Multimodal deep transfer learning for the analysis of optical coherence tomography scans and retinal fundus photographs

Zoi Tsangalidou
Department of Statistics,
University of Oxford

Edwin Fong
Novo Nordisk A/S

Josefine Vilsbøll Sundgaard
Novo Nordisk A/S

Trine Julie Abrahamsen
Novo Nordisk A/S

Kajsa Kvist
Novo Nordisk A/S

Abstract

Deep learning methods are increasingly applied to ophthalmologic scans in order to diagnose and prognosticate eye diseases, cardiovascular or renal outcomes. In this work, we create a multimodal deep learning model that combines retinal fundus photographs and optical coherence tomography scans and evaluate it in predictive tasks, matching state-of-the-art performance with a smaller dataset. We use saliency maps to showcase which sections of the eye morphology influence the model's prediction and benchmark the performance of the multimodal model against algorithms that utilize only the individual modalities.

1 Introduction

In recent years, machine learning has found multiple uses in ophthalmology, including image segmentation [1], disease diagnosis [2, 3] and prognosis [1, 4]. With ever increasing dataset sizes [5], imaging has a key role in the field thanks to the availability of diverse modalities, such as retinal fundus photographs, optical coherence tomography (OCT), anterior segment photographs, and corneal topography. Retinal OCT and fundus photographs are common non-invasive imaging modalities applied in an ophthalmology setting, and they are both part of routine eye examinations. Applications of computer vision algorithms in ophthalmology have primarily focused on retinal fundus photographs [3, 6, 7], although with recent technological advances in OCT devices, this modality is also becoming the object of increased recent attention [1, 8].

OCT scans and fundus photographs are considered complementary diagnostic modalities, as they capture different aspects of the eye's morphology. Fundus images provide a clear view of the posterior pole with the optic nerve head, retinal vasculature and macula, while OCTs are volumetric scans capturing cross-sections of the retinal layers [9]. For retinal disease diagnosis, an ophthalmologist would consider pathological findings in both imaging modalities alongside the patient's medical history. Currently, most automated diagnosis and prognosis algorithms are based on a single imaging modality, with little work done thus far on a joint approach [10, 11]. This is primarily due to limited availability of datasets with paired OCT and fundus imaging.

To address if and how beneficial joint OCT and fundus modelling is for diagnostic or prognostic purposes, we endeavour to assess the performance gains obtained through a multimodal approach compared to the analysis of individual modalities. Even in large-scale population health studies, such as the UKBiobank (UKBB), disease outcomes ranging from various eye conditions (e.g. age-related macular degeneration) to cardiovascular endpoints (e.g. Major Adverse Cardiovascular Event

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(MACE)) are sufficiently rare that developing predictive algorithms poses a difficult challenge. This is further complicated by the fact that predicting occurrence of such outcomes would also be difficult even for an experienced physician [1]. Therefore, in this early work we focus on predicting simpler variables that have previously been shown to be possible to predict, and we evaluate the benefits of a multimodal algorithm. Finally, we assess model interpretability using saliency maps, comparing the multimodal and individual modalities approach.

2 Methods

2.1 Data

We used imaging data, baseline characteristics, and phenotypes from the UKBB (`http://www.ukbiobank.ac.uk`). We considered two outcome variables: sex and systolic blood pressure recorded during the eye imaging visits. The imaging data consist of paired retinal fundus photography and OCT scans. The UKBB performed eye imaging during the initial assessment visit (2006-2010) and the first repeat assessment visit (2012-13). For individuals scanned at both visits, we utilized the initial visit unless it was incomplete (i.e. the individual did not have paired scans for the left or the right eye), in which case we used the repeat visit. The data was randomly split into training, validation and testing sets (with proportions $80\%, 10\%, 10\%$). Each eye was considered a separate data point and we ensured that both eyes of an individual were in the same training/validation/test set.

