



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Cardiovascular Disease Risk Assessment Through Retinal Image Analysis Using Deep Learning

Shaik Matheen¹ · Kuda Nageswara Rao², Dondapati Kalyan³, Dondapati Satish Babu⁴

¹Department of Computer Science and Systems Engineering

Andhra University College of Engineering, Visakhapatnam, Andhra Pradesh, India – 530003. Email: matheen.sk9848@gmail.com

²Department of Computer Science and Systems Engineering

Andhra University College of Engineering, Visakhapatnam, Andhra Pradesh, India – 530003. Email: knraoauce@gmail.com

³Department of Computer Science & Systems Engineering Andhra University College of Engineering (A) Visakhapatnam, India. Email: dondapatikalyan123@gmail.com

⁴dondapatishatishbabu@gmail.com

Abstract

CVD is widely recognized to be the cause of death of almost 17.9 million people each year. This condition results in deaths of about one-third of all people [1]. As retinal vessels share the same developmental origin as coronary and cerebral microvasculature, fundus photography provides an unparalleled chance to assess the condition of cardiovascular system in a noninvasive manner. This study presents a model that allows to screen cardiovascular risk from retinal biomarkers in fundus images. The pipeline combines a clinically motivated preprocessing step — CLAHE contrast enhancement followed by Ben Graham local normalization — with three ImageNet-pretrained backbones (EfficientNetV2B3, InceptionV3, and ResNet50V2), each fine-tuned in two phases and combined through soft-voting of their sigmoid outputs. On the 724-image held-out test set, the ensemble achieved 96.4% accuracy, 0.987 AUC-ROC, 97.9% recall, and 93.5% specificity at the Youden-optimal threshold of 0.5127. The metrics are comparable to the IEEE CSNT 2025 baseline, although the data distributions and evaluation schemes were different in the two studies. Grad-CAM heatmaps allow spatial inspection of model attention, and Test Time Augmentation together with Youden's J threshold optimization improve reliability under realistic screening conditions. All experiments draw on a composite dataset of 7,223 fundus images pulled from eight of the nine sources used in the baseline study; we dropped one source over resolution incompatibility.

Index Terms—cardiovascular risk assessment, retinal biomarker screening, retinal fundus imaging, deep learning, transfer learning, ensemble learning, Grad-CAM interpretability.

I. INTRODUCTION

The most common cause of death is the combination of heart disease, stroke, hypertensive heart disease, and other vascular diseases included under CVD, which is estimated to claim an estimated 17.9 million lives annually according to the World Health Organization [1]. Some of the factors contributing to the difficulty associated with its early diagnosis include its long-term development over several years, sometimes even decades, during which the patient will experience no symptoms until they reach a critical point of heart attack or stroke where the chances for prevention are almost out of question. Assessment of the patient's cardiovascular risks is usually done by using serum lipid profile measurements, blood pressure measurement, electrocardiography, echocardiography, and for higher risk patients coronary angiography and nuclear stress testing. However, the problem with such methods is the fact that they require skilled professionals and specialized equipment and facilities, which are not available in many remote areas and underdeveloped healthcare systems – in some regions the number of specialists to population ratio is less than one for every 10,000 inhabitants [2]. As a result, population-level screening tends to be concentrated in wealthier health systems, and a large share of high-risk individuals go undetected until they present with acute symptoms. Retinal fundus imaging has emerged as a promising complementary screening tool for two reasons. Initially, due to the same origin in the embryo of retinal microvasculature, coronary and cerebral vessels, symptoms of diseases such as hypertension, dyslipidemia, and atherosclerosis may appear in the retina before any other part of the body, and therefore it provides insight into the health of the vasculature [3]. For example, the appearance of symptoms such as arteriovenous nicking, narrowing of arterioles, microaneurysms, and flame haemorrhages have been related to an increased risk of heart disease in numerous epidemiological studies. Second, fundus cameras are cheap and portable, and deep learning classifiers are already used in diabetic retinopathy screening programs [4] — so a

CVD risk classifier could piggyback on that same infrastructure without requiring new hardware. This paper describes a deep learning framework for cardiovascular risk prediction from retinal fundus images, which classifies images as indicating elevated or low cardiovascular risk based on established retinal biomarkers. Building on the IEEE CSNT 2025 baseline reported by Kayser et al. [5], we introduce six changes: (1) a preprocessing pipeline tailored to the domain, using CLAHE and Ben Graham normalization; (2) transfer learning across three complementary pretrained architectures; (3) a two-phase fine-tuning protocol; (4) soft-voting ensemble fusion; (5) Grad-CAM spatial interpretability; and (6) inference-time enhancements motivated by clinical need — Test Time Augmentation and Youden's J threshold optimization. Every single one of these is based on a certain limitation that we have found in the baseline. Section VI and Section VII deal with the architecture of the system and why each decision was made, while Section VIII is about the originality of the framework altogether. Section X presents results and discussion — accuracy, precision, recall, F1-score, ROC analysis, confusion matrix analysis, an ablation study, and Grad-CAM interpretability — followed by the prototype clinical interface in Section XI and limitations and future work in Section XII. In summary, this work contributes: (1) what is, to our knowledge, one of the few reported integrations of these six components into a single retinal cardiovascular risk screening framework; (2) an ablation study that quantifies each component's individual contribution; (3) a domain-context survey (Table V) that situates our results within the broader medical imaging literature, with the caveat that cross-task and cross-dataset comparisons are not strictly comparable; and (4) an explicit discussion of the label-proxy limitation, along with a concrete plan for prospective clinical validation.

II. LITERATURE SURVEY

The concept that a picture of the retina would be able to give information regarding the status of the heart is not something new to us. It has been shown before through large-scale epidemiological studies like the Atherosclerosis Risk in Communities (ARIC) cohort [6] and Blue Mountains Eye Study [7]. This work and similar efforts [8] build upon this body of literature that established the role of the retina as a window to the cardiovascular system. More recently, vision transformer models [9], [10] and multi-task learning approaches that simultaneously classify multiple cardiovascular diseases from a single fundus photograph [11] have appeared in the literature. To the best of our knowledge, the application of such architectures to multi-source cardiovascular disease retinal screening data has not been explored. The work by Poplin et al. [12] provided the first evidence that cardiovascular risk factors could be directly predicted from fundus photographs. Their convolutional model, trained on a large-scale ophthalmology cohort, showed remarkable performance in predicting age, gender, smoking status, blood pressure, and risk of major adverse cardiovascular events. This result was notable in part due to the surprising accuracy reported for the model's performance on these metrics, but also because it demonstrated that the retina contained a wealth of physiological information previously unexploited by clinicians. This finding spurred a line of research into deep learning-based biomarker discovery from retinal images [13], which is expanded upon in this work. Our most direct point of comparison is the IEEE CSNT 2025 paper by Kayser et al. [5], who compiled a nine-source dataset of 8,752 retinal images and trained a custom 16-layer CNN for CVD risk classification, reporting 94.4% training accuracy, 89.6% validation accuracy, 87.4% recall, and an AUC of 0.97. Because that paper did not report a separate held-out test evaluation, our comparisons use their best-reported validation figures. Their results demonstrated that automated retinal cardiovascular screening was feasible, but several gaps remained: no domain-specific preprocessing, a custom CNN trained from random initialization rather than pretrained weights, a fixed 0.5 classification cutoff that does not account for the unequal cost of false negatives versus false positives in a screening context, and no mechanism for clinicians to inspect how the model reached its decisions. As for classification of medical images, the transfer learning approach has often beaten the "from scratch" training, especially when the amount of training data is small. EfficientNetV2 [14] relies on compound scaling, that is, the joint tuning of depth, width, and input resolution of the network rather than their sequential tuning. InceptionV3 [15], on the contrary, uses a split of one large convolution operation into several smaller operations and the parallel processing of the input in multiple inception blocks with several kernel sizes. The pre-activation residual connections of ResNet50V2 [16] ensure proper gradients during fine-tuning and are based on the deep convolution architectures by Simonyan and Zisserman [17]. Graham's local normalization method [18] and CLAHE [19] have been successfully used in diabetic retinopathy competitions. Concerning the explainability of the neural networks, there is an approach that has been proposed by Selvaraju et al. [20] and has gained popularity – Grad-CAM, which allows to examine if a convolutional classifier pays attention to the diagnostic features and not the surrounding noise.

