

RetinaFusionNet: Gated Multimodal Fusion for Classifying Carotid Intima-Media Thickness

Mahir Shadid¹[0000–0001–7317–1714] and Rashed Mustafa²[0000–0001–5123–194X]

¹ International Islamic University Chittagong, Chittagong, Bangladesh

² University of Chittagong, Chittagong, Bangladesh
mahir.shadid@gmail.com¹, rashed.m@cu.ac.bd²

Abstract. Retinal fundus ultrasound-derived CIMT is well-established as a reliable surrogate measure of subclinical atherosclerosis and cardiovascular risk, but its acquisition demands specialized equipment and expertise, limiting its scalability for mass screening. Retinal fundus image analysis and demography can be considered promising complementary methods to predict carotid IMT. In this study, RetinaFusionNet is presented, a novel multimodal deep learning framework combining bilateral fundus images and demography through a learned gating system that adaptively modifies demographic data prior to fusion. All experiments were performed using the publicly available China-Fundus-CIMT dataset containing 2,903 patients according to its officially released patient-level train, validation, and test split. EfficientNet-B0, ResNet50, and DenseNet 121 were investigated in a common-weights dual-eye architecture including automated retina cropping, advanced data augmentation, label smoothing, class weighted loss, horizontal flip test-time augmentation, optimized thresholds on validation, and early stopping. The best gated model, EfficientNet-B0, reached an accuracy of 89.00% and ROC-AUC of 94.94% (95% bootstrap confidence interval: 90.32%-98.38%). The matched ungated models obtained ROC-AUC values of 92.42%, 94.11%, and 88.38% for EfficientNet-B0, ResNet50, and DenseNet121, respectively. The probability averaged from the three gated backbones obtained an ROC-AUC of 93.68%, in contrast to 93.08% of the respective ungated ensemble. These findings demonstrate the possibility of multimodal retinal analysis with demographic fusion for risk classification of non-invasive CIMT, but with caution due to the limited 100-patient test set.

Keywords: Carotid intima-media thickness · Fundus imaging · Multimodal learning · Deep learning · Medical image classification · RetinaFusionNet · Cardiovascular risk prediction

1 Introduction

Cardiovascular disease (CVD) continues to be the primary cause of death across the globe, thus justifying the need for early, non-invasive risk stratification, which has been an aim of preventive cardiology. Carotid intima-media thickness (CIMT), assessed via B-mode ultrasound, has already been recognized as a

surrogate indicator of subclinical atherosclerosis and future cardiovascular events [12, 11]. However, the assessment of CIMT necessitates specific equipment and expertise, restricting its utility in mass screening especially in economically deprived areas.

Fundus photography serves as a quicker, non-invasive, and easy-to-use substitute. Several investigations have demonstrated that deep learning systems are capable of detecting cardiovascular risk factors from fundus images [4], given the anatomical and physiological similarity between retinal and systemic microcirculation [13].

Guo et al. presented the China-Fundus-CIMT dataset, which includes bilateral fundus images, CIMT values, and demographic data for 2,903 individuals. They proposed a SE-ResNeXt50-based multimodal architecture with an improvement in test ROC-AUC from 82.58% to 90.41% by using fundus images along with age and gender via feature concatenation [2]. This research extends their previous work based on a Siamese ResNeXt model [1].

The aim of this work is to study the role of multimodal demographic fusion in RetinaFusionNet, where bilateral fundus images are fused using the same architecture with shared weights, L2 normalized feature averaging, and learnable demographic gate. Models like EfficientNet-B0, ResNet50, and DenseNet121 were compared with automated retina cropping, data augmentation, horizontal flip test time augmentation (TTA), validation based threshold optimization, and probability averaging ensemble prediction methods.

Contributions of this paper include the proposed gated multimodal framework named RetinaFusionNet for the adaptive integration of demographic information along with bilateral retinal features, and a comparison study of different pretrained backbones in the framework of RetinaFusionNet. An ablation study was done to analyze the contributions of demographic fusion, gating, TTA, and ensemble prediction, with the best gated EfficientNet-B0 performing ROC-AUC of 94.94% on the test set.

2 Related Works

Cardiovascular Risk Assessment Using Retinal Imaging: Deep learning has also become more popular in using retinal fundus images to predict cardiovascular risk. In a systematic review conducted by Li et al., most studies have utilized convolutional neural networks, but fewer studies have included clinical parameters and external validations [4]. Such findings can inspire multimodal methods, combining retinal images with demographics based on consistent evaluation methods.

Beyond retinal imaging, multimodal and human-centred healthcare AI increasingly combines heterogeneous patient information while emphasising explainability, privacy, and personalised decision support. Mahmud et al. investigated explainable and privacy-preserving AI for personalised healthcare applications [7], privacy-preserving multi-agent systems for human-centred inclusive healthcare AI [6], and explainable multimodal digital-human modelling for ac-

cessibility and inclusion [5]. These developments support the broader motivation for integrating complementary clinical and imaging information in interpretable healthcare systems. Zhu et al. further reviewed oculomics and its potential for using retinal markers to characterise systemic cardiovascular disease [13].

