DeepAlienorNet: A deep learning model to extract clinical features from colour fundus photography in age-related macular degeneration

Alexis Mathieu, Soufiane Ajana, Jean-Francois Korobelnik, Melanie Le Goff, Brigitte Gontier, Marie-Benedicte Rougier, Cecile Delcourt, Marie-Noelle Delyfer

Received: 5 December 2023 | Accepted: 25 January 2024

DOI: 10.1111/aos.16660

Acta Ophthalmologica

ORIGINAL ARTICLE

DeepAlienorNet: A deep learning model to extract clinical features from colour fundus photography in age-related macular degeneration

Alexis Mathieu[1,2] | **Soufiane Ajana**[1] | **Jean-François Korobelnik**[1,2] | **Mélanie Le Goff**[1] | **Brigitte Gontier**[2] | **Marie-Bénédicte Rougier**[2] | **Cécile Delcourt**[1,2] | **Marie-Noëlle Delyfer**[1,2,3]

[1]Inserm, Bordeaux Population Health Research Center, UMR 1219, University of Bordeaux, Bordeaux, France

[2]Service d'Ophtalmologie, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France

[3]FRCRnet/FCRIN Network, Bordeaux, France

Correspondence
Marie-Noëlle Delyfer, Service d'ophtalmologie du CHU de Bordeaux, Hôpital Pellegrin, Place Amélie Raba Léon, 33000 Bordeaux, France.
Email: marie-noelle.delyfer@chu-bordeaux.fr

Funding information
Ministère de la Santé, Grant/Award Number: PHRC 2012 and PHRC12_157 ECLAIR; Agence Nationale de la Recherche, Grant/Award Number: ANR 2010-PRSP-011 VISA; CFSR Recherche (Club Francophone des Spécialistes de la Rétine); University of Bordeaux; Fondation Voir et Entendre; Théa Pharma; CNSA (Caisse Nationale pour la Solidarité et l'Autonomie)

Abstract

Objective: This study aimed to develop a deep learning (DL) model, named 'DeepAlienorNet', to automatically extract clinical signs of age-related macular degeneration (AMD) from colour fundus photography (CFP).

Methods and Analysis: The ALIENOR Study is a cohort of French individuals 77 years of age or older. A multi-label DL model was developed to grade the presence of 7 clinical signs: large soft drusen (>125 μm), intermediate soft (63–125 μm), large area of soft drusen (total area >500 μm), presence of central soft drusen (large or intermediate), hyperpigmentation, hypopigmentation, and advanced AMD (defined as neovascular or atrophic AMD). Prediction performances were evaluated using cross-validation and the expert human interpretation of the clinical signs as the ground truth.

Results: A total of 1178 images were included in the study. Averaging the 7 clinical signs' detection performances, DeepAlienorNet achieved an overall sensitivity, specificity, and AUROC of 0.77, 0.83, and 0.87, respectively. The model demonstrated particularly strong performance in predicting advanced AMD and large areas of soft drusen. It can also generate heatmaps, highlighting the relevant image areas for interpretation.

Conclusion: DeepAlienorNet demonstrates promising performance in automatically identifying clinical signs of AMD from CFP, offering several notable advantages. Its high interpretability reduces the black box effect, addressing ethical concerns. Additionally, the model can be easily integrated to automate well-established and validated AMD progression scores, and the user-friendly interface further enhances its usability. The main value of DeepAlienorNet lies in its ability to assist in precise severity scoring for further adapted AMD management, all while preserving interpretability.

KEYWORDS
artificial intelligence, deep learning, fundus photography, image interpretation, computer-assisted, macular degeneration

1 | INTRODUCTION

Age-related macular degeneration (AMD) is a leading cause of blindness worldwide, affecting the macula and causing central vision loss as it progresses (Fleckenstein et al., 2021). With the ageing population, the burden of AMD on healthcare systems is expected to increase. By 2040, it is projected that 288 million people will be affected by AMD globally (Wong et al., 2014). The economic burden of visual impairment due to AMD is

Cécile Delcourt and Marie-Noëlle Delyfer contributed equally to this work.

substantial, accounting for $343 billion and constituting 12% of the total global cost of visual impairment. This percentage can be even higher in developed regions, reaching up to 30% (AMD Alliance International, 2010). Deep learning (DL) models offer a promising solution to support physicians and eye care systems in managing advanced AMD. By leveraging DL technology, healthcare professionals can enhance their ability to diagnose and manage AMD, ultimately improving patient outcomes and alleviating the burden on healthcare systems.

Early signs of AMD typically involve the presence of macular drusen and/or pigmentary epithelium alterations, often detected during routine fundus imaging.

The two advanced stages of AMD responsible for most of the visual loss are exudative AMD and atrophic AMD (geographic atrophy). In exudative AMD, fluid and blood leak within the neural retina through choroidal neovascularization, leading to fibrous scarring. Geographic atrophy, on the other hand, is characterized by progressive atrophy of the retinal pigment epithelium, choriocapillaris, and facing photoreceptors. These advanced forms of the disease are responsible for most of the visual impairment associated with AMD (Lim et al., 2012).