III. RESEARCH GAP

Looking across the existing literature, six elements stand out as rarely being combined within a single retinal cardiovascular risk screening framework: (1) ImageNet-pretrained transfer learning has seen little use on composite, multi-source CVD retinal datasets; (2) CLAHE paired with Ben Graham normalization has not been systematically tested in cardiovascular risk pipelines, despite good results in diabetic retinopathy tasks; (3) few studies compare multiple ensemble fusion strategies on this specific problem; (4) Grad-CAM interpretability is rarely built into retinal CVD screening systems; (5) prior work

tends to fix the decision threshold at 0.5, which is a poor choice given that false negatives are clinically far more costly than false positives; and (6) Test Time Augmentation for inference-time stabilization remains largely unexplored in this domain. This work addresses all six within a single, reproducible pipeline.

IV. DATASET ANALYSIS AND DESCRIPTION

All experiments draw on a composite retinal fundus image dataset built from eight of the nine repositories used by Kayser et al. [5]; one source, Messidor-2 (an extension of the Messidor dataset, distinct from it), was excluded because its resolution was incompatible with the rest. After the process of quality screening, 7,223 images were retained and these included five categories of retinal abnormalities which had proven links to cardiovascular diseases. They included microaneurysms, hard exudates, flame shaped hemorrhages, optic disc lesions and changes in the structure of the arterioles. For the selection of our sources, we used the same methodology as Kayser et al. [5]. Stratified partitioning was done in order to maintain the same class balance in the training, validation and test set and thereby ensure a more realistic evaluation process and minimize sampling bias. The sources include Messidor, e-Ophtha, IDRiD, DRIVE, STARE, CHASEDB1, DIARETDB0, and DIARETDB1. With such a variety of institutions as sources, there is bound to be a true variability in terms of camera hardware, sensor resolution, population demographics and lighting conditions. Instead of considering this as some form of noise and removing the same, it actually serves as an added advantage because of the reason stated above.

A. Ground Truth Labeling Protocol

Each image was classified as either CVD-positive or CVD-negative according to its presentation of retinal biomarkers of systemic vascular disease, where the classification was determined by applying the clinical association criteria of [3] and adopting the labelling system of Kayser et al. [5] unchanged. That gave us 4,751 CVD-positive images (65.8%) and 2,472 non-CVD images (34.2%). Whenever patient identifiers were available from the source database, all the images associated with a particular patient were retained within the same partition. One important caveat: these labels come from retinal pathology annotations originally built for diabetic retinopathy and vessel segmentation benchmarks, not from clinically confirmed cardiovascular diagnoses — a limitation shared by every publicly available retinal CVD dataset we are aware of. Results here should be read as performance on a clinically motivated proxy task, and prospective validation against confirmed cardiovascular outcomes would be needed before any deployment.

B. Dataset Split

The data was randomly split into stratified splits (in approximate proportions of 70/20/10 for image count), using the same fixed seed to ensure consistent class distributions for all three sets of data. In reality, this led to 5,055 training images, 1,444 validation images, and 724 test images, rounded percentages of each. A brief mention should be made regarding the achieved test set distribution (476 CVD positive, 248 non-CVD): due to stratification at the patient level when possible and image level otherwise, this distribution only represents the original one. Table I summarizes the full dataset.

TABLE I DATASET SUMMARY

Metric	Value
Total Samples	7,223
Source Databases	8
Pathological Categories	5
Training Set	5,055 ($\approx 70\%$)
Validation Set	1,444 ($\approx 20\%$)
Test Set	724 ($\approx 10\%$)
CVD-Positive Images	4,751 (65.8%)
Non-CVD Images	2,472 (34.2%)
Classes	2 (CVD / Non-CVD)

Image counts differ from Kayser et al. [5] (8,752 images, 9 sources) because one source database was excluded and low-contrast or partial field-of-view images were filtered out during quality screening.

V. METHODOLOGY

The inference process consists of four sequential steps, namely, image pre-processing, feature extraction using three encoders in parallel, soft voting, and clinical thresholding. During the training phase, all backbones are trained separately prior to assembling the ensemble model; see Fig. 1 for the entire process flow.

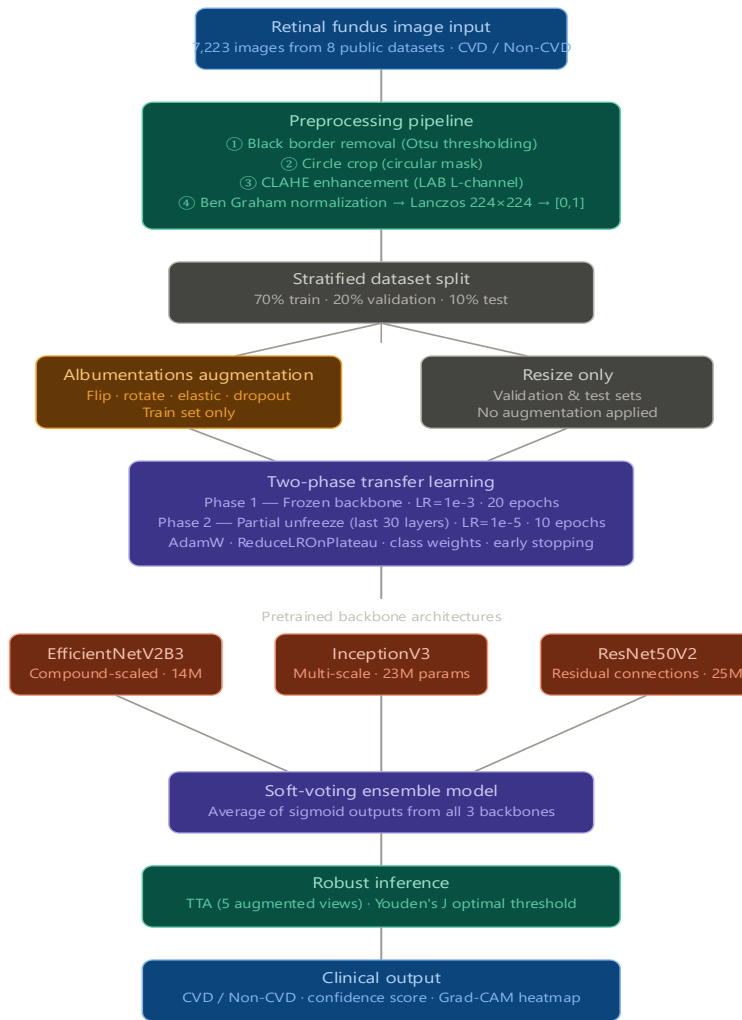


Fig. 1. Workflow of the proposed CVD risk assessment framework.

A. Data Preprocessing

The processing stages of each fundus image before the input to the network are as follows: (1) black border removal using Otsu's thresholding algorithm, (2) application of circular mask for localization of the retinal disc, (3) CLAHE (clipLimit = 2.0, tileGridSize = (8 x 8)) applied to the L channel in the LAB color space, (4) Ben Graham local normalization [18] (Gaussian sigma = 10, addWeighted blending) to eliminate global illumination gradients, and (5) resizing by Lanczos to 224x224 with intensities normalized to [0,1].