Carotid Intima-Media Thickness Prediction: Another application of machine learning techniques is to predict the development of CIMT through use of clinical biomarkers. The work of Zhou et al. [12] and Zhang et al. [11] is an example in which several machine learning models were used to predict CIMT. For fundus-based prediction of CIMT, Gong et al. developed a Siamese ResNeXt model based on retinal images of type 2 diabetes patients [1]. This work was later extended by Guo et al., who released the China-Fundus-CIMT dataset and proposed a multimodal SE-ResNeXt50 benchmark combining fundus images with demographic information [2].

Choice of Backbone in Medical Imaging: Choice of an appropriate backbone architecture continues to be empirical in medical image analysis tasks. According to Kim et al., the EfficientNet-B0 offers a good trade-off in terms of both performance and complexity for classification of retinal images [3]. Ojha et al. evaluated various architectures of CNNs for classifying ocular diseases and also showed that ensembling of multiple backbone architectures increases the model performance [8]. It leads to the comparison of EfficientNet-B0, ResNet50, and DenseNet121 with similar settings.

Bilateral Fundus Fusion: The use of bilateral fundus images processed by shared-weight convolutional branches has emerged as a popular choice for making use of complementary information present in the two eyes. The China-Fundus-CIMT dataset utilizes the shared-weight approach using SE-ResNeXt50 branches [2]. RetinaFusionNet also utilizes the shared-weight model but uses L2-normalization of features for fusion prior to multimodal fusion with gating.

Class Imbalance for Medical Image Classification: Class imbalance in medical image classification tasks is prevalent and tends to reduce the effectiveness of such tasks. Scholz et al. showed that using class-balancing loss functions helps improve robustness under imbalanced classes [10]. As there are about 71% thickened and 29% normal samples in the China-Fundus-CIMT dataset, the RetinaFusionNet applies class-weighted binary cross-entropy loss with label smoothing for class imbalance [2].

Explainability in Medical Deep Learning: Saliency maps are one of the popular post-hoc techniques used for convolutional neural networks in medical imaging as it effectively indicates which areas of the images affect model decisions without any modification to architecture [9]. It has been shown by several recent works that saliency maps can be used for validating if deep learning models use clinically relevant retinal areas for decision making. For instance, Ojha et al. [8] used saliency maps to validate if deep learning models for ophthalmic disease classification use clinically relevant retinal areas or artifacts instead. The same approach is followed here where saliency maps are generated to see if *RetinaFusionNet* uses clinically relevant retinal areas or preprocessing artifacts.

3 Dataset

In the current study, the publicly available China-Fundus-CIMT dataset [2] was utilized. This dataset contains bilateral high-resolution fundus images, CIMT measurements, as well as demographic data (age and gender) of 2,903 participants recruited from the Second Affiliated Hospital of Anhui Medical University from April 2015 to February 2024. Each participant provides two fundus images, one per each eye. CIMT was measured by carotid B-mode ultrasound and dichotomized at the 0.9 mm clinical threshold, resulting in 849 normal (29.2%) and 2,054 thickened (70.8%) samples.

The dataset contains an official patient-level split of 2,603 training patients, 200 validation patients, and 100 test patients, which is used as such in order to allow direct comparison with the benchmark. The test data are balanced (50 normal, 50 thickened cases), while the entire dataset is class imbalanced.

Each patient record also consists of *True age*, its min-max normalization, and gender binary label (0 = female, 1 = male). Demographic statistics of this dataset correspond to the values presented in the original publication (1,862 males and 1,041 females). Sample fundus images are presented in Fig. 1.

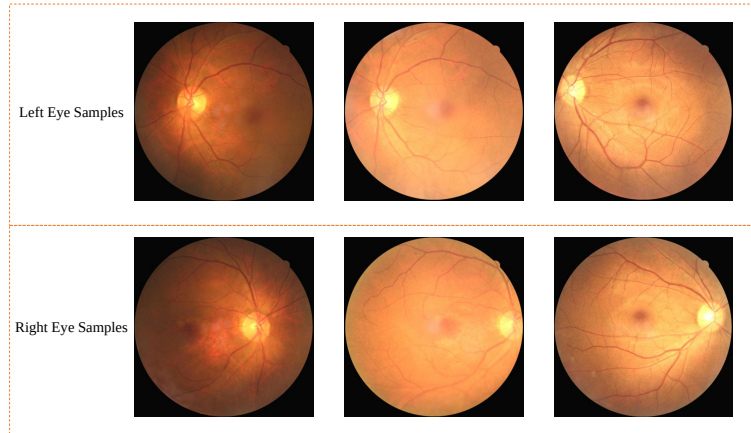


Fig. 1. Samples of left and right eye fundus images of three distinct patients.

