Dataset, Baseline and Evaluation Design for GAVE Challenge

Zhiwei Liu[1,4], Huihui Fang[2(✉)], Shandi Liu[5], Ping Zhang[3], Wenqun Xi[3], Changjian Wu[4], Zihao Zhong[4], Xinyu Fu[6], Qifan Yang[4], Xiangyu Chen[7], Mengxiong Luo[7], Xinya Hu[3], Jiongning Zhao[4], Mingkui Tan[1], Weihua Yang[3], and Yanwu Xu[4,8(✉)]

[1] School of Software Engineering, South China University of Technology, Guangzhou 510006, China
[2] College of Computing and Data Science, Nanyang Technological University, 639798, Singapore huihui.fang@ntu.edu.sg
[3] Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, Shenzhen 518040, China
[4] School of Future Technology, South China University of Technology, Guangzhou 510006, China yxwu@ieee.org
[5] Huaian Huaian Hospital, Huaian 223200, China
[6] School of Biomedical Sciences and Engineering, South China University of Technology, Guangzhou 510006, China
[7] Department of Ophthalmology, Affiliated Hospital of North Sichuan Medical College, Nanchong 637000, China
[8] Pazhou Lab, Guangzhou 510320, China

Abstract. Retinal vessel characteristics serve as crucial biomarkers for screening and diagnosing various diseases. Retinal vessel segmentation, particularly arteriovenous (A/V) segmentation, is a key step in enabling AI-assisted disease screening and diagnosis. Fundus photography, a non-invasive retinal imaging technique, is widely accessible and cost-effective. To advance AI applications in screening and diagnosing conditions such as diabetes and cardiovascular diseases, we collaborated with Medical Image Computing and Computer Assisted Intervention (MICCAI) 2025 to launch the Generalized Analysis of Vessels in Eye (GAVE) Challenge. This initiative provides a dataset with expert annotations for three research tasks: vessel segmentation, A/V segmentation, and quantitative biomarker measurement in fundus images. In the annotation process of the dataset, the fluorescein fundus angiography(FFA) paired with color fundus photos are introduced, which can provide a clearer and more accurate annotation reference than relying on color fundus photos alone. This is the first and biggest dataset to incorporate paired FFA into the vascular annotation of the artery/vein and provides arteriovenous ratio (AVR) parameter annotations. This paper describes the released dataset of 150 color fundus images with corresponding annotations, baseline methods for the three subtasks, and evaluation protocols. The GAVE Challenge is accessible at https://aistudio.baidu.com/competition/detail/1315.

Keywords: Retinal vessel segmentation · Artery/Vein Segmentation · Biomarker · GAVE Challenge.

1 Introduction

Color fundus photography offers an efficient non-invasive approach for observing the human retinal vasculature[5]. Alterations in the retinal vessels are associated with numerous diseases: arterial narrowing indicates hypertension[17], while venous dilation suggests diabetic retinopathy[15]. Diabetes and cardiovascular diseases impose a substantial global healthcare burden[18]. Deep learning models enables disease screening and diagnosis assistance using color fundus photographs[14,21], where accurate vascular segmentation and measurement of vascular morphological parameters (e.g., vessel diameter[8], AVR[1]) are critical steps[10]. Because different artery/vein changes represent distinct pathologies, the precise segmentation of vessels into arteries and veins is important, which enables biomarkers quantification, enhancing the models prediction performance and interpretability[4,20,9].

Currently, the number of publicly available color fundus arteriovenous vessel segmentation datasets is limited, like DRIVE[13], HRF[3], LES-AV[12], RITE[7], less than 50 color fundus images. Moreover, the manual annotation process is fraught with uncertainties[19], especially when annotators rely solely on color fundus photographs to identify and annotate vessels. These challenges have severely restricted the development of related fields. To advance research on arteriovenous vessel segmentation and index measurement, we in collaboration with MICCAI 2025, organized the GAVE Challenge. The aim is to provide a rich and accurately annotated dataset of color fundus photographs for arteriovenous annotation, which can be used for research on vessel segmentation, arteriovenous segmentation, and vascular biomarkers measurement. This paper primarily introduces the 150 color fundus photographs dataset released in the GAVE Challenge, including three subtasks (vessel segmentation, arteriovenous segmentation, and AVR measurement), and the method of using color fundus photos paired with FFA for labeling is described in detail. Then we provide baselines and elaborate on the evaluation methods.