In routine ophthalmologic practice, colour fundus photography (CFP) is the most accessible and best-validated imaging tool for early AMD classification and prediction of progression towards the advanced stages (Ferris et al., 2005). However, the extraction of clinical signs of AMD from CFP is a time-consuming process. Its automation could bring several benefits, such as supporting precise severity scoring for further adapted AMD management, all while preserving interpretability.

Image recognition and classification are complex tasks for computers. It requires not only analysing groups of pixels and searching for specific patterns but also overcoming variations in images (scale, viewpoint, illumination, or partial occlusion with objects). Recently, a new subset of machine learning and artificial intelligence called 'deep learning' has revolutionized state-of-the-art automated medical image recognition. The term 'deep' implies that the model uses multiple layers of artificial neurons, allowing models to learn representations of data with various levels of abstraction.

DL has demonstrated its efficacy in various ophthalmology domains, including the detection of conditions like diabetic retinopathy (Gargeya & Leng, 2017; Gulshan et al., 2016; Jiang et al., 2020; Ting et al., 2017), retinopathy of prematurity (Brown et al., 2018), glaucoma-like disk (Li et al., 2018), and AMD (Burlina et al., 2017; Grassmann et al., 2018; Peng et al., 2019).

However, existing DL models in the field of AMD have primarily focused on diagnosing the disease (Burlina et al., 2017; Li et al., 2019; Luo et al., 2021) or predicting its progression to advanced stages. (Grassmann et al., 2018; Peng et al., 2019, 2020). Unfortunately, these models lack interpretability because of the elusive nature of their inner processes, often referred to as the 'black box effect'. This term describes a scenario in which the underlying mechanisms of a system are unclear or challenging to comprehend, leading to difficulties in understanding how it generates its results. In the medical field, it can raise ethical concerns when treatment choices for a patient derive from a DL model without insight into its operational logic. Interestingly, no prior models have specifically targeted the automated extraction of clinical features of AMD from CFP using DL.

In light of these voids, our study aims to develop 'DeepAlienorNet', a DL model capable of automatically extracting AMD clinical signs from CFP. The model is designed with a user-friendly interface, further enhancing its practicality and usability in clinical settings.

2 | METHODS

2.1 | Study participants

Participants in the ALIENOR (Antioxydants, Lipides Essentiels, Nutrition et Maladies Oculaires) Study are a cohort of French individuals 77 years of age or older recruited from an ongoing population-based study (Three-City [3C] Study) on the vascular risk factors of dementia (3C Study Group, 2003). The ALIENOR Study consists of periodic eye examinations performed on all participants of the 3C Study cohort in Bordeaux, France, since 2006 (Delcourt et al., 2010).

In the present study, we only used the first visit of the ALIENOR (baseline, Alienor 0, from October 2, 2006 to May 23, 2008). We did so to avoid redundancy in images of the same patients throughout successive follow-ups that could lead to overfitting. To be included, eye fundus images needed to have all the following labels correctly assessed as 'present' or 'absent': large soft drusen (>125 μm), intermediate soft (63–125 μm), large area of soft drusen (total area >500 μm), presence of central soft drusen (large or intermediate), hyperpigmentation, hypopigmentation, and advanced AMD (defined as exudative or atrophic AMD).

To represent day-to-day practice, we included all gradable photographs and did not exclude low-quality images (with poor contrast, sharpness, or bad reflection).

Data management and statistical analyses were performed between June 6, 2021 and September 19, 2022.

This research followed the tenets of the Declaration of Helsinki. Participants provided written informed consent. The design of the ALIENOR Study was approved by the Ethical Committee of Bordeaux, France, in May 2006 (Approval ID 2006/10).

2.2 | Eye examination

All eye examinations were performed in the Department of Ophthalmology of Bordeaux University Hospital by experienced technicians. Assessments included the recording of ophthalmic history, slit lamp examination, and 45° retinal photographs using a nonmydriatic fundus camera (TRC-NW6S; Topcon) (Chan et al., 2016; Saunier et al., 2018).

Retinal photographs were interpreted in duplicate by two specially trained technicians. Inconsistencies between the two interpretations were adjudicated by a senior grader (J.-F.K., M.-B.R., C. D, or M.-N.D.). All cases of advanced AMD and other retinal diseases were reviewed and confirmed by retina specialists (J.-F.K.,

M.-B.R., or M.-N.D.). This interpretation of retinal photographs was used as the ground truth in the development and evaluation of our prediction model.

2.3 | Data splitting

Our study used 5-fold cross-validation to split our initial dataset into five subsets. For each model, training was done with a different set of 4 folds, and validation was done on the remaining fold ('Cross-validation', n.d.). The global model performance was estimated by microaveraging the five models (Figure 1). Microaveraging means that we summed individual classes' true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) across the five models. This allowed us to estimate metrics and draw the receiver operating curve (ROC) over the entirety of our dataset.

2.4 | Development of the DL model

DeepAlienorNet was developed with Python 3.8 using Tensorflow and Keras ('Keras: the Python deep learning API', n.d.).

A precise description of the methodology we used to develop our model is explained in Appendix S1. We performed statistical analyses using Scikit-learn.