4 Methodology

4.1 Overall Architecture: RetinaFusionNet

The suggested *RetinaFusionNet* system architecture is shown in Fig. 2 and consists of the combination of bilateral fundus images and patient demographics via

three modules: (i) shared-weight image branch; (ii) gated demographic branch; and (iii) fusion and classification head.

Image branch: Bilateral fundus images are processed by a convolutional backbone of shared weights similar to a Siamese model resulting in two feature vectors. Then, each of these feature vectors undergoes L2-normalization followed by element-wise averaging, forming an image representation, in such a way that both eyes contribute equally despite different magnitudes of the features extracted by the backbone.

Gated demographic branch: Age and gender values are normalized and converted into a demographic embedding using a fully connected branch. Unlike concatenation of the demographic embedding with the image representation, the demographic embedding is multiplied by a learned scalar gate ($g \in (0, 1)$), obtained using both modalities with a lightweight sigmoid-activated network.

Fusion and classification head: The image representation and demographic embedding, both weighted by their corresponding gates, are concatenated and fed into a fully connected classifier in order to predict a binary label of CIMT presence (*normal* or *thickened*).

4.2 Training Configuration and Evaluation Protocol

The training images were augmented using synchronized horizontal flips, random rotations ($\pm 12^\circ$), brightness/contrast adjustment, affine transformation, coarse dropout, and light grid distortion, preserving correspondence between both eyes. Validation and test images were only resized and normalized. To address the 71%/29% class imbalance, label-smoothed weighted binary cross-entropy was used:

$$y' = y(1 - \epsilon) + 0.5\epsilon,$$

with $\text{pos_weight} = n_{\text{normal}}/n_{\text{thickened}}$. The backbones were fine-tuned in an end-to-end manner with *AdamW* using a learning rate of 3×10^{-5} for the backbone that was pretrained and 3×10^{-4} for the fusion head. Linear warm-up with cosine decay was employed with *LambdaLR*, with a weight decay of 5×10^{-4} , dropout of 0.30, label smoothing of 0.05, and gradient clipping of 1.0. A fixed seed of 42 and validation ROC-AUC-based early stopping were used, retaining the best validation-AUC checkpoint.

The uncertainty was computed based on 2,000 patient-level bootstrap samples with random state 42. The test set has one sample per patient, containing demographic information and images for both eyes. Therefore, the row-level bootstrapping procedure is essentially equivalent to the patient-level sampling. The horizontal-flip TTA was done by taking the average prediction of the original and horizontally flipped image. The threshold values of 0.5 and the validation-tuned value that maximizes the F1 score were used, where the latter was directly used on the test set.

EfficientNet-B0, ResNet50, and DenseNet121, pre-trained on ImageNet, were considered as shared-weight encoders for images. Fundus images were gray-scaled solely to detect black borders and crop them out; the RGB cropped images were