2 Dataset

The 150 color fundus photographs are derived from a dataset collected during a previous fundus disease study at Shenzhen Eye Hospital in China. This study spanned six years and included individuals aged 18–60 years without obvious ocular pathologies. From the original dataset, a subset featuring high-quality imaging and comprehensive data modalities was carefully selected to serve as the annotated dataset. Each data instance encompasses a color fundus photograph of one subject's eye, a FFA image depicting the complete arterial filling of the corresponding eye, and at least one FFA image demonstrating the complete filling of all fundus blood vessels of the same eye. The objective is to utilize the FFA images in these two states to facilitate the discrimination and labeling of arteries and veins in the fundus. In contrast to color fundus photographs, FFA images exhibit superior imaging quality. They can clearly image the boundary

of vessel, the morphology of arterioles and venule, and the relative positional relationships among near vessels. In addition, the principle underlying fluorescein angiography allows physicians to acquire color fundus photographs with exclusive arterial filling and those with the complete filling of both arteries and veins during the angiography procedure. This approach, compared to relying solely on color fundus photographs, enables annotating physicians to more precisely and effortlessly distinguish arterial and venous vessels, particularly in the case of minute vessels and those regions that are challenging to differentiate due to imaging - related factors. The images were acquired using a variety of ophthalmic devices, including the Topcon NW400, Canon CR - 2 AF, KeHe VX - 10i, Zeiss VIUCAM200, and other commonly employed fundus cameras in ophthalmic hospitals. This study was approved by Institutional Review Board of Shenzhen Eye Hospital (Approval 2024KYPJ013) and adhered to the Declaration of Helsinki. The original acquired images were stored in JPG format at a resolution of 1536×1024. The dataset images are saved in PNG format at the same resolution of 1536×1024. An illustrative example of one such image is presented in Fig. 1.

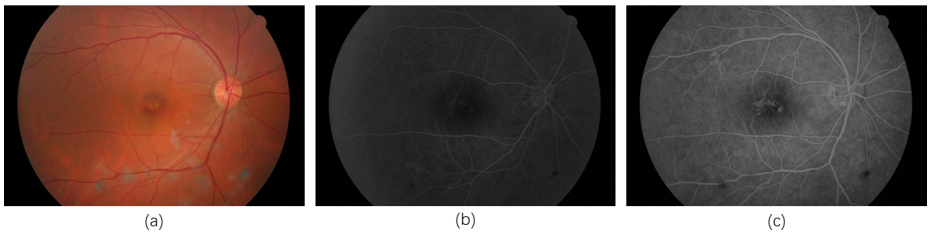

Fig. 1. Illustration of color fundus photography and paired FFA angiography process. (a) represents the color fundus photo, (b) represents the FFA images with filled arteries, and (c) represents the FFA image with both arteries and veins filled

The GAVE dataset provides segmentation masks for retinal vessels in color fundus images, arterial and venous classification labels, and average AVR measurement values for the top 4 vessels around the optic disc. The manual annotation team consists of one quality control doctor with higher seniority, three primary annotating doctors, and five image annotation support personnel. These 150 images were randomly divided into three subsets, each annotated by one doctor, supported by assistants and supervised throughout by the quality control doctor. During annotation, the three annotating doctors used the color fundus images and FFA images in each case to assist and followed a unified annotation standard and specification. Specifically, each image included color fundus images, corresponding FFA images with arteries filling and the FFA with both arteries and veins filled , and a pre-segmented mask using a fine-tuned pre-trained deep learning model. The FFA images and the pre-segmented mask were used to help

the doctors make judgments and annotations. The annotating doctors first observed the FFA filled arteries images and those with both arteries and veins filled, identified venous vessels, especially the tiny vessels at the ends, the interlaced vessels, and the parts that are difficult to observe and distinguish in the color fundus images. For some occlusions due to imaging or lesions, the FFA images could also assist the doctors in identifying and annotating the vessels. Then the annotating doctors used the annotation tools to draw the edges of the venous vessels and obtained the vessel masks by filling the contours. The same process was completed for the annotation of arteries with the cooperation of annotating doctors and annotation support personnel, and the doctors supervised and checked the annotations. The separately annotated arterial and venous vessel masks were merged and processed further after being checked and verified by the doctors. For overlapping areas of arterial and venous vessels, separate annotations were made to ensure the continuity of the single class vessel masks, and the overlapping parts were handled separately in the final annotation results. In the final segmentation masks, red represents arteries, blue represents veins, and green represents the overlapping parts of the vessels, we present three examples in Fig. 2.