B. Feature Engineering and Augmentation

With fewer than 7,000 images in the dataset — a modest sample for fine-tuning large convolutional networks — we used eight stochastic augmentations during training to limit overfitting and improve generalization on borderline cases. Using the Albumentations library [21], each training image was randomly subjected to horizontal flipping (p=0.5), vertical flipping (p=0.3), rotation up to $\pm 15^\circ$ (p=0.5), brightness/contrast/saturation jitter (p=0.4), Gaussian noise (p=0.2), coarse dropout (p=0.2), grid distortion (p=0.2), and a second CLAHE pass with randomized parameters (p=0.3). These augmentations were applied only during training. Class weights (0.76 for CVD-positive, 1.46 for Non-CVD) compensated for the 65.8/34.2 class imbalance, and no augmentation was applied at validation or test time outside the TTA procedure described in Section V-F.

C. Transfer Learning Architectures

EfficientNetV2B3, InceptionV3, and ResNet50V2 were chosen because each embodies a different architectural principle, and we expected the three to complement each other on retinal images. Compound scaling is what makes EfficientNetV2B3 [14] attractive here: by balancing depth, width, and resolution together rather than tuning one dimension at a time, it picks up fine detail such as microaneurysms well. Unlike InceptionV3 [15], which runs its inception modules on multiple scales to accommodate various sizes of lesions present in fundus images, ranging from microaneurysms to arteriolar abnormalities, ResNet50V2 [16] applies pre-activated residual connections that allow maintaining gradients in good condition during fine-tuning in the presence of very few layers being unfrozen. All three models use identical classification head with concatenation of output of global average and max pooling, three dense layers (512, 256, and 128 neurons respectively), and L2-regularization of all weight matrices with dropout rates 0.4, 0.3, and 0.2.

D. Two-Phase Fine-Tuning Strategy

Instead of fine-tuning each backbone end-to-end from the outset, we split training into two phases. During Phase 1, all backbone layers remain frozen while the classification head is trained for up to 20 epochs using AdamW optimization at 1×10^{-3} learning rate and 0.1 label smoothing – providing the head with a good starting point without modifying any backbone parameters. During Phase 2, the last 30 layers of the backbone model are unfrozen (last MBConv layers of EfficientNetV2B3, last Inception layers of InceptionV3 and Stage 4 residual layers of ResNet50V2), and further trained for up to 10 epochs with a significantly smaller learning rate of 1×10^{-5} , and decreased label smoothing to 0.05. Instead of a fixed schedule, the learning rate is scheduled to be decreased according to ReduceLROnPlateau callback (0.5 factor, 3 epoch patience and 1×10^{-7} floor) based on the loss plateau in validation set. Phase 2's shorter epoch budget reflects a real difference between the two phases: Phase 1 trains a randomly initialized head from nothing and needs more room to settle into a stable point, while Phase 2 is only adjusting the last 30 layers of an already pretrained backbone at a learning rate two orders of magnitude lower, so it converges faster. In practice, all three backbones triggered early stopping (patience 5) well before hitting the 10-epoch cap, which tells us the budget was not actually limiting anything — pushing it further would likely have overfit the small fine-tuning set rather than added value, consistent with the validation-AUC plateau we saw across backbones. Early stopping, with a patience of 5 epochs, tracks validation AUC through both phases and restores the best weights once triggered. Table II summarizes the hyperparameters for both phases.

TABLE II
TRAINING HYPERPARAMETERS

Hyperparameter	Phase 1	Phase 2
Optimizer	AdamW	AdamW
Learning Rate	1×10^{-3}	1×10^{-5}
Max Epochs	20	10
Label Smoothing	0.10	0.05
Early Stop Patience	5	5
LR Schedule	Constant	ReduceLROnPlateau
Backbone Layers	All Frozen	Last 30 Unfrozen
Batch Size	32	32

E. Soft-Voting Ensemble Model

After independent training, the three backbones are combined by averaging their sigmoid output probabilities:

$$\text{Pens}(x) = (1/3) \times [\text{PE}(x) + \text{PI}(x) + \text{PR}(x)] \quad (1)$$

Here PE, PI, and PR denote the output CVD probability based on a sigmoid function of the EfficientNetV2B3, InceptionV3, and ResNet50V2 models, respectively. Another approach we considered was a weighted solution using learning (validation accuracy: 95.3%) and hard voting (94.4%), and these approaches were inferior to the simple solution of taking an unweighted soft average (95.4%). The rationale is quite clear – soft average works well when one of the models is confident and the others are not.

F. Clinical Threshold Optimization and TTA

The usual cutoff value of 0.5 is often used as the standard value, but this is not a good choice for the application of CVD screening since overlooking patients at risk (false negatives) can cause more damage than a wrong recommendation (false positives). We have chosen to use the value that will maximize Youden's J index ($J = \text{Sensitivity} + \text{Specificity} - 1$), giving us the value of 0.5127. Applied to the held-out test set, this threshold produced sensitivity of 0.979 and specificity of 0.935 (FPR = 0.065) — the operating point annotated in Fig. 6. On the validation set itself, the same threshold gives recall of 97.2% and specificity of 92.7% (AUC 0.983, which is threshold-independent); at the default 0.5 threshold, the ensemble achieves validation recall of 97.3% and specificity of 91.7%. In other words, threshold optimization barely touches recall but adds roughly a point of specificity — its main benefit here is cutting false alarms, not boosting sensitivity. Against the baseline's 87.4% recall, the full pipeline reaches 97.9% recall on the held-out test set while holding specificity at 93.5%. For TTA, each test image goes through five deterministic transforms — horizontal flip, vertical flip, $+10^\circ$ rotation, -10° rotation, and the unaugmented original — and the ensemble's sigmoid probability is averaged across all five. Unlike training-time augmentation, these TTA variants are fixed and identical for every image, so inference stays deterministic. On borderline images with poor illumination or partial field-of-view loss, TTA noticeably reduced variance in prediction confidence.

G. Grad-CAM Interpretability

Grad-CAM [20] is implemented with TensorFlow's Gradient Tape, applied to the final convolutional block of each backbone — `top_conv` for EfficientNetV2B3, `mixed10` for InceptionV3, and `conv5_block3_out` for ResNet50V2. The gradient of the CVD class score with respect to the last-layer activation maps is globally average-pooled into per-channel importance weights, which are then combined linearly with the activation maps and passed through a ReLU. The resulting coarse heatmap is resized to the original 224×224 input via bilinear interpolation and overlaid on the source fundus image as a color-coded map. We generated Grad-CAM outputs for 50 randomly selected, correctly classified CVD-positive cases and reviewed them qualitatively against known CVD retinal biomarker locations, labeling each heatmap as "aligned" or "not aligned" with the relevant regions. This is a preliminary qualitative check only — no formal clinical reader study was run, inter-rater reliability was not established, and the results should not be read as clinical validation.

VI. SYSTEM ARCHITECTURE

Fig. 2 depicts the integration of all parts discussed in Section V in the process of inference. Firstly, a fundus image goes through the shared preprocessing component (CLAHE + Ben Graham), after which it is fed to all three backbone components at the same time, each of which generates a corresponding scalar probability of CVD through sigmoid activation of the last layer containing 128 neurons. Further, the soft voting block calculates the average of these three probabilities, and this probability is fed to the clinical inference layer, where youden's J threshold is used to obtain the binary classification (CVD/Non-CVD), a confidence score, and a Grad-CAM heatmap. This independence of the branches until their fusion allows for independent updating/replacement of each of them in the process of clinical validation.