All experiments were conducted on a server with 2 Intel® Xeon® Gold SKL-6130 @ 2.1 GHz, using a NVidia® P100 GPU for training and testing, with 192 Gb available in RAM memory.

We also followed the minimum information about clinical artificial intelligence modelling (MI-CLAIM) checklist to provide transparent reporting of our DL model (Figure S1) (Norgeot et al., 2020).

2.5 | Threshold calculation

Our algorithm provides predictions for the presence of each clinical sign by assigning a probability value. However, to interpret these probabilities, they need to be converted into discrete class labels using a threshold. The commonly used default threshold of 50% may not perform well in cases where there is a significant class imbalance. To address this issue, we employed a simple approach of fine-tuning the threshold (Lipton et al., 2014). Given the highly imbalanced nature of our dataset, with more normal eye fundus images than pathological ones, we determined the optimal threshold for each clinical sign using the ROC analysis (Figure 2a). By calculating the Youden index (sensitivity + specificity − 1) for different threshold values between 0 and 1, we identified the threshold that yielded the highest Youden index value (Figure 2b). This approach allows us to optimize the performance of our algorithm by classifying each clinical sign accurately.

2.6 | Grad-CAM

Gradient-weighted class activation mapping (Grad-CAM) finds which parts of the image led a convolutional neural network to its final decision. This method produces thermal maps representing the activation classes on the images received as input. It is a crucial step to implement in DL models to reduce the black box effect and to understand which part of the images drives the predictions (Buhrmester et al., 2021; Selvaraju et al., 2020). We used Keras official documentation to get the code and implement Grad-CAM in our model (Chollet, n.d.).

Figure S3 shows various examples of original images associated with corresponding Grad-CAM images.

2.7 | User interface

We designed DeepAlienorNet to be easily usable for any ophthalmologist with minimum skills in computer sciences. As a result, we developed a user-friendly interface using Pygubu, an application tool used for the development of user interfaces for Python's Tkinter module.

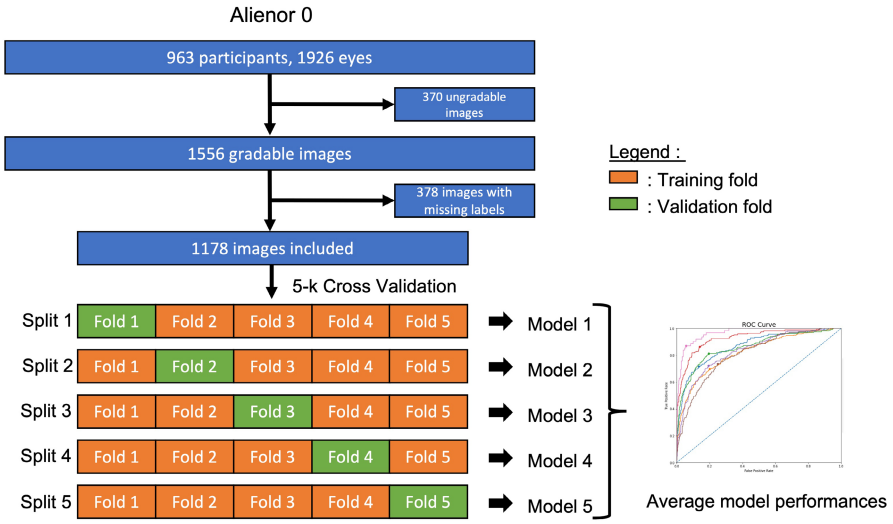

FIGURE 1 Flow chart and cross-validation. 1566 images were available at baseline, and 1178 images were included. We then do 5-k cross-validation: the dataset is split into five smaller sets named 'fold'. Then, five models are trained independently.