We preprocessed fundus photographs by automatically detecting a circular mask using a Hough transform [12], resizing and applying a center crop to obtain a 587×587 RGB image. Images for which this mask could not be detected were discarded. OCT scans are volumetric so we first randomly selected three contiguous slices in the middle of the volume and treated them as an RGB image with channels resized to dimensions 512×512. Pixel values were then normalized to ImageNet data statistics in order to apply transfer learning. Multiple transforms were used for data augmentation; random vertical flips, rotations, and erasing [13] for fundus photographs, and random horizontal flips and pixel value inversions for OCT scans. We further explored using Contrast Limited Adaptive Histogram Equalization (CLAHE)[14] as a preprocessing step to emphasize vasculature.

2.2 Model and architecture

We assessed the single and joint modality architectures in both sex and hypertension prediction. For comparability of results, we framed both as binary classification tasks and defined hypertension as a systolic blood pressure measurement exceeding 140 mmHg [15]. We trained separate models for each task using transfer learning and employed the following architectures:

- **Single modality (OCT or fundus only).** InceptionV3 [16] pretrained on ImageNet with the last classification layer replaced by a linear fully connected layer with one output neuron.

- **Multimodal/joint modalities (OCT and fundus).** Two separate InceptionV3 models pretrained on ImageNet, where the last classification layer is replaced by a linear fully connected layer with 50 output neurons. The two 50-dimensional vector outputs are then concatenated and propagated through a final linear layer with a single output node.

Our experiments showed that transfer learning substantially increases convergence speed and that adjusting all weights, rather than freezing some as is customary when applying pretrained models in computer vision tasks, results in substantial performance gains consistently with previous work on OCT and fundus scans [3, 17, 18]. Therefore, we opted to train the entire network across all models and experiments. We experimented with other backbone architectures such as VGG-16 [19] and ResNet-50 [20], but obtained superior performance with InceptionV3. Unless otherwise stated, we used an Adagrad optimizer [21] with a learning rate of 0.001, binary cross entropy loss and batch sizes of 16 and 8 for single modality and joint modality experiments respectively. Training a single model required approximately 140 hours on an Amazon Web Services NVIDIA® V100 Tensor Core GPU, and we implemented early stopping if the validation loss had not decreased for 10 epochs.

2.3 Interpretability

We produced saliency maps using image-specific class saliency visualization [22] in order to visualize the information that the single and joint modality models leveraged for their predictions. For each pixel, the maximum gradient value across all three colour channels is extracted and the result is plotted as a heatmap overlayed on the original image.

3 Results

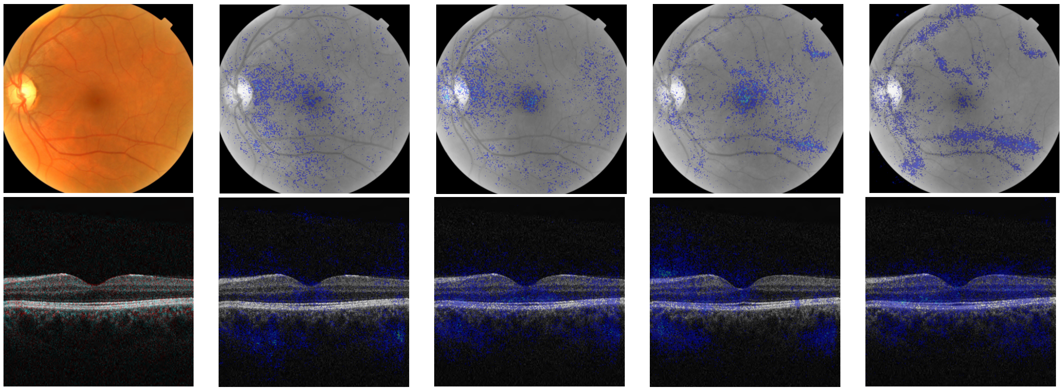
Our dataset consisted of 130 558 scans from 66 041 participants, of which 29 954 were males and 36 087 females. The average age and systolic blood pressure were 57.4 years (s.d.=8.3) and 140.3 mmHg (s.d.=19.7), respectively, and did not differ across the training, validation and testing sets (Table A1). We present performance metrics for both prediction tasks in Table 1 and Figures A1 - A4.