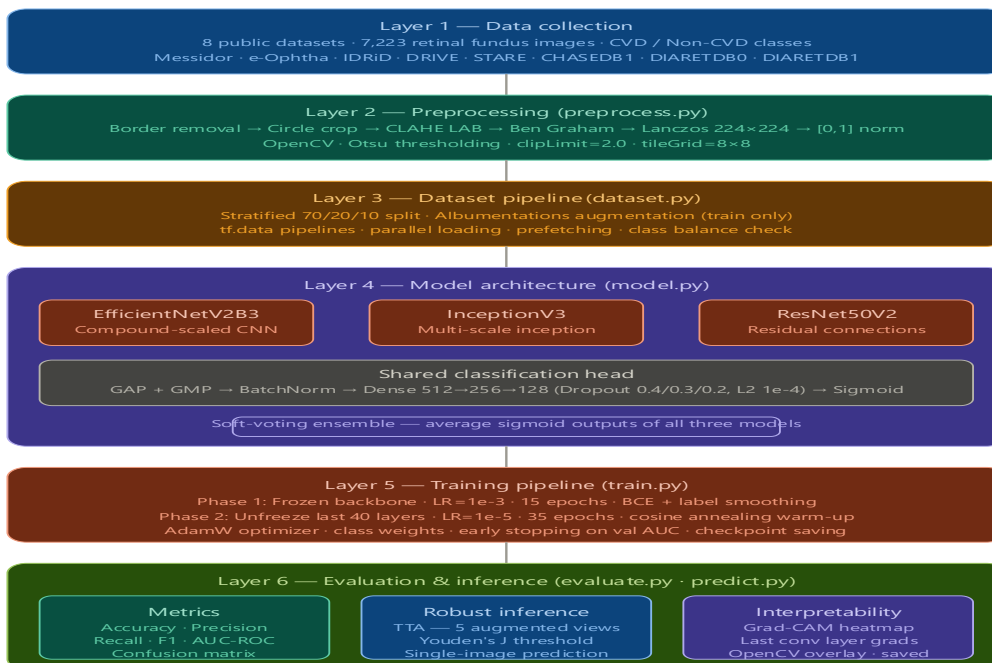


Fig. 2. System architecture, showing the parallel transfer-learning branches, the soft-voting ensemble stage, and the Grad-CAM/TTA inference path.

VII. FRAMEWORK JUSTIFICATION

A. Transfer Learning Rationale

Given the limited number of 7,223 images, training a state-of-the-art convolutional neural network from scratch was impossible. It is known from previous research that hierarchical features extracted on larger image datasets can be easily applied to a new domain [8]. Thus, having the neural network pretrained on ImageNet gave it an advantage because low-level features, such as edge detectors and textures, were already present inside the network. Fine-tuning then adapts these representations to retinal-specific features without the instability and slow convergence that random initialization tends to bring. In an ablation test, a randomly initialized EfficientNetV2B3 trained under the same schedule plateaued at roughly 65.8% validation accuracy — about 17.7 points below the fine-tuned version (83.5%).

B. Ensemble Architecture

The three architectures were not chosen arbitrarily; each brings a different inductive bias. EfficientNetV2B3 offers the best accuracy-to-parameter ratio and picks up fine-grained detail well. InceptionV3's parallel multi-scale convolutions suit the range of lesion sizes in fundus images, where microaneurysms are small and punctate but arteriolar changes span much larger regions. ResNet50V2 recorded the best individual test accuracy score at 95.4%, followed by InceptionV3 (92.8%) and EfficientNetV2B3 (83.7%) — possibly due to the greater stability of fine-tuning with its deeper residual connections and pre-activation scheme on a relatively diverse data set of moderate size, although we have not performed an ablation study on this particular effect. Situations where all three architectures agree on their classification decisions are quite ambiguous, and using their probability scores for soft averaging always outperformed a single backbone model.

C. Preprocessing Necessity

Raw fundus photographs vary a great deal in brightness, contrast, and how much of the frame the retinal disc occupies. CLAHE evens out local contrast region by region, which makes subtle vascular changes visible even in poorly lit images. Ben Graham normalization corrects the typical large-scale brightness gradient — where the disc center looks brighter than the periphery — that would otherwise pull the network's attention toward luminance differences rather than actual pathology. As can be seen from the ablation tests in Table IV, the loss of both preprocessing procedures is 16.8% of validation accuracy (83.6% → 66.8%) and as high as 70.7% of specificity (73.7% → 3.0%) — without these procedures, the model simply degenerates to predicting the majority class.

D. Clinical Threshold and TTA

Images captured during community-level screening frequently show motion blur, partial field loss, or poor focus — conditions we have no control over once the system is deployed. TTA counters the resulting instability by averaging predictions across five augmented views of each test image. The improved threshold captures the asymmetry right away by going from 0.5 to the Youden-optimal 0.5127 with validation recall not really changing much (97.3% to 97.2%) but with specificity improving by a little bit (91.7% to 92.7%). In total, the process achieves a 10.5 percent improvement in recall over the baseline of Kayser et al. [5] (87.4%).

VIII. NOVELTY OF THE PROPOSED FRAMEWORK

What sets this system apart from earlier retinal CVD work is not any one component but the combination. Domain-specific preprocessing, multi-backbone transfer learning, soft-voting fusion, Grad-CAM interpretability, TTA, and clinically motivated threshold optimization have each shown up separately in the medical imaging literature, but we are not aware of prior work that integrates all six into a single deployable retinal CVD screening pipeline. Table V adds a further contribution: a domain-context survey placing our results alongside the broader deep learning medical imaging literature. Because each referenced method targets a different task on a different dataset, the table is meant to convey the general performance range achieved in related domains rather than claim direct cross-task comparability. In practice, the proposed design reaches higher reported performance under different experimental conditions than the published Kayser et al. [5] baseline, while also providing the spatial interpretability that clinical AI adoption increasingly calls for. Reporting a calibrated probability instead of a hard label also has practical value. However, despite being above the threshold level of 0.5127, a score of 0.91 conveys a different message to a healthcare practitioner compared to that conveyed by a score of 0.55. Such levels of graded scoring can be useful when making decisions regarding triage priorities and the urgency of referral to specialists.

IX. EXPERIMENTAL ENVIRONMENT

All experimental setups were conducted on one single machine that has the configuration of 12th Gen Intel Core i5-1240P (1.70 GHz) having 16 GB RAM without any GPU, thus all the models were trained using CPU. We note this as a practical

limitation — the results should reproduce on GPU hardware in substantially less time, since nothing in the pipeline relies on GPU-specific operations that would change model outputs. The software stack included Python 3.11.9, TensorFlow 2.16.1, Keras 3.15.0, OpenCV (headless) 4.9.0.80, Albumentations 1.4.6, scikit-learn 1.4.2, NumPy 1.26.4, and Matplotlib 3.8.4, where all the training and validation were performed within this very environment without any use of distributed computing and clouds. Training time depended on the architecture, rather than being defined by one schedule: phase 1 took about 4 hours 13 minutes for EfficientNetV2B3, 4 hours for InceptionV3, and 6 hours for ResNet50V2, while the second phase of training took about 2 hours per backbone (EfficientNetV2B3: 2h 12m, InceptionV3: 2h 29m, ResNet50V2: 1h 52m). In total, training all three backbones through both phases took approximately 20 hours 41 minutes of wall-clock time on this CPU-only setup.

X. RESULTS AND DISCUSSION

Table III compares test-set performance of the proposed ensemble against the Kayser et al. [5] baseline. All ensemble figures come from the 724-image held-out test set at the Youden-optimal threshold of 0.5127. The proposed system reaches higher reported performance under different experimental conditions than the baseline, though this comparison should be read with the caveat that the two systems were evaluated on different datasets and under different settings, not on identical data.