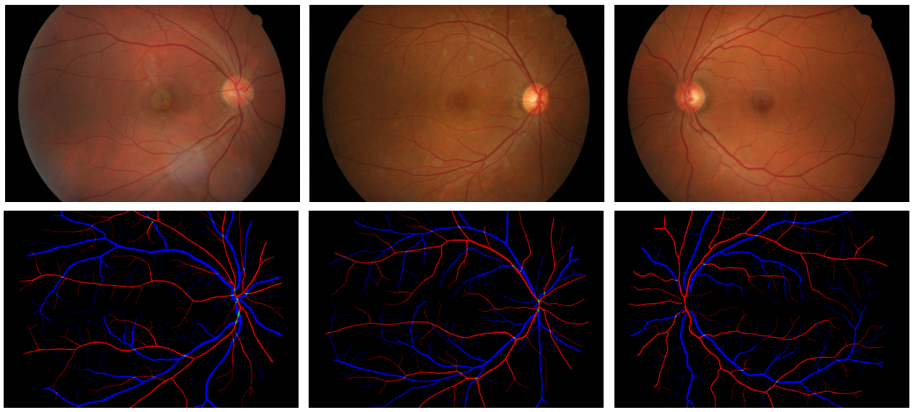

Fig. 2. Three examples of color fundus images with their corresponding pixel-wise manual annotations in GAVE dataset.

The data on the ratio of arteriovenous diameters is based on the labels made by doctors on the arteriovenous vessels and the optic disc. By using both labels, the diameters of the arteriovenous vessels around the optic disc are obtained, and the diameters of the largest arteries and veins are calculated. Finally, the AVR of the top four groups of vessels are calculated as the final reference standard. An example optic disk label and its corresponding AVR annotation are shown in Fig. 3. Finally, in the released dataset, 150 color fundus photos and corresponding arteriovenous labels, AVR measurements are included. The 150 images in the

dataset will be randomly divided into three groups. In the preliminary stage, 50 labeled images will be provided for model training, another 50 labeled images will be used for validation, and the remaining 50 images will be used for the final test.

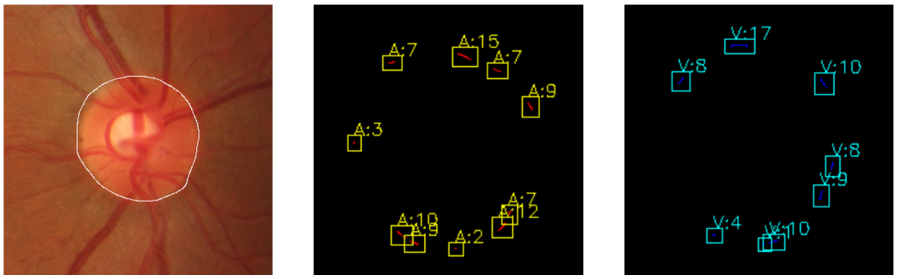

Fig. 3. Measurment process of AVR. Figure (a) presents manually labeled optic disc contour with 1 pixel width, (b) and (c) show arterial diameter section surface and venous diameter section obtained based on optic disc contour and vascular labeling, respectively. Numbers around bounding box represent the number of pixels of the blood caliber width.

3 Baseline

We design a baseline model for all three challenge sub-tasks. Inspired by [11], we propose a U-shaped model with recursive framework for vessel segmentation and arteriovenous segmentation simultaneously. As shown in Fig. 4, we introduce dense skip-connection and Vision Transformer(ViT) for better multiscale representative fusion and enhancing global context extraction, respectively. Through dense skip links, patches of different sizes from other layers can be aggregated together to serve as supplementary, thereby reducing information loss. To compensate for the limited receptive fields of convolutional layers, we incorporate a pre-trained ViT module in the deepest network stage to capture long-range dependencies. The recursive framework allows the model to continually improve the segmentation results of the previous model by correcting classification errors in the input results; we set the recursion depth 7 in experiment. During training, the weight of the loss of the first layer is the highest because the subsequent iterations are all based on its results. A decaying weight scheme is applied to the loss terms of deeper network layers during iterative training, which the weight equals to the ratio of its layer index to the cumulative sum of preceding layers.

The baseline is implemented in PyTorch and trained on an NVIDIA A6000 GPU (48GB memory). Input images are resized to 70% of their original dimensions to accommodate GPU memory constraints. We use pre-processing following [11]. For GAVE dataset, 50 images were used for training and 50 for testing.

During training, we use an Adam optimizer with learning rate = 10^{-4} and early stopping.