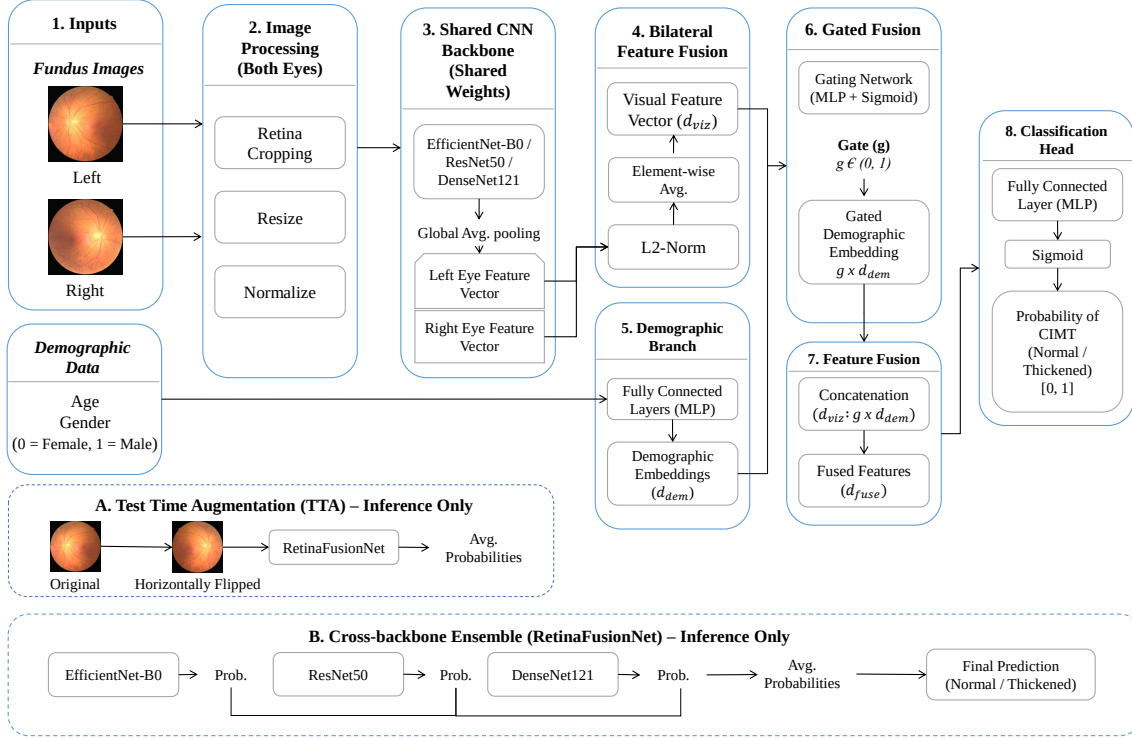


Fig. 2. Proposed RetinaFusionNet framework overview. Fundus images from both eyes are encoded using a shared-weight CNN backbone network, then merged using gated demographic feature vectors before classification with test-time augmentation and cross-backbone network ensembling.

then fed into the model. Evaluation was done using accuracy, precision, recall, F1-score, and ROC-AUC, with 95% ROC-AUC intervals being bootstrapped at patient level. Gradient-based saliency mapping was also performed on test images to check if the model was paying attention to retinal structures instead of background areas [8, 9].

5 Experimental Setup

All experiments used the official patient-level China-Fundus-CIMT split (2,603/200/100 patients) without re-splitting or resampling. The 100-patient test set was reserved for final evaluation, while checkpoint selection, validation-F1 threshold tuning, and ensemble construction used only the training and validation sets; each model was evaluated on the test set once. All backbones used identical pre-processing, augmentation, optimization, loss, and early-stopping settings.

EfficientNet-B0, ResNet50, and DenseNet121 were evaluated with gated demographic fusion and in the fundus-only setting. Results are reported at both the default threshold of 0.5 and the validation-tuned threshold. Unless stated otherwise, horizontal-flip TTA was applied. Probability averaging of the three backbones formed the *RetinaFusionNet Ensemble*.

Implementation used Python with PyTorch and TorchVision, with OpenCV, Albumentations, and Pillow for image processing and augmentation. NumPy, SciPy, scikit-learn, and Pandas were used for evaluation and statistical analysis, while saliency maps were generated using PyTorch automatic differentiation.

Table 1. Experimental environment.

Component	Specification
Programming Language	Python 3.13.0
Deep Learning	PyTorch 2.12.0+cpu, TorchVision 0.27.0+cpu
Image Processing	OpenCV 4.13.0, Albumentations 2.0.8, Pillow 12.2.0
Numerical Computing	NumPy 2.4.6, SciPy 1.17.1
Evaluation	scikit-learn 1.8.0, Pandas 3.0.3
CPU	Intel Core i5-11400H
GPU	NVIDIA GeForce RTX 3050 Ti Laptop (4 GB VRAM)
Memory	16 GB RAM
Operating System	Windows 11

6 Results

6.1 Backbone Comparison

Table 2 presents a comparison of RetinaFusionNet with three different backbone networks and the published SE-ResNeXt50 multimodal benchmark [2]. The results have been reported with test-time augmentation (TTA) at both the default (0.5) and validation-tuned decision thresholds.

All configurations of RetinaFusionNet outperform the public SE-ResNeXt50 baseline regarding accuracy and ROC-AUC metric apart from ResNet50 that performs 0.29% worse than the SE-ResNeXt50 in ROC-AUC. Among the tested architectures, EfficientNet-B0 demonstrated the best results being the leader in validation-tuned accuracy (89.00%, as good as the ensemble), F1 score (89.52%) and ROC-AUC (94.94%, 95% CI: 90.32-98.38%) for both all possible configurations of RetinaFusionNet and the ensemble, where the lower bound of the confidence interval is comparable with the public SE-ResNeXt50 point estimate of ROC-AUC (90.41%). The three-backbone ensemble was still competitive (93.68% ROC-AUC, 95% CI: 88.32-97.71%) improving specificity at the tuned threshold against the single backbone models but not beating EfficientNet-B0 on this test split. Although promising, this comparison should be interpreted cautiously because of the small test set.

Table 2. RetinaFusionNet performance comparison based on various backbone architectures in the test set of 100 patients. ROC-AUC 95% confidence intervals were estimated via bootstrapping 2,000 patient samples.

Model	Acc _{0.5}	Acc _T	Precision	Recall	F1	ROC-AUC (95% CI)
SE-ResNeXt50 [2]	–	81.00	–	88.00	80.91	90.41
RetinaFusionNet (EfficientNet-B0)	86.00	89.00	85.45	94.00	89.52	94.94 (90.32–98.38)
RetinaFusionNet (ResNet50)	81.00	86.00	83.33	90.00	86.54	90.12 (82.93–95.42)
RetinaFusionNet (DenseNet121)	84.00	83.00	78.95	90.00	84.11	93.23 (88.00–97.40)
RetinaFusionNet Ensemble	84.00	89.00	88.24	90.00	89.11	93.68 (88.32–97.71)

Note: Each RetinaFusionNet network is equipped with the suggested gated demographic fusion. Acc_{0.5} refers to accuracy at the traditional threshold of 0.5, whereas Acc_T refers to accuracy at the tuned threshold.