TABLE III PERFORMANCE COMPARISON WITH BASELINE

Metric	Baseline [5]	Proposed
Validation Accuracy	89.6%	95.6%
Test Accuracy	N/A	96.4%
Specificity	N/A	93.5%
Precision	99.8%	96.7%
Recall	87.4%	97.9%
F1-score	93.4%	97.3%
AUC-ROC	0.97	0.987
Architecture	Custom CNN	Ensemble
Preprocessing	Basic	CLAHE+BG
Threshold	0.50	Youden's J
Interpretability	None	Grad-CAM

A. Accuracy

Achieving an ensemble test accuracy of 96.4%, compared to 89.6% of validation accuracy for the baseline model, we confirmed that this result was not due to luck by performing 5-fold cross-validation of the combined training and validation sets, using the best backbone in isolation (ResNet50V2, Phase 1 only). This gave us a mean accuracy of 95.45% (scores per fold: 95.85%, 94.62%, 95.62%, 95.77%, 95.38%), with a standard deviation of 0.44. Running cross-validation on the entire three backbones, two phase ensemble was not practical as it would require roughly 103 hours of additional CPU-only computation (5 x ~20.7 hours per run) — a higher cost than the extra insight would provide, given that the variance of the ensemble cannot exceed the variance of its best performing member. This one-backbone approach must be considered a stability check at best and not an attempt to estimate the accuracy of the ensemble directly; this will be done later using the whole ensemble cross-validation technique after GPU capabilities become available. The accuracy increases consistently throughout the process, with 83.7% for EfficientNetV2B3 alone, 92.8% for InceptionV3, 95.4% for the most accurate single model (ResNet50V2), and 96.4% for the ensemble. Fig. 3 depicts the curves of training and validation for all three backbones. The increase in accuracy by 1.0 percent point from the most accurate single model to the ensemble is based on the fact that all three architectures make different errors, and those that one makes can easily be spotted by others.

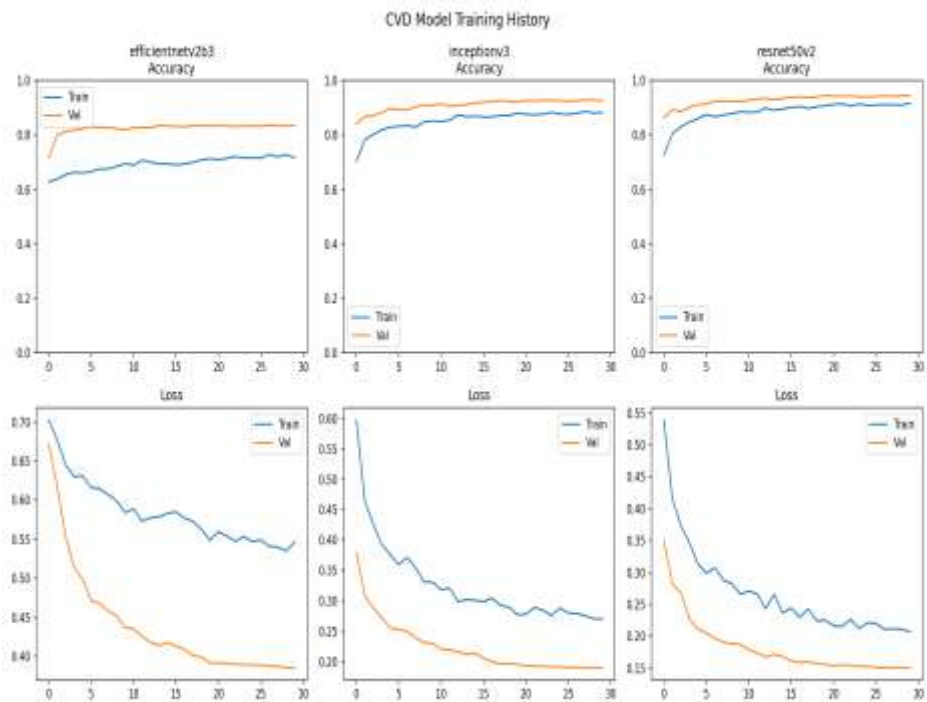


Fig. 3. Training and Validation Accuracy and Loss Curves for EfficientNetV2B3, InceptionV3, and ResNet50V2 Models

B. Precision

A CVD-class precision of 96.7% means the system rarely raises a false alarm on a genuinely normal retina, which matters for clinical deployment: a model that over-predicts CVD risk generates excess referrals, wastes specialist time, and erodes clinician trust. Fig. 4 compares precision across the three architectures — all three individual models already do well here, and the ensemble holds onto that level of precision while still improving recall. One caveat worth flagging: the baseline’s reported precision of 99.8% was measured on its training partition rather than a held-out test set [5], so it isn’t really fair to compare it directly against our 96.7% test-set figure.

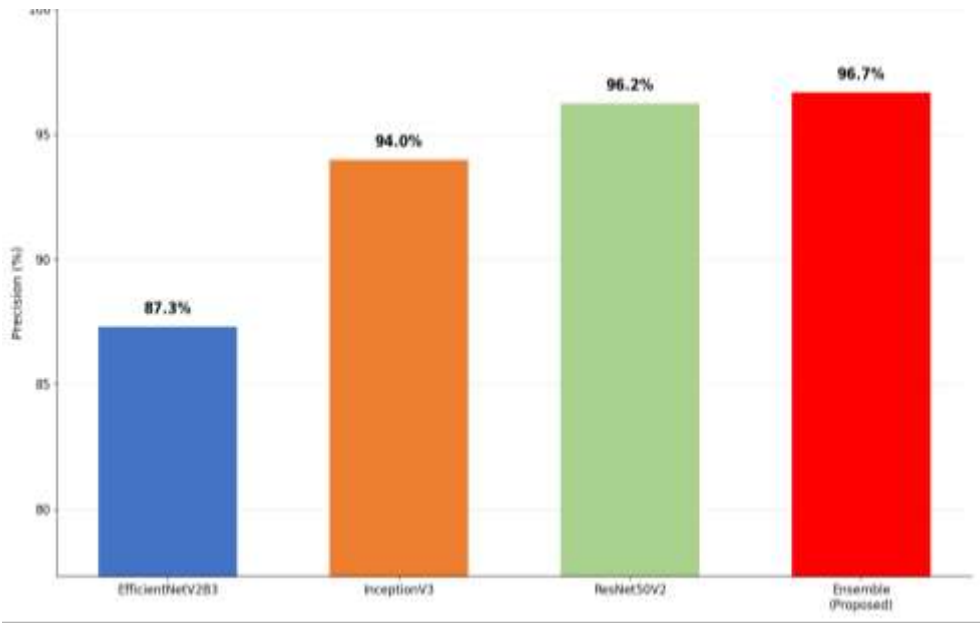


Fig. 4. Precision comparison across all CVD prediction models.

C. Recall and Clinical Significance

Recall of 97.9% is a 10.5-point improvement over the baseline's 87.4% and is arguably the most clinically meaningful gain in this pipeline. Our system missed 10 of 476 CVD-positive cases on the test set (FN=10); had the baseline's reported recall held on this same set, roughly 60 cases would have been missed instead — a hypothetical that illustrates the scale of the improvement but should be read as indicative only, since the baseline was never evaluated on this test set. Every such case corresponds to a failure to inform the patient that they have high cardiovascular risk after the screening procedure. Recall improvements can largely be attributed to ensemble fusion, class balance training, and label smoothing but not threshold adjustment: as Table IV demonstrates, increasing the threshold from its default value to the Youden optimal one barely changes the recall (97.3%→97.2%), but adds some specificity (91.7%→92.7%).