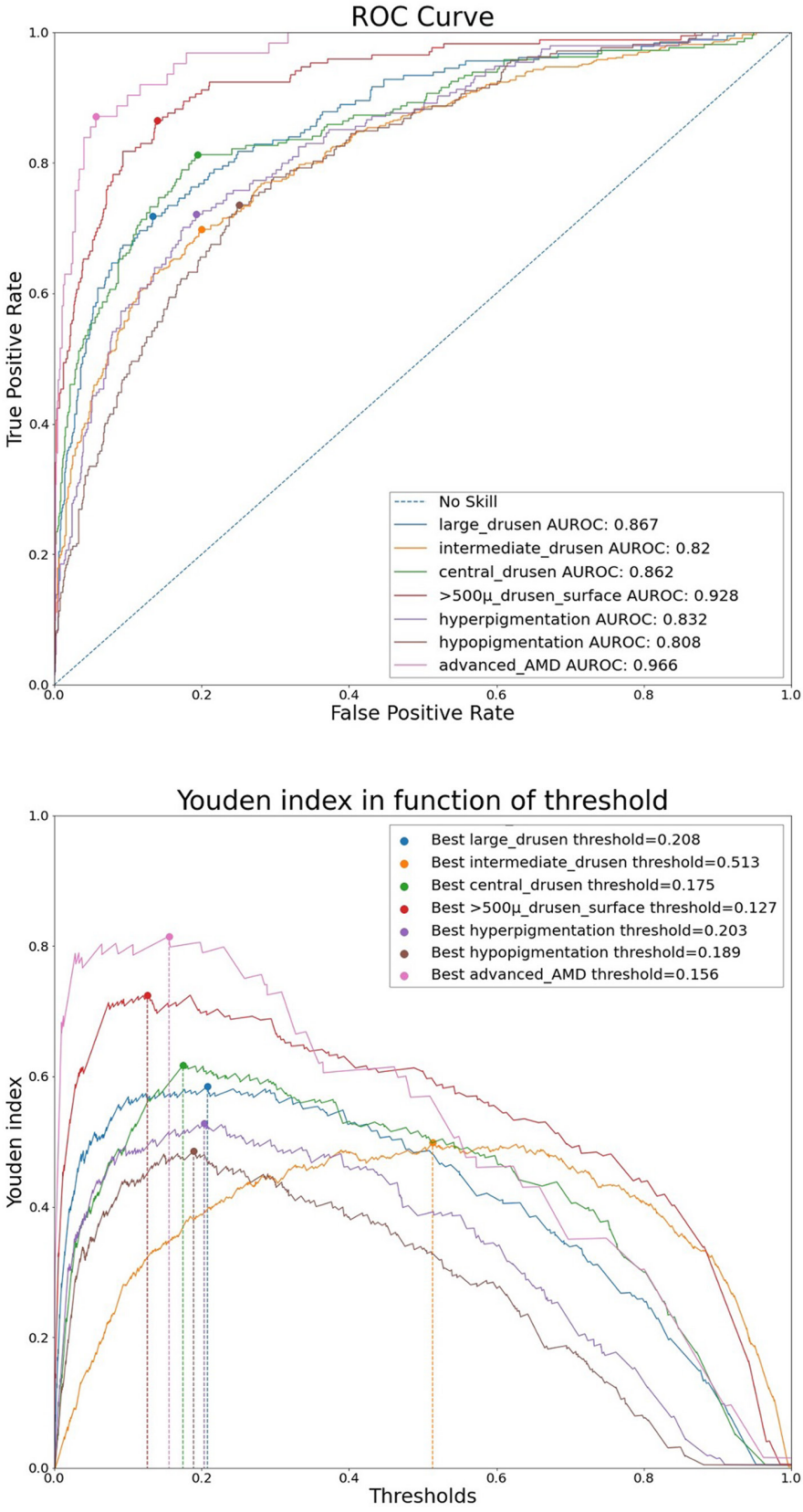

FIGURE 2 ROC curve and Youden index as a function of threshold. AMD, age-related macular disease; AUC, area under the curve; ROC, receiver operating curve.

2.8 | Metrics

To evaluate our model, we used various metrics to be as comparable as possible with previous and future models. We chose well-known metrics in medical fields such as sensitivity, specificity, and AUROC. We also calculated Cohen's kappa coefficient, which has been reported in previous DL publications.

These metrics are computed for each clinical sign, and 'overall metrics' are calculated by macroaveraging. The macro-averaged score is calculated by taking the arithmetic mean of all the per-clinical-sign scores

(Leung, 2022). We made this choice because, from our point of view, even if the dataset is imbalanced, each clinical sign should be equally considered when estimating the overall performance. The 95% confidence intervals were computed using stratified bootstrapping (Carpenter & Bithell, 2000).

3 | RESULTS

3.1 | Sample selection

Of the 963 participants in the 3C Study cohort in Bordeaux who participated in the baseline eye examination, 1926 eye fundus images were available and 1556 were gradable. A total of 1178 images met the inclusion criteria, with all labels graded. Table 1 shows the characteristics of the included eye fundi.

3.2 | Threshold finding

Based on the Youden index, the following thresholds were estimated (Figure 2): 20.8% for large drusen (>125 μm), 51.3% for intermediate drusen (65–125 μm), 12.7% for large areas of soft drusen, 17.5% for central soft drusen, 20.3% for hyperpigmentation, 18.9% for hypopigmentation and 15.6% for advanced AMD.

TABLE 1 Characteristics of images in the ALIENOR dataset.

Clinical sign	No. of participants	% Total
Normal images	515	44
Large drusen	184	16
Intermediate drusen	552	47
Large area of drusen[a]	170	14
Central drusen	214	18
Hypopigmentation	212	18
Hyperpigmentation	194	16
Advanced AMD	62	5
Total included images	1178	100

Abbreviation: AMD, age-related macular degeneration.
[a]Total area >500 μm.

3.3 | Predicting AMD clinical signs

Using the 5-fold cross-validation, we estimated the global performances of DeepAlienorNet on our entire dataset (n=1178) (Table 2).

The overall sensitivity, specificity and AUROC of our model across the seven clinical signs were 0.77 (95% confidence interval [CI]: 0.72–0.82), 0.83 (95% CI: 0.81–0.85) and 0.87 (95% CI: 0.85–0.89), respectively. Figure 2a shows the ROC for each clinical sign. The average Cohen's kappa was 0.48 (95% CI: 0.44–0.53).