Following clinical practice experience, the diameters of the four largest vessels around the optic disc are used as the source for calculating the AVR. To measure AVR, we first use the pre-trained optic disc segmentation model [6] to segment the color fundus photo to obtain the contour of the optic disc, and then process it with the segmentation results of the second task to obtain the artery and vein diameters at the boundary of the optic disc, and calculate the average diameter of the top 4 arteries and veins to obtain the final AVR. The codes of our baseline are available at `https://github.com/liuzw20/GAVE`.

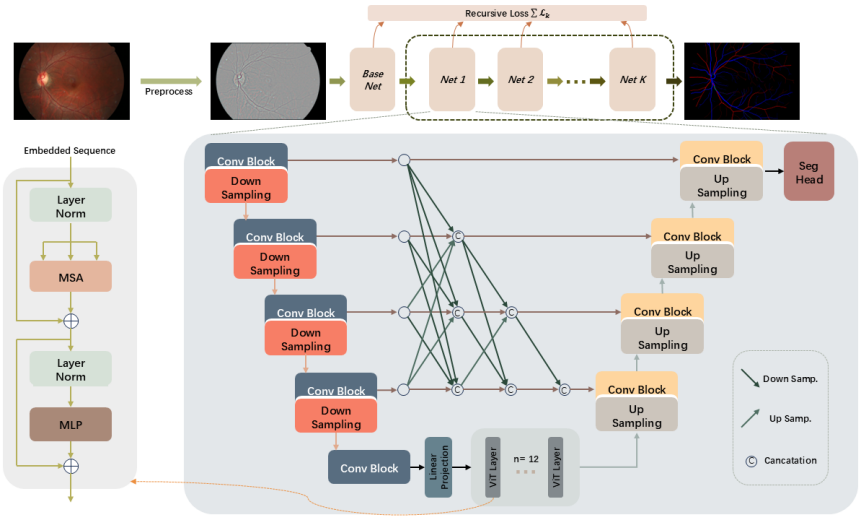

Fig. 4. The model structure of baseline

4 Evaluation

4.1 Task 1: Retinal Vessel Segmentation

To measure the accuracy of the segmentation region, we use the Dice Similarity Coefficient(DSC), which is a commonly used metric for evaluating segmentation tasks. DSC $= \frac{2\cdot|X\cap Y|}{|X|+|Y|}$, where X represents the set of segmented target pixels in the ground truth, Y represents the set of segmented pixels in the prediction result, $|X\cap Y|$ denotes the number of pixels in the intersection between X and Y, $|X|$ and $|Y|$ denote the number of elements (i.e., pixels) in X and Y, respectively. To assess the accuracy of the segmentation results on vascular boundaries, we use the 95% Hausdorff Distance (HD95) as the evaluation metric. HD95$(X, Y) =$

$\max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}_{95\%}$, where X and Y denote boundary points sets of the groundtruth and the prediction, respectively. $d(x,y)$ is the Euclidean distance between the points $x \in X$ and $y \in Y$. $\inf_{y\in Y} d(x,y)$ represents the minimum distance from the point x to the set Y. In addition, to evaluate tubular structure segmentation, such as blood vessels, the centerline Dice (clDice) index [16] is applied to assess topological connectivity. Our baseline achieves DSC of 0.7369, HD95 of 2.2361 and clDice of 0.2611. When designing the scoring calculation, we set the proportions of Dice score, HD95 score, and clDice score to 0.4, 0.3, and 0.3 according to the importance of metrics, respectively.

$$\text{Score}_{\text{task1}} = 10 \times \left(0.4 \times \text{DSC} + 0.3 \times \frac{3}{3 + \text{HD95}} + 0.3 \times \text{clDice}\right) \tag{1}$$