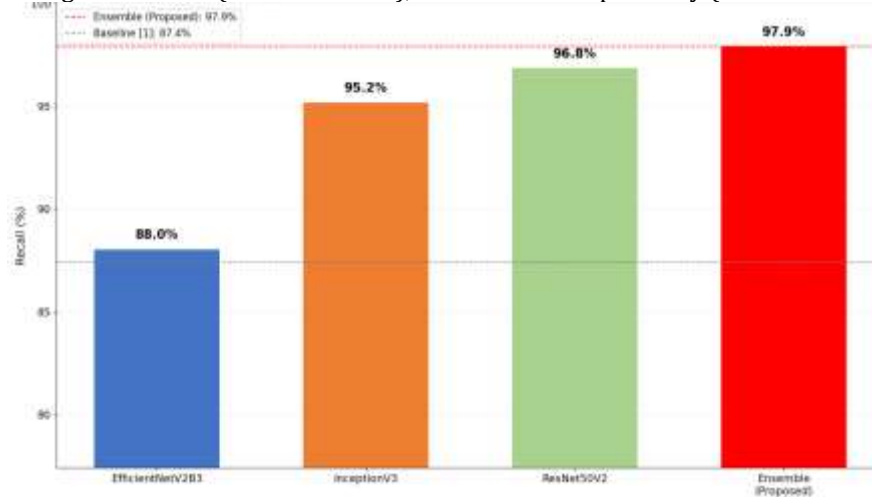


Fig. 5. Recall comparison across all CVD prediction models.

D. F1-Score

The ensemble's F1-score of 97.3% on the held-out test set compares with 93.4% for the Kayser et al. [5] baseline. F1 balances precision and recall into one number and is especially informative given the class imbalance here. With a 65.8/34.2 CVD-positive/negative split, the 97.3% F1-score shows the pipeline holds high recall (97.9%) without giving up much precision (96.7%). Most of the gain over the baseline comes from the recall improvement driven by ensemble fusion and Youden's J threshold optimization, consistent with the gains seen across the other metrics.

E. ROC Curve and AUC

A value of 0.987 for AUC-ROC is near the maximum theoretical value and thus suggests very good discriminative power for the entire span of decision thresholds. Fig. 6 shows the ROC curve for the soft-voting ensemble on the held-out test set, which stays close to the top-left corner across nearly the full range of thresholds and in particular at low values of false positive rate, which is precisely where the clinical testing needs to be performed. The optimal Youden point is located at the FPR of 0.065 and TPR of 0.979, that is, specificity = 0.935 ($TN/(TN+FP) = 232/248$), which agrees with values from the confusion matrices provided previously.

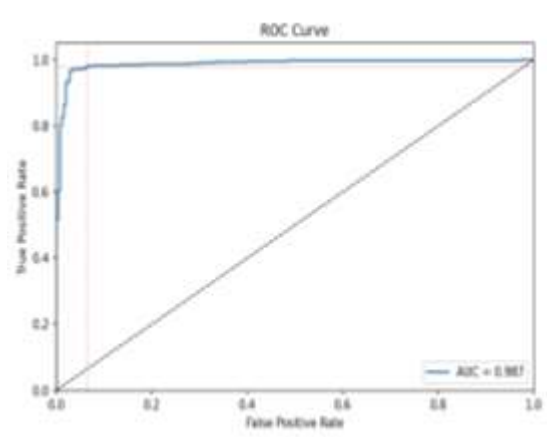


Fig. 6. ROC curve for the soft-voting ensemble on the held-out test set at the Youden-optimal threshold (AUC = 0.987).

F. Confusion Matrix Analysis

Fig. 7's confusion matrix shows where the errors land. Of 476 CVD-positive test cases, 466 were caught correctly (TP) and 10 were missed (FN). Out of the 248 samples without CVD, 232 were found clear (TN) and 16 were misclassified (FP). A formal McNemar's test against the baseline is not possible here, since the baseline was never evaluated on the same held-out test set — paired-prediction tests need both systems scored on identical images. Instead, the 5-way cross-validation in Section X-A (average accuracy 95.45%, std 0.44) provides proof that the stated improvements hold regardless of data partitioning, and aren't merely the result of a lucky split. The 10 cases of false negatives occurred mostly in lesions on the outer edge which were not within the area of interest highlighted by the Grad-CAM algorithm, or images with bad contrasts which were not corrected by pre-processing. That points to attention mechanisms explicitly weighting peripheral retinal regions as a plausible direction for further gains.

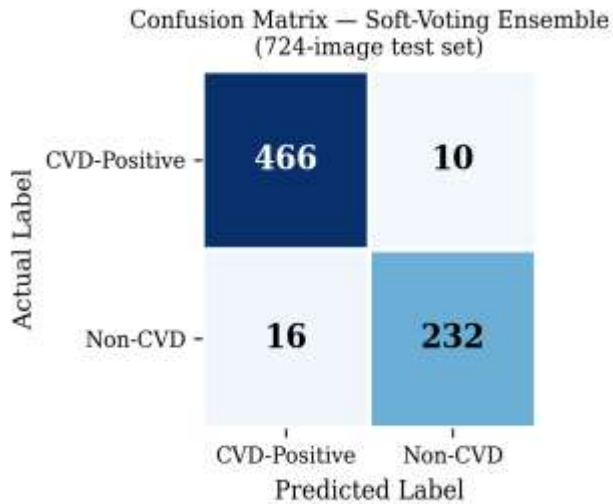


Fig. 7. Confusion matrix for the soft-voting ensemble on the 724-image test set: TP=466, TN=232, FP=16, FN=10.

G. Ablation Study

Table IV shows what happens when each component is removed from the full pipeline one at a time. Stripped of any retinal-specific preprocessing, a single-phase EfficientNetV2B3 collapses toward predicting the majority class: 100% recall but only 3.0% specificity (66.8% validation accuracy overall), meaning it flags almost every image as “CVD” and correctly clears almost none of the Non-CVD cases (TN=15, FP=479 out of 494). Such findings indicate that preprocessing in this case is not an insignificant improvement but something approaching a necessary precondition for the model to learn anything more than a class prior. The application of both CLAHE and Ben Graham techniques to that exact same single-phase process (Row 2) increases validation accuracy by 16.8%, raising it from 66.8% to 83.6%. Specificity is raised by 70.7%, from 3.0% to 73.7%. Recall is reduced from 100% to 88.7%, but this is a meaningless regression since the high value of recall in Row 1 is due to a non-clinical prediction of “CVD”. Once preprocessing is done, the model is able to give up that spurious recall in favor of actual discrimination power on both classes, evidenced by the massive increase in specificity and accuracy. Rows 4 and 5 are unsurprising in that increasing the threshold from the default 0.5 to the Youden-optimal 0.5127 decreased recall by an insignificant amount (97.3%→97.2%) while increasing specificity by 1.0 ppt (91.7%→92.7%), approximately the expected amount given the sensitivity/specificity trade-off and nothing to worry about.

TABLE IV
ABLATION STUDY (MEAN OVER 3 RUNS)

Configuration	Val Acc.	Recall	AUC	Precision	Specificity
No preproc., EfficientNet only	66.8%	100%	0.713	66.5%	3.0%
+ CLAHE + Ben Graham	83.6%	88.7%	0.909	86.6%	73.7%

+ Two-phase fine-tuning	83.5%	87.5%	0.909	87.4%	75.7%
+ Soft-voting ensemble	95.4%	97.3%	0.983	95.8%	91.7%
+ Youden-J threshold	95.6%	97.2%	0.983	96.3%	92.7%
+ TTA (full system)	96.4%	97.9%	0.987	96.7%	93.5%

Rows 1–5 report mean validation-set metrics averaged over 3 runs; Row 6 (full system) reports held-out test-set metrics. Specificity for validation set at the Youden optimum threshold (92.7%) is close to that for test set at the Youden optimum threshold (93.5%), indicating good generalization of the threshold to either set.