When predicting sex, we observed increased performance with the multimodal approach compared to only using the individual modalities (AUC=0.97 vs AUC= 0.93), a comparable result to a fundus-only model published by Poplin et al. [17], which used a much larger dataset and excluded low-quality scans, while we used all images irrespective of quality.[1] The increased joint model performance could arise from the model extracting complementary information from the fundus and OCT imaging, or because one image modality can compensate for the other when it is not of good quality.

Table 1: Performance metrics on the held-out test set across models and tasks.

AUC: Area under the Receiver Operating Characteristic curve. PR-AUC: Area under the precision-recall curve.

Task & Model	Accuracy	AUC	PR-AUC	Specificity	Sensitivity	F1 score
Sex - Fundus	85.2%	0.932	0.847	88.7%	81.1%	83.1%
Sex - OCT	85.5%	0.934	0.898	88.4%	82.0%	83.6%
Sex - Joint	**91.0%**	**0.969**	**0.931**	**91.8%**	**90.0%**	**90.0%**
Hypertension - Fundus	67.9%	0.747	0.718	**69.5%**	66.2%	66.6%
Hypertension - OCT	61.2%	0.647	0.606	60.3%	62.2%	60.9%
Hypertension - Joint	**68.5%**	**0.754**	**0.724**	64.2%	**73.1%**	**69.3%**



Original

Sex single mod.

Sex multimodal

Hypertension single mod.

Hypertension multimodal

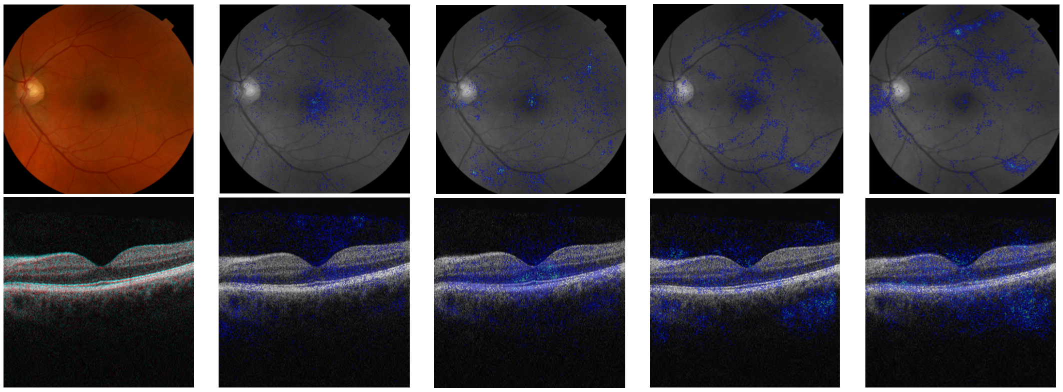
Figure 1: Saliency maps of scans from female without hypertension.

We thus obtained saliency maps (Figures 1, 2, A5 and A6) to assess whether our model was assigning increased importance to different parts of the image when using a single modality compared to using both, as this would provide evidence that the improved performance is not just due to image quality.

[1] For reference, previous published models predicting sex reported AUC=0.97 (fundus) [17], AUC=0.80 (fundus) [23] and AUC=0.908 (OCT) [24]

Indeed, in the fundus-only sex model, the foveal and optic nerve regions were given high importance, but in the joint model, signal was distributed more evenly throughout the retinal fundus photograph as the OCT image provides a much more detailed view of the fovea. For both the single and joint modality OCT saliency maps, we observed a high importance of the fovea and retinal layer thickness.