6.2 Ablation Study: Matched

Ablation studies were carried out using the same patient-level test protocol and final pre-processing pipeline to evaluate the contribution of demographic data, gated fusion, bilateral images, test-time augmentation (TTA), ensembling, and random seed. The respective ROC-AUC scores of these models are shown in Table 3. All runs except for the additional random seed runs were done with random seed 42. In the following notation, $T = 0.5$ indicates the standard classification threshold. However, due to the computation time constraint, the additional random seed runs were done only for the EfficientNet-B0 setup with seeds 64 and 92.

These results reveal that adding demographic data and gated fusion yields better results than the bilateral image-only approach. Moreover, even the single-eye models have comparable performances. The TTA and ensembling comparisons assess the contribution of the inference modules independently. The extra runs based on random seeds also reveal the impact of random seed on the performance estimates using the test set of 100 patients.

6.3 Training Dynamics and Overfitting

The highest validation ROC-AUC value of each backbone was achieved during the first five epochs: efficientnet-b0 backbone during epoch 5 (0.8688), ResNet50 backbone during epoch 2 (0.8766), and DenseNet121 backbone during epoch 3 (0.8537), as illustrated in Fig. 3. The training loss value of EfficientNet-B0 was reduced from 0.3511 to 0.2351, while its validation loss value was minimized at epoch 3 (0.3505), and then increased to 0.4018 at epoch 12. The validation ROC-AUC value of EfficientNet-B0 also fell to 0.8340 by epoch 12 although there were improvements during the training. Therefore, the checkpoint with the highest validation ROC-AUC was retained for subsequent evaluation.

6.4 Qualitative and Quantitative Explainability Analysis

Saliency maps for the two eyes of an enhanced case using RetinaFusionNet with EfficientNet-B0 are depicted in Fig. 4. The probability value for this case is

Table 3. Ablation analysis of RetinaFusionNet on the 100-patient test set.

Configuration	Demo.	Gate	TTA	Ensemble	ROC-AUC
Image-only (bilateral)	No	No	Yes	Yes	0.8564
Left-eye only	Yes	Yes	Yes	Yes	0.9314
Right-eye only	Yes	Yes	Yes	Yes	0.9236
Multimodal, ungated	Yes	No	Yes	Yes	0.9307
Multimodal, ungated (ResNet50)	Yes	No	Yes	No	0.9411
RetinaFusionNet (EfficientNet-B0)	Yes	Yes	Yes	No	0.9493
RetinaFusionNet (TTA, $T = 0.5$)	Yes	Yes	Yes	Yes	0.9500
RetinaFusionNet (no TTA, $T = 0.5$)	Yes	Yes	No	Yes	0.9544
RetinaFusionNet Ensemble	Yes	Yes	Yes	Yes	0.9367
Seed = 64 (EfficientNet-B0)	Yes	Yes	Yes	Yes	0.9143
Seed = 92 (EfficientNet-B0)	Yes	Yes	Yes	Yes	0.8940

Note: $T = 0.5$ denotes the conventional classification threshold. Unless otherwise specified, configurations use seed 42.

0.910. The saliency is primarily found inside the retinal areas as opposed to the black background.

Saliency analysis of six representative thickened instances revealed that the saliency was predominantly focused on the retinal foreground. The mean saliency of the foreground was $86.37\% \pm 2.41\%$, with the ratio of foreground to background being 1.82 ± 0.45 (Table 4). The top 10% of saliency mass contained $92.45\% \pm 3.97\%$ within the retinal foreground.

Table 4. Quantitative saliency analysis across representative thickened cases. Values are mean $\pm$ SD.

Metric	Mean	SD
Foreground saliency (%)	86.37	2.41
Background saliency (%)	13.63	2.41
Foreground/background density ratio	1.82	0.45
Top-10% foreground saliency (%)	92.45	3.97
Top-10% background saliency (%)	7.55	3.97

The Wilcoxon signed-rank test performed on $n = 6 \times 2 = 12$ eye samples revealed a statistically significant difference between the foreground and background saliencies ($W = 78.0$, $p = 0.000244$), indicating that the model’s saliency was predominantly concentrated within retinal regions.