The prediction performances were particularly high for advanced AMD (AUROC=0.97, sensitivity=0.86, specificity=0.94, Cohen's κ=0.57) and large areas of soft drusen (AUROC=0.93, sensitivity=0.86, specificity=0.86, Cohen's κ=0.56) and somewhat lower for other clinical signs (AUROC ranging from 0.81 to 0.83).

The confusion matrix for each clinical sign with corresponding TPs, TNs, FPs, and FNs is shown in Figure S2.

3.4 | Sensitivity analysis

Figure S4 shows some examples of well-classified and misclassified images. Figure S1a,b exposes cases where a central shadow could mislead the model in the diagnosis. Interestingly, in Figure S4a, the model still finds the relevant zone of interest for the classification, whereas in Figure S4b, this recognition could not be achieved. Figure S4c is an example of a poor-quality picture that can be seen in general practice. Even if the model fails its prediction, it, however, here recognizes a relevant zone of interest in the image. Figure S4d,e shows well-classified images. A high degree of confidence in predictions is associated with great contrast, colours and the absence of reflection or shadow in the background. It has to be noted that, in Figure S4d, the heatmap does not show any zone of interest as the image is strictly normal.

Overall, we see that images on which the model makes more than 2 or 3 mistakes are more likely to be those with poor contrast and sharpness or reflection. Conversely, images with a higher degree of confidence are more likely to be of great quality.

TABLE 2 Performances of clinical signs prediction in the ALIENOR dataset.

Clinical sign	Sensitivity (95% CI)	Specificity (95% CI)	AUROC (95% CI)	Cohen's kappa (95% CI)
Large drusen	0.72 (0.66–0.77)	0.87 (0.85–0.89)	0.87 (0.85–0.89)	0.49 (0.44–0.54)
Intermediate drusen	0.69 (0.66–0.73)	0.80 (0.77–0.83)	0.82 (0.80–0.84)	0.50 (0.46–0.55)
Central soft drusen	0.81 (0.77–0.86)	0.81 (0.78–0.82)	0.86 (0.83–0.89)	0.48 (0.44–0.52)
Large area of soft drusen	0.86 (0.81–0.91)	0.86 (0.85–0.88)	0.93 (0.91–0.95)	0.56 (0.52–0.60)
Hyperpigmentation	0.72 (0.66–0.78)	0.81 (0.78–0.83)	0.83 (0.80–0.86)	0.42 (0.36–0.47)
Hypopigmentation	0.73 (0.69–0.78)	0.75 (0.72–0.77)	0.81 (0.78–0.83)	0.36 (0.31–0.41)
Advanced AMD	0.86 (0.77–0.94)	0.94 (0.93–0.96)	0.97 (0.95–0.98)	0.57 (0.51–0.62)
Overall	0.77 (0.72–0.82)	0.83 (0.81–0.85)	0.87 (0.85–0.89)	0.48 (0.44–0.53)

Abbreviations: AMD, age-related macular degeneration; AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating curve.