For hypertension, the fundus-only model is on par with the joint model (AUC= 0.747 vs AUC= 0.754) and outperforms using only OCT (AUC= 0.647). Results from using CLAHE for preprocessing to further emphasize vessels are omitted as the performance did not substantially improve. Our hypothesis is consistent with these results - we expect blood vessel morphology to be the most informative eye scan feature to predict hypertension and vasculature is primarily visible in fundus images (Figure 1 and 2). The joint model saliency maps appear to trace the vessel morphology more accurately than the fundus-only model, potentially because the model shifts its focus from the fovea region (which is better captured in the OCT) towards the vasculature, an observation consistent with the sex prediction saliency maps.



Original

Sex single mod.

Sex multimodal

Hypertension single mod.

Hypertension multimodal

Figure 2: Saliency map of scans from male with hypertension.

4 Discussion and next steps

We believe this is the first attempt to combine OCT scans and retinal fundus photography in a dataset as large as UKBB. The results show the importance of a multimodal approach, since these modalities contain complementary information and jointly guide ophthalmologists in clinical practice. Our sex prediction model achieves performance on par with state-of-the-art approaches [17, 23, 24, 25] while using a smaller training dataset ($130k$ vs $1.7mil$ images in Poplin et al. [17]) and without excluding low-quality scans. We expect excluding low-grade images to further improve accuracy. However, manually obtaining image quality scores is a resource intensive task and we believe it is valuable to develop sufficiently accurate models that are directly applicable to datasets arising from standard clinical practice.

We selected sex and hypertension as target variables, since we expect these to differ in terms of whether a multimodal approach is beneficial and therefore provide a good proof-of-concept for our study. On one hand, it has been previously shown that it is possible to predict sex from both retinal fundus photographs and OCT scans [17, 23, 24, 25] so a multimodal model may be able to capture and combine information for both. On the other hand, hypertension is expected to be visible through the retinal vasculature which is primarily captured in retinal fundus photographs, and we thus expect an OCT scan to not provide additional useful information for the prediction.

We show the added value of a joint model depends on the task, which illustrates that it is crucial to make modelling choices based on clinical domain knowledge. Specifically for cardiovascular or renal outcomes, we anticipate fundus-only models to be sufficient, as blood vessel morphology should be the most important feature. However, for eye disease, the multimodal approach may outperform individual modalities, as it incorporates the optic nerve, fovea, vasculature, as well as retinal layer thickness and structure into the prediction. As a next step, we plan to explore the prediction of eye diseases and establish where and to what extent multimodality is beneficial.

We further establish that the fovea, optic nerve and pigmentation are key for predicting sex, while vasculature is a key determinant of hypertension. It is therefore likely that saliency maps could provide novel biological insights when our multimodal approach is applied to eye, renal or cardiovascular diseases. It would be particularly useful to assess which parts of the two modalities are exploited and whether these are complementary in the joint model. For example, we discovered that the fundus-only model focused mainly on the optic nerve and the fovea when predicting sex, but when the OCT scan was also provided as input, the importance shifted towards pigmentation, as the OCT captures the fovea in more detail than the fundus image.

To conclude, we have taken the first steps in establishing a multimodal approach combining fundus and OCT scans, which we hope to further develop in order to better diagnose and prognose eye disease, while simultaneously obtaining valuable biological insight.

Acknowledgements

We would like to thank our colleagues Anna Menacher and Carina Slidsborg for their helpful comments on our draft, and Sile Hu and Joanna Howson for assisting with data access. We would like to thank the UK Biobank participants for their contribution to the study which was conducted under application 53639.

References

[1] Jason Yim, Reena Chopra, Terry Spitz, Jim Winkens, Annette Obika, Christopher Kelly, Harry Askham, Marko Lukic, Josef Huemer, Katrin Fasler, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine*, 26(6):892–899, 2020.