6.5 Demographic Analysis of the Learned Gating Mechanism

The consistency of the learned demographic gating was estimated by examining gate values based on disease status, gender, and age on the test set of 100 patients

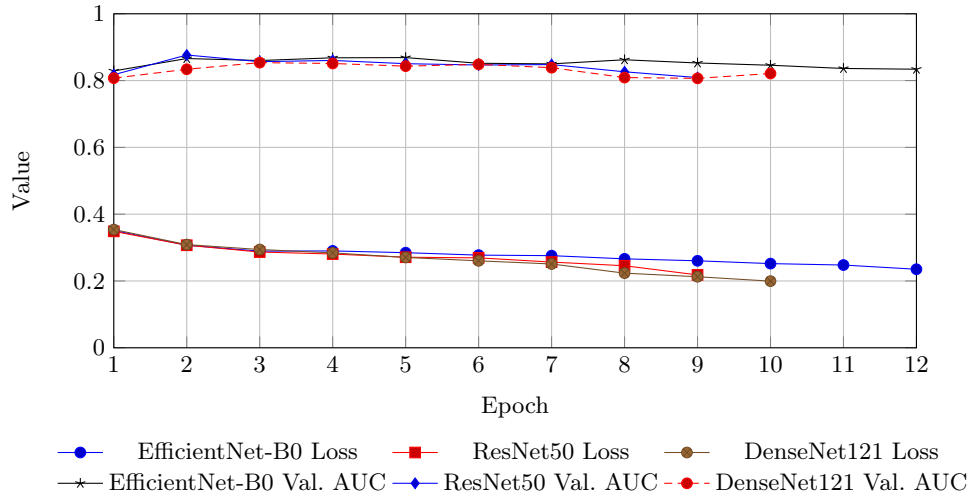


Fig. 3. Training and validation ROC-AUC curves for the three backbones. The EfficientNet-B0, ResNet50, and DenseNet121 achieved their maximum values of ROC-AUC on the validation set at epoch 5 (0.8688), epoch 2 (0.8766), and 3 (0.8537), respectively.

for each of the backbones. The median age of the test set was 47 years, and therefore the two age groups, 14–47 and 48–73 years, were created.

Table 5 shows that the gate value is the learned demographic weighting with its mean and standard deviation indicating its central tendency and variability in each group. High gate values were maintained consistently in all groups. The overall mean gate values were 0.9975, 0.9884, and 0.9958 for EfficientNet-B0, ResNet50, and DenseNet121, respectively. Differences in terms of disease status were negligible; for instance, EfficientNet-B0 gave 0.9974 ± 0.0015 and 0.9975 ± 0.0016 for normal and thickened cases, respectively.

Similarly, the sex and age exhibited small variations. The average gate value for the EfficientNet-B0 model was 0.9975 ± 0.0015 and 0.9975 ± 0.0016 for female and male patients, respectively, while 0.9975 ± 0.0014 and 0.9975 ± 0.0017 values were achieved for young and old individuals, respectively. ResNet50 and DenseNet121 models showed similar performance in the same way. These results indicate that the learned gate remained consistently active across the examined disease, sex, and age subgroups.

7 Discussion

7.1 Effect of Validation-Tuned Decision Thresholds

The impact of choosing a threshold in place of the traditional 0.5 threshold on the performance of models is shown in Table 6. The test-set accuracy of

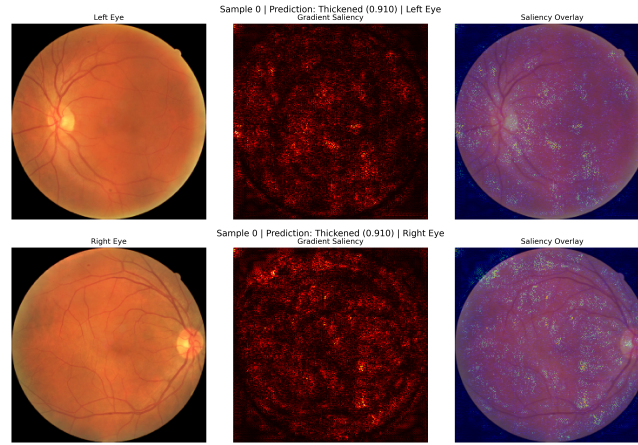


Fig. 4. Gradient-based saliency maps for both eyes from the same thickened case (probability: 0.910). Saliency is focused more on the retina than on the background.

EfficientNet-B0 model is increased from 86.00% to 89.00% and the F1 score from 87.50% to 89.52%. For the case of ResNet50, accuracy is increased from 81.00% to 86.00%, whereas the F1 score is increased from 82.57% to 86.54%. The test-set accuracy of DenseNet121 model is decreased from 84.00% to 83.00%, while F1 score is decreased from 85.45% to 84.11%. The most significant increase in test-set accuracy is observed for the ensemble from 84.00% to 89.00%, and the F1 score is increased from 85.45% to 89.11%.

The above outcomes show that the use of validation-tuning for threshold setting was beneficial for classification of test set data for EfficientNet-B0, ResNet50, and the ensemble model, but not for DenseNet121. Thus, both traditional and tuned thresholds are reported (Table 6).