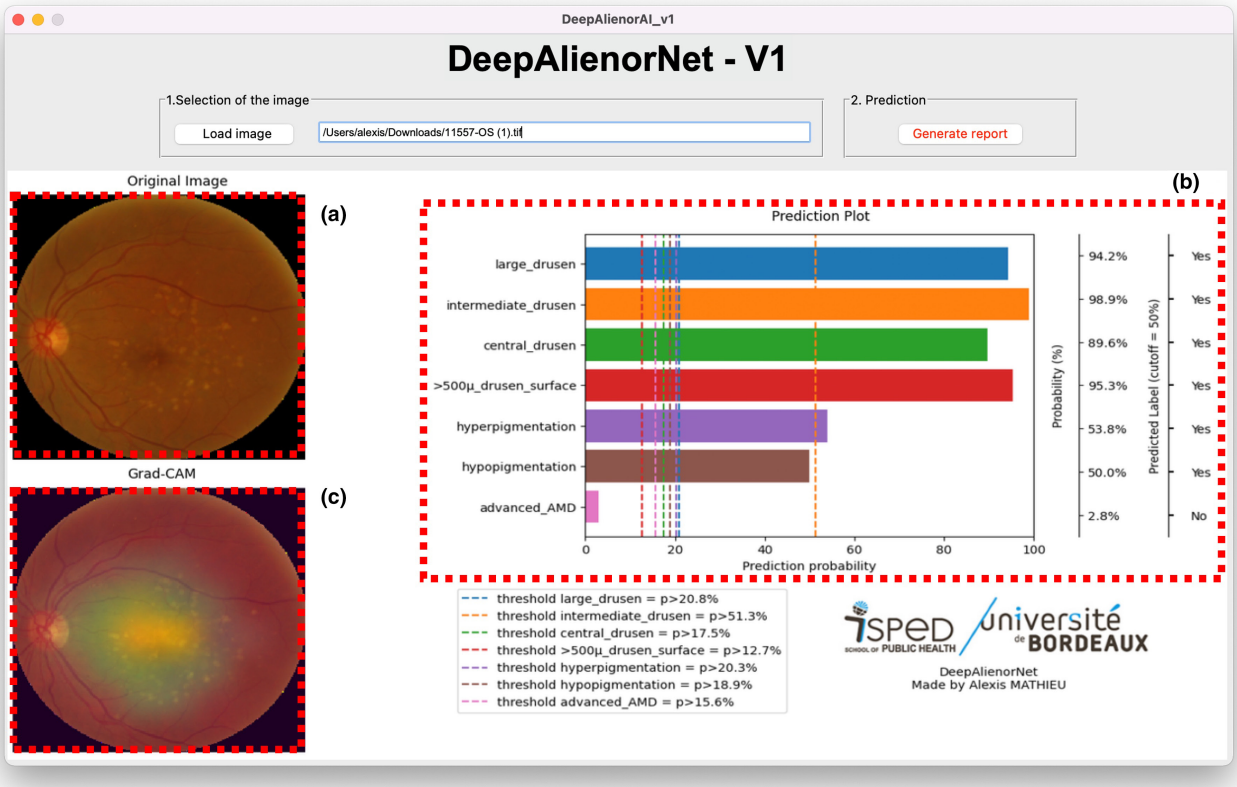

FIGURE 3 DeepAlienorNet Interface. The original image is displayed in full size to better interpret the clinical signs (a). The prediction results are shown as horizontal graph bars to represent the probabilities of clinical sign presence. Vertical interrupted lines represent the threshold for each clinical sign (b). A heatmap is displayed in the lower part of the window to provide better interpretability (c).

3.5 | User interface

Figure 3 exposes the user interface of DeepAlienorNet. A full-sized original image is present to help ophthalmologists control the results (Figure 3a). On the right part of the window, a horizontal bar graph shows the presence probability for each clinical sign (Figure 3b). Thresholds for each clinical sign are shown with vertical interrupted lines. The algorithm concludes that a sign is present if its probability is higher than its associated threshold. In the lower-left part of the window, we can find Grad-CAM (Figure 3c). This heatmap helps to understand which part of the image the algorithm used to predict labels.

4 | DISCUSSION

Among the 1178 included images, DeepAlienorNet achieved an overall sensitivity, specificity, and AUROC of 0.77, 0.83, and 0.87, respectively. The model demonstrated particularly strong performance in predicting advanced AMD and large areas of soft drusen.

Existing DL models in the field of AMD have predominantly focused on diagnosing the disease (Burlina et al., 2017; Li et al., 2019; Luo et al., 2021) or predicting its progression to advanced stages (Burlina et al., 2017; Grassmann et al., 2018; Peng et al., 2019). However, a significant limitation of these models is their lack of interpretability, often attributed to the black box effect. Notably, there is a gap in the literature concerning DL models that specifically target the automated extraction of clinical features from CFP for AMD.

Prioritizing the extraction of clinical features, rather than solely focusing on diagnoses or risk predictions, holds the potential to enhance interpretability and address ethical concerns associated with DL models. Furthermore, integrating such DL models with well-established and validated AMD progression scores, such as the AREDS Severity Scale (Ferris et al., 2005) or Macutest (Ajana et al., 2021), can enable seamless automation of the assessment process. For instance, we can imagine a system where, instead of just requesting a DL model to predict the AREDS score directly, we input images of both eyes of a patient. DeepAlienorNet then extracts the clinical signs present in each eye and merges the outcomes to calculate the AREDS score. This approach harmonizes the robustness of the AREDS score with the interpretability of DeepAlienorNet, all within an automated pipeline designed for ophthalmologists.

In addition to extracting clinical signs, DeepAlienorNet goes a step further by generating heatmaps that indicate the specific regions of the image used to predict these signs. This innovative approach offers two significant advantages. First, it helps mitigate the black box effect, providing clinicians with interpretability and insight into the model's decision-making process. This enables a better understanding of how the model arrives at its predictions (Buhrmester et al., 2021; Selvaraju et al., 2020). Second, studies in

diabetic retinopathy have shown that the automatic extraction of clinical signs from CFP, combined with clinical feature heatmaps, can enhance the confidence and accuracy of diagnosis in a supervised reading environment (Sayres et al., 2019). By incorporating heatmaps, DeepAlienorNet empowers clinicians by providing visual cues that highlight the regions of interest and contribute to more reliable and precise assessments.

Another noteworthy advantage of DeepAlienorNet is its user-friendly interface. Unlike many deep learning models that require expertise in Python coding to be applied in clinical practice, we specifically designed DeepAlienorNet with this limitation in mind. Our objective was to create an interface that simplifies the utilization of the model, particularly in day-to-day clinical practice. By addressing the inherent complexity of deep learning models, we hope to make DeepAlienorNet more accessible to healthcare professionals who may not possess extensive programming knowledge.

Despite the promising results obtained, our model has several limitations that need to be considered. First, it currently relies solely on CFP, while Optical Coherence Tomography (OCT) has become a crucial tool in the diagnosis and follow-up of AMD. This limitation arises from the lack of extensive databases containing well-labelled CFP and OCT images required to train a model that incorporates both inputs. However, it is worth noting that most of the currently validated progression scores, such as the AREDS score, have been developed solely based on CFP. These scores depend on identifying anomalies like hyper- or hypopigmentation, which cannot be distinguished using OCT. Nevertheless, future enhancements to our model will prioritize addressing this limitation, aiming to closely emulate the thought process of ophthalmologists in real-life scenarios by leveraging both CFP and OCT.