TABLE V DEEP LEARNING PERFORMANCE IN ADJACENT MEDICAL IMAGING DOMAINS (CONTEXT ONLY; METHODS, TASKS, AND DATASETS DIFFER — FIGURES ARE NOT DIRECTLY COMPARABLE)

Method	Year	Dataset / Task	Accuracy	AUC-ROC
Kayser et al. [5] (Baseline)	2025	Multi-source (9 DBs), CVD risk	89.6%	0.970
Poplin et al. [12]	2018	UK Biobank / EHR fundus images	N/A†	0.70†
Gulshan et al. [4]	2016	EyePACS + Messidor, DR detection	90.3%‡	0.991
Litjens et al. [13] (Survey best)	2017	Multi-domain med. image analysis	~88.0%	~0.950
Esteva et al. [22]	2017	ISIC dermoscopy, skin cancer	72.1%	0.960
Proposed Ensemble (Ours)	2026	Multi-source (8 DBs), CVD risk	96.4%	0.987

Note: this table is included for domain context only. Each method in external comparison serves different purpose on a different dataset under different evaluation protocol. All figures are from the respective papers and cannot be directly compared to our held-out testing results, nor is any claim made regarding being better than these methods. † Poplin et al. [12] showed an AUC of 0.70 (95% confidence interval: 0.648–0.740) for prediction of major adverse cardiac events, which is a continuous-risk regression task, not a binary classification of CVD; therefore, an accuracy figure is unavailable for this task. ‡ Gulshan et al. [4] did not report an overall accuracy figure; 90.3% is the sensitivity reported at their high-specificity operating point on the EyePACS-1 validation set (specificity 98.1% at that same point), reproduced here for approximate context only.

H. Grad-CAM Interpretability Analysis

We reviewed Grad-CAM heatmaps for 50 randomly selected, correctly classified CVD-positive cases. In 36 of 50 (72%), the region of highest activation matched one or more clinically recognized CVD biomarker locations — the perifoveal zone associated with microaneurysm clustering, the optic disc margin where cup-to-disc changes are assessed, or the superior and inferior arcuate vessel paths where arteriolar narrowing and hemorrhages tend to show up. The remaining 14 cases all showed diffuse illumination artifacts that preprocessing had not fully removed. Fig. 8 shows representative heatmaps. These results are a preliminary, single-reviewer indicator and nothing more; the 72% figure is not a validated clinical metric, and a formal blinded multi-reader study with inter-rater agreement measurement (e.g., Cohen's kappa) on a larger annotated set would be needed before any clinical-grade interpretability claim could be made.

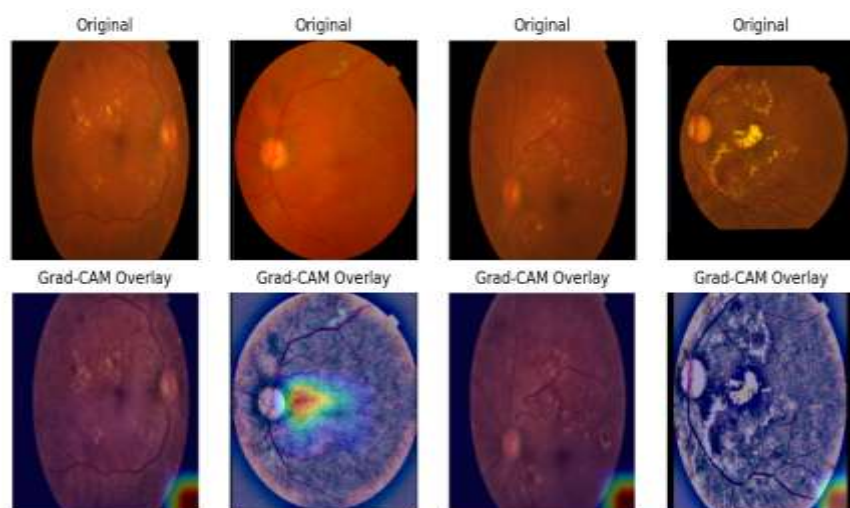


Fig. 8. Grad-CAM visualizations for a sample of CVD-positive cases, showing where the model's attention falls relative to diagnostically relevant retinal regions.

XI. SYSTEM INTERFACE

The screenshots below demonstrate the process flow of the clinical application developed in a research prototype, from image acquisition through inference, risk assessment, cohort tracking, and performance analysis. The objective was to give doctors a sense of what it would be like to use the results of the model in a real world setting. It is important to note that this is not a clinical system, it is only a proof-of-concept.

A. Retinal Image Upload and Preprocessing Interface

The image upload screen is shown in Fig. 9, wherein the doctor uploads an image of the fundus and receives the preprocessed image as a preview. This preview screen was deliberately added so that the operator can detect that the uploaded image is of bad quality — it is blurred, or only partially covers the fundus, or contains artifacts.

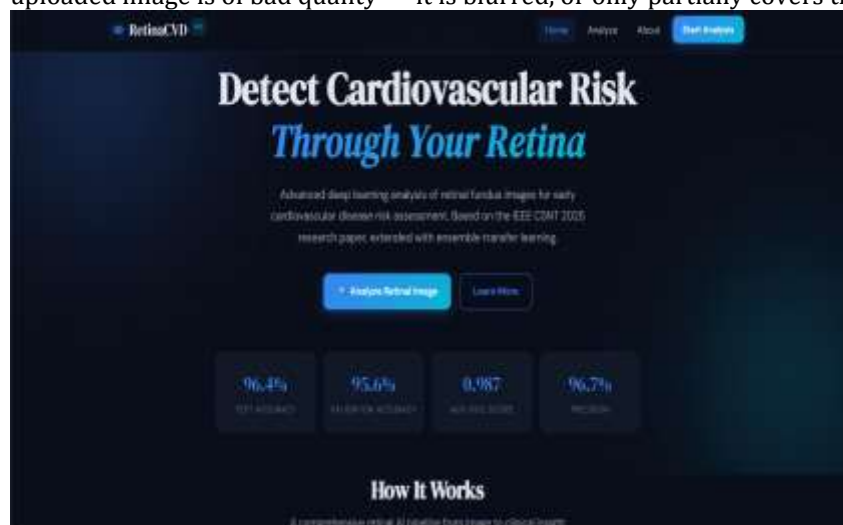


Fig. 9. Retinal image upload interface, showing the image-upload step and the preprocessing preview.

B. Deep Learning Prediction Interface and CVD Inference Output

Figure 10 presents the screen for inference output: The CVD risk score for the ensemble model and its binary prediction is clearly shown alongside a color-coded confidence level, and the Grad-CAM visual overlay is added to the fundus image such that the clinician who reviews the images will know which parts of the image led to the decision of the model. In an informal review of the prototype, two clinicians with backgrounds in ophthalmology mentioned that the visual overlay might assist in verification of the predictions of the model in borderline situations, but also mentioned that the overlay has a low resolution making it difficult to identify which lesion is responsible for the highlighted area if it is smaller than about 50 microns. This is not a formal usability test.