7.2 Calibration and Diagnostic Performance

Confusion matrices are given for the decision thresholds both at default (0.5) and tuned values, along with calibration scores (Brier score, ECE; number of bins = 10) that were calculated using the predicted probabilities from the test set. Sensitivity and specificity are provided for the tuned decision threshold, complementing the accuracy/F1 measures presented in Table 7.

The calibration process is largely the same for all backbone models, with the Brier score ranging between 0.1088 and 0.1309 and ECE values between 0.1185 and 0.1350. The ensemble performed the best in terms of specificity (0.880) and number of false positives (6), as opposed to 8–12 for each backbone individually. However, it was not the most sensitive model or had the largest ROC-AUC.

Table 5. Distributions of learned gate values for the subgroups based on status of the disease, gender, and age were calculated from the 100 patients in the test group. Data is presented as mean $\pm$ SD. Age groups were defined by the median age in the test set (47).

Model	Subgroup	N	Gate
<i>Disease status</i>			
EfficientNet-B0	Overall	100	0.9975 ± 0.0015
	Normal	50	0.9974 ± 0.0015
	Thickened	50	0.9975 ± 0.0016
ResNet50	Overall	100	0.9884 ± 0.0047
	Normal	50	0.9893 ± 0.0052
	Thickened	50	0.9875 ± 0.0040
DenseNet121	Overall	100	0.9958 ± 0.0028
	Normal	50	0.9963 ± 0.0027
	Thickened	50	0.9953 ± 0.0029
<i>Sex</i>			
EfficientNet-B0	Female	37	0.9975 ± 0.0015
	Male	63	0.9975 ± 0.0016
ResNet50	Female	37	0.9880 ± 0.0045
	Male	63	0.9886 ± 0.0048
DenseNet121	Female	37	0.9955 ± 0.0027
	Male	63	0.9960 ± 0.0029
<i>Age</i>			
EfficientNet-B0	14–47	50	0.9975 ± 0.0014
	48–73	50	0.9975 ± 0.0017
ResNet50	14–47	50	0.9890 ± 0.0052
	48–73	50	0.9878 ± 0.0040
DenseNet121	14–47	50	0.9961 ± 0.0031
	48–73	50	0.9956 ± 0.0025

8 Constraints

Some of the limitations of this study include: First, the test set is limited to just 100 patients, reducing the accuracy of the point estimates and any potential statistical comparison. While patient level 95% confidence intervals were obtained using bootstrap sampling, no formal paired significance test could be done since patient level predictions from the SE-ResNeXt50 [2] publication were unavailable.

Second, the dataset used, China-Fundus-CIMT, is collected from one single medical center from Anhui Province, China. Hence, its ability to generalize to other populations/devices/imaging protocols remains to be seen.

Table 6. Effect of validation-tuned decision thresholds on test-set performance.

Model	Threshold	Acc.	Prec.	Recall	F1
EfficientNet-B0	0.500	86.00	79.03	98.00	87.50
	0.550	89.00	85.45	94.00	89.52
ResNet50	0.500	81.00	76.27	90.00	82.57
	0.570	86.00	83.33	90.00	86.54
DenseNet121	0.500	84.00	78.33	94.00	85.45
	0.545	83.00	78.95	90.00	84.11
Ensemble	0.500	84.00	78.33	94.00	85.45
	0.615	89.00	88.24	90.00	89.11

Note: The threshold of 0.500 is the conventional classification threshold. Tuned thresholds were selected using the validation set and applied unchanged to the test set.

Table 7. Confusion matrices, sensitivity/specificity, and calibration metrics on the test set.

Model	Threshold	TN	FP	FN	TP	Sens.	Spec.	Brier	ECE
EfficientNet-B0	0.50	37	13	1	49	0.980	0.740	0.1088	0.1350
	Tuned (0.550)	42	8	3	47	0.940	0.840		
ResNet50	0.50	36	14	5	45	0.900	0.720	0.1309	0.1185
	Tuned (0.570)	41	9	5	45	0.900	0.820		
DenseNet121	0.50	37	13	3	47	0.940	0.740	0.1187	0.1275
	Tuned (0.545)	38	12	5	45	0.900	0.760		
Ensemble	0.50	37*	13*	3*	47*	0.940	0.740	0.1156	0.1350
	Tuned (0.615)	44	6	5	45	0.900	0.880		

*At the default 0.5 threshold, the ensemble’s hard predictions coincide exactly with DenseNet121’s on all 100 test cases, despite differing underlying probabilities.

Third, all three backbones were trained with the same settings without specific hyperparameter optimization for each backbone. Further tuning may lead to a change in their performance.

Lastly, while this study did conduct an extensive ablation study, it did not isolate and analyze every single pre-processing step and training step separately, including but not limited to, the retinal cropping and label smoothing steps.