Another limitation to consider regarding our model is the relatively small dataset size. Although we were able to develop a DL model using 1178 images, this number pales in comparison to non-medical DL architectures like ImageNet, which comprises 14 million labelled images. This limitation is not exclusive to ophthalmology but is applicable to other specialized medical fields as well. To advance DL in AMD and ophthalmology in general, it is crucial to establish larger and more comprehensive datasets (Alzubaidi et al., 2020, 2021).

Last, our model's performance may be constrained by the use of a single device (TRC-NW6S; Topcon, Japan) for capturing the 45° CFP images used during the training process. Consequently, the generalizability of our model to other CFP images obtained from different devices, particularly those with wider fields of view, may be limited, potentially leading to decreased performance.

However, one of the strengths of our study lies in the population-based nature of our sample and the inclusion of all gradable images, regardless of their quality. This approach minimizes selection bias, as databases created in clinical settings may overrepresent high-quality images and typical cases, and thus overestimate the classification performances in real life. Additionally, we used a rigorous grading scheme to ensure that the model was developed and tested based on valid CFP classifications, further bolstering the robustness of our study.

A last important consideration is that DeepAlienorNet's DL nature inherently raises ethical concerns. First, while its purpose is to aid clinicians, ethical dilemmas may arise if its suggestions override or unduly influence ophthalmologists' judgement. There should be a balance between the use of AI as a supportive tool and preserving the autonomy and expertise of healthcare providers. Second, employing AI models like DeepAlienorNet in healthcare requires informed consent from patients regarding the use of their medical data. Ensuring patient data privacy and security against potential breaches or misuse is crucial. Finally, DL models need continual monitoring and updates to adapt to new data, technologies, or emerging biases. Neglecting regular model evaluations and updates might result in outdated or inaccurate predictions, impacting patient care.

5 | CONCLUSION

DeepAlienorNet demonstrates promising performance in automatically identifying clinical signs of AMD from CFP, offering several notable advantages. High-interpretability of our model mitigates the black box effect and alleviates ethical concerns, as the use of black box models in AMD diagnosis can result in decreased transparency, potentially leading to inaccuracies in patient care. Additionally, the model can be easily integrated to automate well-established and validated AMD progression scores, and the user-friendly interface further enhances its usability. The main value of DeepAlienorNet lies in its ability to assist in precise severity scoring for further adapted AMD management, all while preserving interpretability.

ACKNOWLEDGEMENTS

The ALIENOR study was supported by Théa Pharma, Fondation Voir et Entendre, University of Bordeaux, Agence Nationale de la Recherche (ANR 2010-PRSP-011 VISA), CFSR Recherche (Club Francophone des Spécialistes de la Rétine), CNSA (Caisse Nationale pour la Solidarité et l'Autonomie) and the French Ministry of Health (PHRC, 2012, PHRC12_157 ECLAIR). Théa Pharma participated in the design of the ALIENOR study, but none of the sponsors participated in the collection, management, statistical analyses, interpretation of the data, or preparation, review or approval of the present manuscript. The authors are grateful for the support provided by the French clinical research network, FRCRnet/FCRIN.

CONFLICT OF INTEREST STATEMENT

A. Mathieu: AbbVie. J-F. Korobelnik: AbbVie, Bayer, KangHong, Novartis, Roche, Théa. M-B. Rougier: AbbVie, Bausch&Lomb, Bayer, Novartis, Théa. C. Delcourt: AbbVie, Bausch&Lomb, Novartis, Roche, Théa. M-N. Delyfer: AbbVie, Bayer, Horama, Horus Pharma, Novartis, Roche, Théa.

ORCID

Alexis Mathieu https://orcid.org/0000-0002-0638-4354
Jean-François Korobelnik https://orcid.org/0000-0002-4438-9535
Cécile Delcourt https://orcid.org/0000-0002-2099-0481
Marie-Noëlle Delyfer https://orcid.org/0000-0002-3127-8617

REFERENCES

3C Study Group. (2003) Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, 22, 316–325.

Ajana, S., Cougnard-Grégoire, A., Colijn, J.M., Merle, B.M.J., Verzijden, T., de Jong, P.T.V.M. et al. (2021) Predicting progression to advanced age-related macular degeneration from clinical, genetic, and lifestyle factors using machine learning. *Ophthalmology*, 128, 587–597.

Alzubaidi L, Al-Amidie M, Al-Asadi A, Humaidi AJ, Al-Shamma O, Fadhel MA, et al. (2021) Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13, 1590.

Alzubaidi L, Fadhel MA, Al-Shamma O, Zhang J, Santamaría J, Duan Y, et al. (2020) Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, 10, 4523.

AMD Alliance International. (2010) *The global economic cost of visual impairment*. Australia: Access Economics Pty Limited.