[2] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[3] Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S Corrado, Lily Peng, Dale R Webster, Naama Hammel, Yun Liu, and Avinash V Varadarajan. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1):18–27, 2020.

[4] Jooyoung Chang, Ahryoung Ko, Sang Min Park, Seulggie Choi, Kyuwoong Kim, Sung Min Kim, Jae Moon Yun, Uk Kang, Il Hyung Shin, Joo Young Shin, et al. Association of cardiovascular mortality and deep learning-funduscopic atherosclerosis score derived from retinal fundus images. *American Journal of Ophthalmology*, 217:121–130, 2020.

[5] Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1):e51–e66, 2021.

[6] Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *Plos one*, 14(9):e0222025, 2019.

[7] Friso G Heslinga, Josien PW Pluim, AJHM Houben, Miranda T Schram, Ronald MA Henry, Coen DA Stehouwer, Marleen J Van Greevenbroek, Tos TJM Berendschot, and Mitko Veta. Direct classification of type 2 diabetes from retinal fundus images in a population-based sample from the maastricht study. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 383–388. SPIE, 2020.

[8] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[9] Matthew D Davis, Susan B Bressler, Lloyd Paul Aiello, Neil M Bressler, David J Browning, Christina J Flaxel, Donald S Fong, William J Foster, Adam R Glassman, Mary Elizabeth R Hartnett, et al. Comparison of time-domain oct and fundus photographic assessments of retinal thickening in eyes with diabetic macular edema. *Investigative ophthalmology & visual science*, 49(5):1745–1752, 2008.

[10] Tae Keun Yoo, Joon Yul Choi, Jeong Gi Seo, Bhoopalan Ramasubramanian, Sundaramoorthy Selvaperumal, and Deok Won Kim. The possibility of the combination of oct and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. *Medical & biological engineering & computing*, 57(3):677–687, 2019.

[11] Xingxin He, Ying Deng, Leyuan Fang, and Qinghua Peng. Multi-modal retinal image classification with modality-specific attention network. *IEEE Transactions on Medical Imaging*, 40(6): 1591–1602, 2021.

[12] HK Yuen, John Princen, John Illingworth, and Josef Kittler. Comparative study of hough transform methods for circle finding. *Image and vision computing*, 8(1):71–77, 1990.

[13] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[14] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3): 355–368, 1987.

[15] American heart association. `https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings`. Accessed: 2022-09-23.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[17] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

[18] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[23] Marion R Munk, Thomas Kurmann, Pablo Marquez-Neila, Martin S Zinkernagel, Sebastian Wolf, and Raphael Sznitman. Assessment of patient specific information in the wild on fundus photography and optical coherence tomography. *Scientific reports*, 11(1):1–10, 2021.

[24] Kuan-Ming Chueh, Yi-Ting Hsieh, and Sheng-Lung Huang. Prediction of gender from macular optical coherence tomography using deep learning. *Investigative Ophthalmology & Visual Science*, 61(7):2042–2042, 2020.

[25] Edward Korot, Nikolas Pontikos, Xiaoxuan Liu, Siegfried K Wagner, Livia Faes, Josef Huemer, Konstantinos Balaskas, Alastair K Denniston, Anthony Khawaja, and Pearse A Keane. Predicting sex from retinal fundus photographs using automated deep learning. *Scientific reports*, 11 (1):1–8, 2021.

