE also fell into this category (also presenting a positive DL output for non-mydriatic image and a negative DL output for the mydriatic image). In the manuscript, we have emphasized that these kinds of situations corresponded to exceptions (10 cases out of 210) since, in the majority of cases, DL outputs were concordant for pairs of images comprising non-mydriatic and mydriatic images.

Changes in the text: We have added the following line in the Results section **(see Pages 9-10, lines 280 - 282)**:

"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."

Comment 8: Adding Scores to Figures 1 to 3: Kindly consider adding all scores from Tables 1 and 2 to the captions of Figures 1 to 3. Additionally, incorporate insights from the activation areas of the heatmap according to medical knowledge to enhance the interpretation of DL results.

Reply 8: Thank you for the comment. We have added the scores as suggested, as well as incorporated insights from activated heatmap areas to the captions of Figures 1-3. Please see our answer to Comment 1 for indication of the changes in the text.

Reviewer B

this papers studies the effect of variance in image acquisition of patient eyes to determine DR from CNN based classifiers using class activation maps. The paper concludes with observations regarding the significance of heatmaps to determine the

robustness of a CNN based classifier.

Comments-

Comment 1: The base model is finetune on limited data and no comparison has been showed regarding the diagnostic accuracy of the classifier with respect to other published classifier for similar applications. It is recommended the authors shows some comparison to published results from other works for DR classification.

Reply 1: Thank you for your comment. We agree that putting our results into the context of recently published studies will contribute to a more comprehensive article. We have added a new paragraph in the Discussion section, as well as some additional references.

Changes in the text: We have added the following paragraph in the Discussion section **(see Page 12, lines 350 - 373)**:

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 ™ camera and the Selena + ™ 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 ™ 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 ™ camera and the EyeArt ™ 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 Nunes do Rio and colleagues (24). The images were captured with the Zeiss Visuscout ™ camera and

analysed with the Zeiss VISUHEALTH-AI DR ™ 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 ungradability 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.

We have also added the following references:

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; 8:1–6.

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(4):620-627.

23-Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. *Eye (Lond).* 2018;32(6):1138-1144.

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(3):614.

Comment 2: Since this is a black box based approach, it is difficult to understand the motive of this paper. Are the authors trying to show using black box approaches for clinical decision making is not the right choice based on these results or are the authors suggesting this should be one of the ways to analyze robustness of black box approaches please clarify.

Reply 2: Thank you for your comment. The objective of this research, as stated in the abstract, was to "explore discrepancies in heatmaps (…) to gain insights into the deep learning (DL) decision process." Our experiments´ results led to the conclusion present also in the abstract: "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".

Hence, we wrote in the Conclusion section that "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."

In that sense, we have concluded that heatmaps are valuable tools to enhance explainability of DL systems. Such message was emphasized also in the following passages:

- Highlight Box: "A successful relationship was established among objective parameters extracted from heatmaps and deep learning output discrepancies."; "Objective variables extracted from heatmaps helped shed light onto output discrepancies in automatic diabetic retinopathy detection."; "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."

- Discussion: "Our findings point to the consistent role of GradCam heatmaps in explaining DL outputs, thus helping to shed light on the algorithmic decision process."

We have added a sentence in the last paragraph of the introduction section, in order to reinforce the motive of our research.

Changes in the text: We have added a sentence in the last paragraph of the introduction section **(see Page 5, line 149)**:

"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 composed of fundus images of individuals with diabetes, to gain insights into the deep learning (DL) decision process.

Comment 3: The visual results are minimal.

Reply 3: Thank you for raising this point. We have added a new figure (Figure 4), with the respective caption, showing the comparison of heatmaps in a concordant pair.

Changes in the text: We have added the following sentence in order to present the new Figure: **(see Page 10, line 302)**":

"Figure 4 depicts an example of a concordant pair"

Figure 4 Title - Example of a concordant pair of retinal images.
Figure 4 Captions - 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.

Comment 4: The major flaw with this paper is there is no biomarker based truth to compare the heatmaps objectively. The activation does not hold much significance unless the activation is happening in the region which is well defined to have a biomarker. Just looking at differences does not provide any clinical observations.

Reply 4: Thank you for your comment. Even though the automatic system reported herein was designed and trained for identification of classes, as opposed to specific lesions or biomarkers, in general, activation areas corresponded to clinical lesions, as seen both in Figure 1 and the new Figure 4, which display typical examples where the heatmap highlights clinically significant lesions. Our experiments explored differences of objective parameters, such as the sum of activations, the number of activated areas and deep learning scores. For the experiments, we did not explore differences among pairs of images related to clinical observations of biomarkers, as clinical observations from the reading center corresponded to the classification of diabetic retinopathy into severity levels by expert readers, and such labeling was used for the creation of the ground truth.

In that sense, cases displayed in Figures 2 and 3, along with all the other discordant pairs explored in our experiments, are exceptions to an overall high diagnostic performance of the algorithm (sensitivity 89.78%, specificity 96.26%, area under the ROC curve 0.952).

In order to emphasize that the automatic system was designed and trained for image identification into classes, we have modified some sentences in the Methods section, subsection 2.1 Automated Detection of DR

Changes in the text:

"We employed a DL system which performs image identification into classes, the Diabetic Retinopathy Alteration Score (DRAS)" **(see Page 6, line 178)**; and

"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. ” **(see Page 7, line 193)**

Comment 5: Lastly, the authors should show how inference time augmentation will effect the performance of the base classifier - seems like inference time augmentation should increase accuracy and then the authors can analyze all the heatmaps generated from all the augmentations during inference to provide a substantiated observations

Reply 5: Thank you for the comment. We agree that inference time augmentation, or test-time augmentation, is a technique that could improve the performance of the system. However, that would fall beyond the scope of the present study, since our objective was to gain insights into the DL decision process. We have included a new sentence at the conclusion paragraph, stating that future research should address techniques to further increase the performance of image classification, such as test-time augmentation.

Changes in the text: We have modified our conclusion paragraph **(see Page 13, lines 401-402)**:

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 (26). Future research should also evaluate explainability of other AI modalities besides computer vision, such as speech recognition and natural language processing.

New reference:

26 - Seth P, Gupta A, Mishra S, Bhandhari A. UATTA-ENS: Uncertainty Aware Test Time Augmented Ensemble for PIRC Diabetic Retinopathy Detection. 2022. arXiv:2211.03148v2
https://doi.org/10.48550/arXiv.2211.03148