Brown, J.M., Campbell, J.P., Beers, A., Chang, K., Ostmo, S., Chan, R.V.P. et al. (2018) Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmology*, 136, 803–810.

Buhrmester, V., Münch, D. & Arens, M. (2021) Analysis of explainers of black box deep neural networks for computer vision: a survey. *Machine Learning and Knowledge Extraction*, 3, 966–989.

Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E. & Bressler, N.M. (2017) Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmology*, 135, 1170–1176.

Carpenter, J. & Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141–1164.

Chan, H., Cougnard-Grégoire, A., Delyfer, M.-N., Combillet, F., Rougier, M.B., Schweitzer, C. et al. (2016) Multimodal imaging of reticular pseudodrusen in a population-based setting: the Alienor study. *Investigative Ophthalmology & Visual Science*, 57, 3058–3065.

Chollet, F. (n.d.) Keras documentation: Grad-CAM class activation visualization. https://keras.io/examples/vision/grad_cam/

Cross-Validation: Evaluating Estimator Performance. (n.d.) Scikit-Learn. https://scikit-learn.org/stable/modules/cross_validation.html

Delcourt, C., Korobelnik, J.-F., Barberger-Gateau, P., Delyfer, M.N., Rougier, M.B., Le Goff, M. et al. (2010) Nutrition and age-related eye diseases: the Alienor (Antioxydants, Lipides Essentiels, Nutrition et maladies OculaiRes) Study. *The Journal of Nutrition, Health & Aging*, 14, 854–861.

Ferris, F.L., Davis, M.D., Clemons, T.E., Lee, L.Y., Chew, E.Y., Lindblad, A.S. et al. (2005) A simplified severity scale for age-related macular degeneration: AREDS Report No. 18. *Archives of Ophthalmology (Chicago, Ill: 1960)*, 123, 1570–1574.

Fleckenstein, M., Keenan, T.D.L., Guymer, R.H., Chakravarthy, U., Schmitz-Valckenberg, S., Klaver, C.C. et al. (2021) Age-related macular degeneration. *Nature Reviews Disease Primers*, 7, 31.

Gargeya, R. & Leng, T. (2017) Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124, 962–969.

Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M.E., Linkohr, B. et al. (2018) A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, 125, 1410–1420.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A. et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316, 2402–2410.

Jiang, H., Xu, J., Shi, R., Yang, K., Zhang, D., Gao, M. et al. (2020) A multi-label deep learning model with interpretable grad-CAM for diabetic retinopathy classification. In: *Presented at the 2020 42nd annual international conference of the IEEE engineering in medicine biology society (EMBC)*, pp. 1560–1563.

Keras: the Python deep learning API (n.d.) https://keras.io/.

Leung, K. (2022) Micro, macro & weighted averages of F1 score, clearly explained. Medium.

Li, F., Chen, H., Liu, Z., Zhang, X. & Wu, Z. (2019) Fully automated detection of retinal disorders by image-based deep learning. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 257, 495–505.

Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T. & He, M. (2018) Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 125, 1199–1206.

Lim, L.S., Mitchell, P., Seddon, J.M., Holz, F.G. & Wong, T.Y. (2012) Age-related macular degeneration. *The Lancet*, 379, 1728–1738.

Lipton, Z.C., Elkan, C. & Narayanaswamy, B. (2014) Thresholding classifiers to maximize F1 score. *ArXiv14021892 Cs Stat*.

Luo, X., Li, J., Chen, M., Yang, X. & Li, X. (2021) Ophthalmic disease detection via deep learning with a novel mixture loss function. *IEEE Journal of Biomedical and Health Informatics*, 25, 3332–3339.

Norgeot, B., Quer, G., Beaulieu-Jones, B.K., Torkamani, A., Dias, R., Gianfrancesco, M. et al. (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26, 1320–1324.

Peng, Y., Dharssi, S., Chen, Q., Keenan, T.D., Agrón, E., Wong, W.T. et al. (2019) DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, 126, 565–575.

Peng, Y., Keenan, T.D., Chen, Q., Agrón, E., Allot, A., Wong, W.T. et al. (2020) Predicting risk of late age-related macular degeneration using deep learning. *npj Digital Medicine*, 3, 111.

Saunier, V., Merle, B.M.J., Delyfer, M.-N., Cougnard-Grégoire, A., Rougier, M.B., Amouyel, P. et al. (2018) Incidence of and risk factors associated with age-related macular degeneration: four-year follow-up from the ALIENOR study. *JAMA Ophthalmology*, 136, 473–481.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N. et al. (2019) Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 126, 552–564.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128, 336–359.

Ting, D.S.W., Cheung, C.Y.-L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A. et al. (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318, 2211–2223.

Wong, W.L., Su, X., Li, X., Cheung, C.M.G., Klein, R., Cheng, C.-Y. et al. (2014) Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health*, 2, e106–e116.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mathieu, A., Ajana, S., Korobelnik, J.-F., Le Goff, M., Gontier, B., Rougier, M.-B. et al. (2024) DeepAlienorNet: A deep learning model to extract clinical features from colour fundus photography in age-related macular degeneration. *Acta Ophthalmologica*, 00, 1–8. Available from: https://doi.org/10.1111/aos.16660