9 Conclusion

This work presents RetinaFusionNet, a multimodal framework for CIMT risk classification using bilateral fundus images and demographic data using a learnable gating mechanism. EfficientNet-B0, ResNet50, and DenseNet121 were compared under identical training conditions, along with a probability averaging ensemble.

The framework uses automated retinal image cropping, L2 normalized bilateral feature averaging, gated demographic fusion, data augmentation, class

balanced label smoothed loss, horizontal flip TTA and validation-based threshold tuning. EfficientNet-B0 performed best achieving the highest ROC-AUC of 94.94% and tuned accuracy of 89.00%, while the ensemble model got 93.68% ROC-AUC and 89.00% accuracy. Gated models consistently surpassed the respective ungated and fundus-only baseline versions, validating the usefulness of demographic fusion.

Dynamics of training showed that validation ROC-AUC peaks at early epochs for all three backbone networks, followed by improved training performance and decreased validation performance. This indicates early overfitting and implies that the backbone capacity increase alone does not help to generalize better on this particular dataset. Ensemble model stayed competitive but did not surpass the most efficient backbone network.

Future research should involve multi-center external validation, increased sample size of the independent test set, additional component level ablation study and inclusion of more clinically relevant features.

Data and Code Availability

The dataset used in this paper, called China-Fundus-CIMT, can be accessed from the platform Springer Nature Figshare as described in the original data descriptor [2]. Code for RetinaFusionNet, including data processing and training, testing steps are available at: <https://github.com/mahirshadid2point0/RetinaFusionNet>.

References

1. Gong, A., Fu, W., Li, H., Guo, N., Pan, T.: A siamese resnext network for predicting carotid intimal thickness of patients with t2dm from fundus images. *Frontiers in Endocrinology* **15**, 1364519 (2024)
2. Guo, N., Fu, W., Li, H., Zhang, H., Li, T., Zhang, W., Zhong, X., Pan, T., Sun, F., Gong, A.: High-resolution fundus images for ophthalmomomics and early cardiovascular disease prediction. *Scientific data* **12**(1), 568 (2025)
3. Kim, S., Yun, W., Oh, J., Lee, S.: Efficient deep learning approaches for processing ultra-widefield retinal imaging. In: *MICCAI Challenge on Ultra-Widefield Fundus Imaging for Diabetic Retinopathy*, pp. 125–134 (2024)
4. Li, L.Y., Isaksen, A.A., Lebiecka-Johansen, B., Funck, K., Thambawita, V., Byberg, S., Andersen, T.H., Norgaard, O., Hulman, A.: Prediction of cardiovascular markers and diseases using retinal fundus images and deep learning: a systematic scoping review. *European Heart Journal-Digital Health* **5**(6), 660–669 (2024)
5. Mahmud, M., Baslyman, M., et al.: Inclusive ai for digital human modelling: A perspective on explainable multimodal digital twins for accessibility and inclusion. In: *International Conference on Human-Computer Interaction*. pp. 1–16. Springer (2026)
6. Mahmud, M., Shaffi, N., Kaiser, M.S., et al.: Privacy-preserving multi-agent systems for human-centred inclusive healthcare ai. *AI Magazine* **47**(3), e70087 (2026). <https://doi.org/https://doi.org/10.1002/aaai.70087>

7. Mahmud, M., et al.: Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **13309 LNCS**, 356 – 370 (2022). https://doi.org/10.1007/978-3-031-05039-8_26
8. Ojha, M.K., Kumar, T., Rai, A., Roy, A., et al.: Deep learning-based ophthalmic disease classification with ensemble modeling and grad-cam visualization. *INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES* pp. 1403–1417 (2026)
9. Raghavan, K., B, S., v, K.: Attention guided grad-cam: an improved explainable artificial intelligence model for infrared breast cancer detection. *Multimedia Tools and Applications* **83**(19), 57551–57578 (2024)
10. Scholz, D., Erdur, A.C., Buchner, J.A., Peeken, J.C., Rueckert, D., Wiestler, B.: Imbalance-aware loss functions improve medical image classification. In: *Medical imaging with deep learning* (2024)
11. Zhang, Y., Song, M., Wang, H., Du, L.: Increased carotid intima-media thickness in pediatric nephrotic syndrome: A meta-analysis. *Biomolecules and Biomedicine* **26**(3), 399 (2025)
12. Zhou, A., Chen, K., Wei, Y., Ye, Q., Xiao, Y., Shi, R., Wang, J., Li, W.D.: Machine learning-based prediction of carotid intima-media thickness progression: a three-year prospective cohort study. *Frontiers in Medicine* **12**, 1593662 (2025)
13. Zhu, Z., Wang, Y., Qi, Z., Hu, W., Zhang, X., Wagner, S.K., Wang, Y., Ran, A.R., Ong, J., Waisberg, E., et al.: Oculomics: Current concepts and evidence. *Progress in Retinal and Eye Research* **106**, 101350 (2025)