Checklist

1. For all authors...
    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
    (b) Did you describe the limitations of your work? [Yes] See discussion.
    (c) Did you discuss any potential negative societal impacts of your work? [N/A] No direct clinical and therefore societal impact from this work
    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
    (a) Did you state the full set of assumptions of all theoretical results? [N/A]
    (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The pretrained models are available through Pytorch and we have provided sufficient information to reproduce the results. The UKBB data is not is licensed under application number TODO.
    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The experiments require substantial runtime and due to limited resources, we were not able to run enough replicates to obtain valid confidence intervals. We did however repeat all experiments at least 3 times and obtained consistent results.
    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    (a) If your work uses existing assets, did you cite the creators? [Yes]
    (b) Did you mention the license of the assets? [Yes] See acknowledgements for mention of the specific UKBB application in the non-anonymised version.
    (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The UKBiobank study is a well-known resource and informed consent was obtained from all participants at the time of recruitment.
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The UKBiobank study is an anonymised research study and all participants provided informed consent at the time of recruitment.
5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

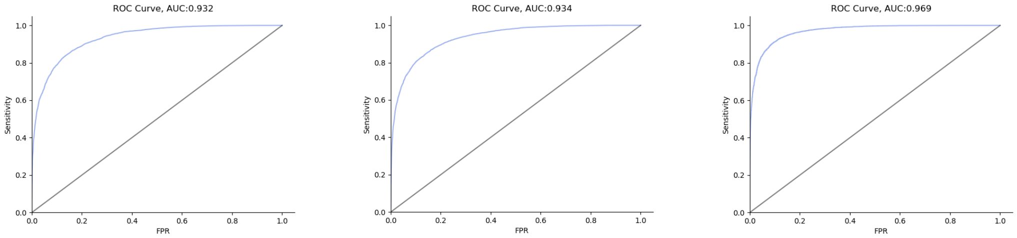
A.1 Tables

Table A1: Baseline characteristics of UKBB participants.

	Training set	Validation set	Testing set
Number of participants	66 041	8 255	8 255
Number of eyes	130 558	16 318	16 325
Males	29 954	3 815	3 708
Females	36 087	4 440	4 547
Age (s.d.)	57.4 (8.3)	57.4 (8.3)	57.7 (8.3)
Systolic blood pressure (s.d.)	140.3 (19.7)	140.1 (19.8)	140.2 (19.6)

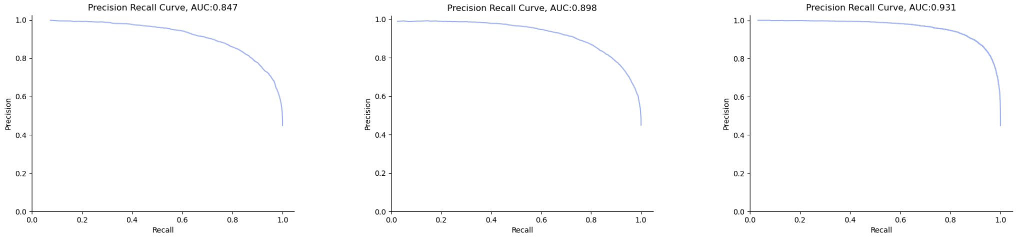
A.2 Figures

A.2.1 Model performance



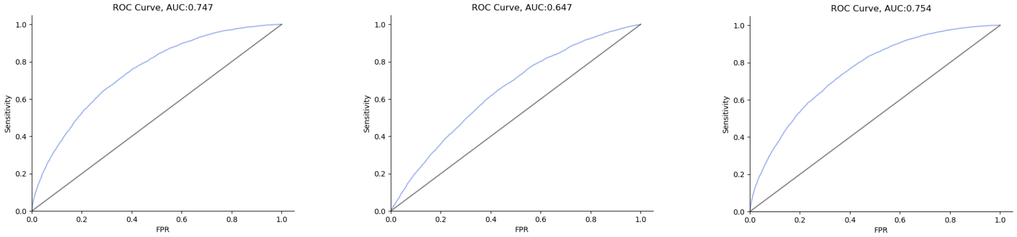
Fundus only OCT only Joint

Figure A1: Sex receiver operating characteristic curves.