4.2 Task 2: Artery/Vein Segmentation

For artery/vein segmentation, we also choose DSC. Three evaluation indicators from different perspectives of vascular classification attributes are selected, Sen $= \frac{TP}{TP+FN}$, Spe $= \frac{TN}{TN+FP}$, Acc $= \frac{TP+TN}{TP+FN+FP+TN}$, where TP, TN, FP, and FN denote true positive, true negative, false positive and false negative, respectively. Inspired by [2], we used infeasible (INF) and correct (COR) path percentage to evaluate the topological connectivity of the A/V segmentation map. When calculating INF and COR, the algorithm randomly samples two points on the real centerline and finds their nearest corresponding points in the predicted mask. It then compares whether the shortest path lengths between the two pairs are consistent (allowing for a 10% error) to evaluate the topological connectivity of the predicted mask. If the two points are not connected in the prediction, it is recorded as an infeasible path. If the path length is significantly deviated(more than 10%), it is judged as an incorrect connection; otherwise, it is considered a correct connection. Because too long paths indicate missing links, whereas too short ones indicate hallucinated connections. We randomly sample 100 paths during the testing phase. Finally, by statistically analyzing the proportions of the three types of paths (infeasible, shorter/larger, and correct), the topological consistency of the prediction results is quantified. The higher value of COR and the lower value of INF implies better performance. The results of each specific evaluation metric are shown in Table 1. When designing the score calculation, we set the proportions of the segmentation evaluation indicators, classification evaluation indicators, and topological connectivity evaluation indicators in the total score to 0.3, 0.3, and 0.4, respectively.

$$\begin{aligned}\text{Score}_{\text{task2}} = 10 \times \big(&0.3 \times \text{DSC} + 0.1 \times (\text{Sen} + \text{Spe} + \text{Acc}) \\ &+ 0.2 \times (1 - \text{INF}) + 0.2 \times \text{COR}\big)\end{aligned} \tag{2}$$

4.3 Task 3: Arteriovenous Ratio Automatic Measurement

In order to assess the performance of regression task, We adopt the Mean Absolute Error (MAE), which is a commonly used metric for evaluating regression

Table 1. Evaluation results for artery and vein segmentation.

	DSC	Sen	Spe	Acc	INF ↓	COR
Artery	0.6009	0.4981	0.9952	0.9807	0.8478	0.1511
Vein	0.7383	0.7448	0.9879	0.9775	0.7048	0.2924

tasks. We also used the Symmetric Mean Absolute Percentage Error (SMAPE) to calculate the difference between the predicted values and the target values. $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, $\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$, where n is the number of samples, y_i is the groundtruth value, and $\hat{y}_i$ is the predicted value. The baseline achieves MAE of 0.2511 and SMAPE of 34.02%. When designing the score calculation, we set the ratio of the MAE score to the SMAPE score to 0.5 and 0.5 respectively.

$$\text{Score}_{\text{task3}} = 10 \times (0.5 \times \frac{0.37665}{0.37665 + \text{MAE}} + 0.5 \times (1 - \frac{\text{SMAPE}}{2})) \tag{3}$$

Based on the evaluation criteria, our baseline achieves a score of 5.4605 for the vessel segmentation task, a score of 5.4918 for the artery vein segmentation task and a score of 7.1495 for the final AVR measurement task. Arteriovenous segmentation has a higher level of task complexity and clinical value. Therefore, when calculating the total score of the task, we give task 2 a higher weight:

$$\text{Score}_{\text{round}} = 0.3 \times \text{Score}_{\text{task1}} + 0.4 \times \text{Score}_{\text{task2}} + 0.3 \times \text{Score}_{\text{task3}} \tag{4}$$

where $\text{round} \in \{\text{preliminary}, \text{final}\}$. Based on the task total score formula, the score of our baseline model on the preliminary set is 5.9797. Since the leaderboard of the preliminary competition is visible to all players, players can adjust the model parameters or strategies to obtain the best prediction on the preliminary set. To prevent players' results from overfitting on the preliminary dataset and obtaining higher scores, we assign a lower weight to the preliminary score when calculating the total challenge score. Therefore, the total score is:

$$\text{Score}_{\text{total}} = 0.3 \times \text{Score}_{\text{preliminary}} + 0.7 \times \text{Score}_{\text{final}} \tag{5}$$

5 Conclusion

In this article, we introduce the GAVE challenge at MICCAI 2025. We designed three subtasks based on the new GAVE dataset, including color fundus image vessel segmentation, arteriovenous segmentation, and automatic measurement of the AVR. In the dataset annotation process, we innovatively introduced two paired FFA images for accurate annotation, which represented the state of FFA images arteries filled and both arteries and veins filled, respectively. To the best of our knowledge, this is the first and largest vessel segmentation dataset which is annotated in this way and provides AVR labels as well. We further proposed a novel recursive framework and measurement method as baseline for GAVE dataset. Finally, we illustrate specific metrics for the evaluation of three tasks.