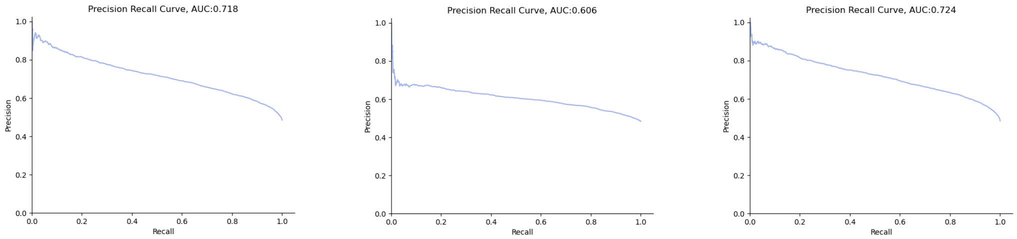
Fundus only OCT only Joint

Figure A2: Sex precision recall curves.



Fundus only OCT only Joint

Figure A3: Hypertension receiver operating characteristic curves.



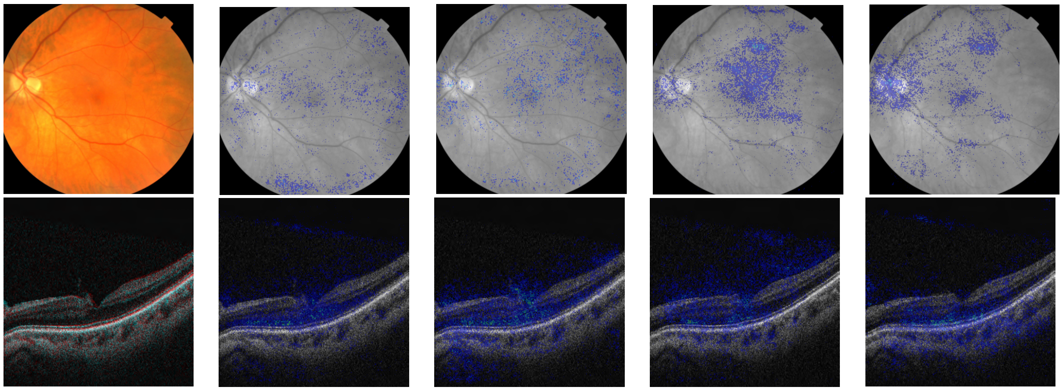
Fundus only

OCT only

Joint

Figure A4: Hypertension precision recall curves.

A.2.2 Saliency maps



Original

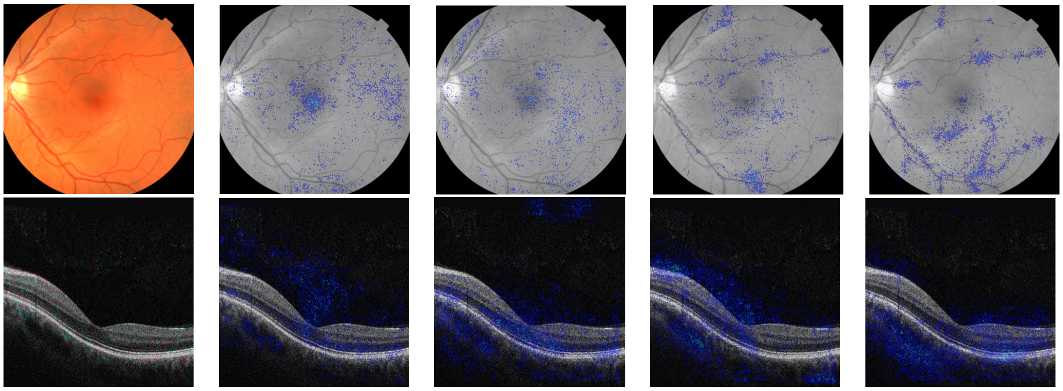
Sex single mod.

Sex multimodal

Hypertension single mod.

Hypertension multimodal

Figure A5: Saliency maps of scans from female with hypertension.



Original

Sex single mod.

Sex multimodal

Hypertension single mod.

Hypertension multimodal

Figure A6: Saliency maps of scans from male without hypertension.