References

1. Akbar, S., Hassan, T., Akram, M.U., Yasin, U.U., Basit, I.: Avrdb: annotated dataset for vessel segmentation and calculation of arteriovenous ratio. In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). pp. 129–134. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2017)
2. Araújo, R.J., Cardoso, J.S., Oliveira, H.P.: A deep learning design for improving topology coherence in blood vessel segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. pp. 93–101. Springer (2019)
3. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust vessel segmentation in fundus images. International journal of biomedical imaging **2013**(1), 154860 (2013)
4. Chen, W., Yu, S., Ma, K., Ji, W., Bian, C., Chu, C., Shen, L., Zheng, Y.: Tw-gan: Topology and width aware gan for retinal artery/vein classification. Medical Image Analysis **77**, 102340 (2022)
5. Fogel-Levin, M., Sadda, S.R., Rosenfeld, P.J., Waheed, N., Querques, G., Freund, B.K., Sarraf, D.: Advanced retinal imaging and applications for clinical practice: A consensus review. Survey of ophthalmology **67**(5), 1373–1390 (2022)
6. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018. Lecture Notes in Computer Science, vol. 11071, pp. 1597–1605. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_177
7. Hu, Q., Abràmoff, M.D., Garvin, M.K.: Automated separation of binary overlapping trees in low-contrast color retinal images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16. pp. 436–443. Springer (2013)
8. Ikram, M.K., de Jong, F.J., Vingerling, J.R., Witteman, J.C., Hofman, A., Breteler, M.M., de Jong, P.T.: Are retinal arteriolar or venular diameters associated with markers for cardiovascular disorders? the rotterdam study. Investigative ophthalmology & visual science **45**(7), 2129–2134 (2004)
9. Kang, W., Li, B., Papma, J.M., Jiskoot, L.C., Deyn, P.P.D., Biessels, G.J., Claassen, J.A., Middelkoop, H.A., Flier, W.M.v.d., Ramakers, I.H., et al.: An interpretable machine learning model with deep learning-based imaging biomarkers for diagnosis of alzheimer's disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 69–78. Springer (2023)
10. Mookiah, M.R.K., Hogg, S., MacGillivray, T.J., Prathiba, V., Pradeepa, R., Mohan, V., Anjana, R.M., Doney, A.S., Palmer, C.N., Trucco, E.: A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. Medical Image Analysis **68**, 101905 (2021)
11. Morano, J., Aresta, G., Bogunović, H.: Rrwnet: Recursive refinement network for effective retinal artery/vein segmentation and classification. Expert Systems with Applications **256**, 124970 (2024)
12. Orlando, J.I., Barbosa Breda, J., Van Keer, K., Blaschko, M.B., Blanco, P.J., Bulant, C.A.: Towards a glaucoma risk index based on simulated hemodynamics from

fundus images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 65–73. Springer (2018)
13. Qureshi, T.A., Habib, M., Hunter, A., Al-Diri, B.: A manually-labeled, artery/vein classified benchmark for the drive dataset. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. pp. 485–488. IEEE (2013)
14. Rahman, A.U., Alsenani, Y., Zafar, A., Ullah, K., Rabie, K., Shongwe, T.: Enhancing heart disease prediction using a self-attention-based transformer model. Scientific Reports **14**(1), 514 (2024)
15. Safi, H., Safi, S., Hafezi-Moghadam, A., Ahmadieh, H.: Early detection of diabetic retinopathy. Survey of ophthalmology **63**(5), 601–608 (2018)
16. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16560–16569 (2021)
17. Tapp, R.J., Owen, C.G., Barman, S.A., Welikala, R.A., Foster, P.J., Whincup, P.H., Strachan, D.P., Rudnicka, A.R., Eye, U.B., Consortium, V.: Associations of retinal microvascular diameters and tortuosity with blood pressure and arterial stiffness: United kingdom biobank. Hypertension **74**(6), 1383–1390 (2019)
18. Wong, N.D., Sattar, N.: Cardiovascular risk in diabetes mellitus: epidemiology, assessment and prevention. Nature Reviews Cardiology **20**(10), 685–695 (2023)
19. Wu, J., Fang, H., Zhu, J., Zhang, Y., Li, X., Liu, Y., Liu, H., Jin, Y., Huang, W., Liu, Q., et al.: Multi-rater prism: Learning self-calibrated medical image segmentation from multiple raters. Science Bulletin **69**(18), 2906–2919 (2024)
20. Zhao, Y., Xie, J., Zhang, H., Zheng, Y., Zhao, Y., Qi, H., Zhao, Y., Su, P., Liu, J., Liu, Y.: Retinal vascular network topology reconstruction and artery/vein classification via dominant set clustering. IEEE transactions on medical imaging **39**(2), 341–356 (2019)
21. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)