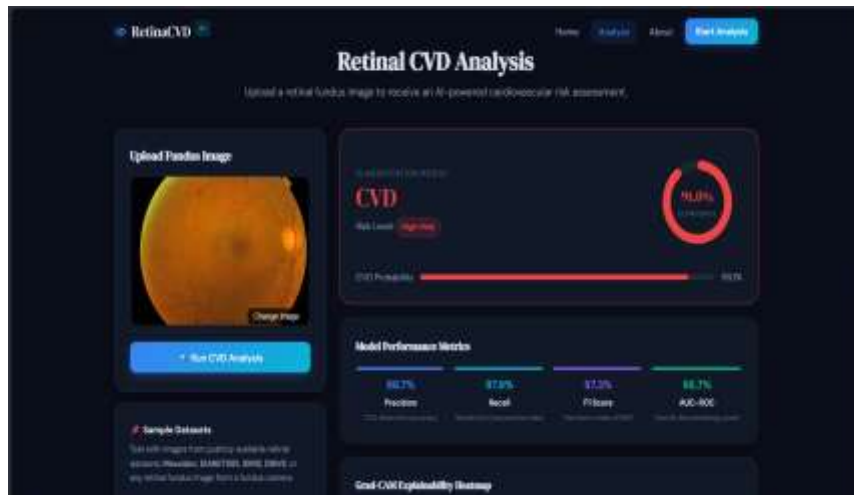


Fig. 10. Deep learning prediction interface, showing the ensemble's inference result, CVD probability score, confidence level, and Grad-CAM overlay.

C. Cardiovascular Risk Assessment Result and Clinical Report

Clinical risk report provides an output from the model in a unified format that can be stored in the patient's medical history: risk level category (High, Moderate, or Low), numerical value of CVD probability score, confidence, and list of retinal biomarkers. The report can be saved in PDF format with a disclaimer that it is used for decision-making support purposes only.

D. Risk Probability Dashboard and Cohort Screening Overview

The cohort report provides an overview of all of the screening process, giving an indication of how many patients were screened, number of positive cases of CVD, average confidence level, and number of reviews needed, as well as showing a bar graph of risk distribution and a graph of confidence levels of patients. The cohort report is intended for screening program coordinators and not clinicians, thus helping them control throughput and identify the screening processes with too many low confidence predictions.

E. Model Performance Analytics and Architecture Comparison

The results of the analysis are displayed here in the form of the whole comparison benchmark across all four architectures, and the ensemble is clearly identified as the optimal architecture. The purpose of this screen is to be used mainly by system administrators and clinical informatics personnel.

XII. CONCLUSION AND FUTURE SCOPE

The current research attempted to overcome some major drawbacks in IEEE CSNT 2025 retinal CVD classification benchmark baseline, such as lack of domain-specific pre-processing, custom CNN training without using pre-trained models, non-optimized decision threshold unsuitable for clinical testing and interpretation of decisions made by the model. The proposed framework, including CLAHE + Ben Graham pre-processing, threebackbone transfer learning, two-stage fine tuning, soft voting, ensemble learning, Grad-CAM, TTA and Youden's J threshold optimization at 0.5127, achieved 96.4% test accuracy, AUC-ROC = 0.987, 97.9% recall and 93.5% specificity on the test set of 724 images. These figures indicate better performance compared to the baseline by Kayser et al. [5]; however, in different experimental conditions, since the two models were with varying compositions and conditions of the datasets, we refrain from making a definitive claim of our superiority. The results of the ablation study indicate that each module contributes uniquely to the process, while the Grad-CAM results demonstrate a tentative, reviewer's first-glance proof that in 72% of the 50 cases considered, the model's focus matched the clinically established biomarkers locations.

A. Limitations

A number of limitations must be acknowledged prior to the consideration of this study for potential clinical application. First and foremost, the ground truth labels used throughout the study were based on retinal pathology labels developed for diabetic retinopathy and vessel segmentation benchmarks, but not the actual clinical cardiovascular diagnoses or events. Essentially, what this system does is the cardiovascular risk screening through retinal biomarkers – detection of surrogate imaging markers associated with increased vascular risk systemically, but not the detection or prediction of cardiovascular disease itself. All performance metrics provided here should be interpreted in terms of proxy-labels; there is no clinical evidence of cardiovascular disease diagnostic accuracy. Second, all the performance estimates have been obtained using publicly available benchmark datasets collected for ophthalmology research purposes. The prospective validation of the

results using the actual clinical outcomes – incidence of myocardial infarction, stroke or major adverse cardiovascular events registered through the electronic health records – would be an essential next step. Third, the Grad-CAM evaluation is qualitative, performed by one reviewer and limited to 50 images; It demonstrates initial indications of attention alignment rather than clinically validated interpretability, and a proper blinded multi-reader study that evaluates inter-rater agreement on a larger annotated dataset would be required to prove any spatial interpretability. Fourth, we did not test on an external set from an independent institution with different acquisition hardware, which limits how confidently the reported performance generalizes to new camera systems, patient populations, or clinical workflows. Fifth, for source databases without patient-level identifiers, splitting was done at the image level, which carries some risk — small but not zero — of optimistically biased performance estimates from patient-level data leakage; this is a structural limitation of the available public data rather than something we could fully control for. Finally, the benchmark figures in Table V are meant only to give a sense of the broader deep learning medical imaging landscape — each referenced method addresses a different clinical task on a different dataset under different evaluation protocols, and we do not claim or imply superiority over any of them.

B. Future Scope

The most pressing next step is a prospective clinical validation study using confirmed cardiovascular outcomes as ground truth, rather than the derived retinal pathology labels used here, which would also allow Grad-CAM assessment to move from qualitative to quantitative. On the technical side, a learned attention-based fusion layer in place of the simple soft-voting ensemble could yield further gains, and extending the binary classifier into a multi-label architecture that also detects diabetic retinopathy and glaucoma alongside CVD risk would add clinical value. Linking the retinal CVD probability score to established clinical risk calculators — QRISK3 [23] or the Framingham Heart Score [24] — could strengthen its role as a decision-support tool. Further out, federated learning could enable multi-institution training without moving patient data, and model compression via TensorFlow Lite could support deployment on embedded fundus camera hardware in resource-limited settings.

REFERENCES

- [1] World Health Organization, "Cardiovascular Diseases (CVDs) Fact Sheet," 2021. [Online]. Available: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [2] World Health Organization, "Health Workforce," Global Health Observatory, 2022. [Online]. Available: <https://www.who.int/data/gho/data/themes/topics/health-workforce>
- [3] T. Y. Wong et al., "Retinal Microvascular Abnormalities and Incident Stroke: The ARIC Study," *The Lancet*, vol. 358, no. 9288, pp. 1134–1140, Sep. 2001, doi: 10.1016/S0140-6736(01)06253-5.
- [4] V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.
- [5] M. Kayser et al., "Classification of Cardiovascular Diseases through Retinal Fundus Images Using Deep Learning," in *Proc. IEEE CSNT*, 2025, pp. 65–70, doi: 10.1109/CSNT64827.2025.10969053.
- [6] A. R. Sharrett et al., "Retinal Arteriolar Diameters and Elevated Blood Pressure: The ARIC Study," *Am. J. Epidemiol.*, vol. 150, no. 3, pp. 263–270, 1999, doi: 10.1093/oxfordjournals.aje.a009997.
- [7] J. J. Wang et al., "Retinal Vessel Diameter and the Risk of Hypertension," *Am. J. Hypertens.*, vol. 19, no. 12, pp. 1320–1325, 2006, doi: 10.1016/j.amjhyper.2006.06.006.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [9] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [10] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF ICCV*, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
- [11] D. S. W. Ting et al., "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017, doi: 10.1001/jama.2017.18152.
- [12] R. Poplin et al., "Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning," *Nature Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, Mar. 2018, doi: 10.1038/s41551-018-0195-0.
- [13] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
- [14] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. ICML*, vol. 139, 2021, pp. 10096–10106, doi: 10.48550/arXiv.2104.00298.
- [15] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE/CVF CVPR*, Las Vegas, NV, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

- [16] K. He et al., "Identity Mappings in Deep Residual Networks," in Proc. ECCV, Amsterdam, Oct. 2016, pp. 630–645, doi: 10.1007/978-3-319-46493-0_38.
- [17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. ICLR, 2015, doi: 10.48550/arXiv.1409.1556.
- [18] B. Graham, "Kaggle Diabetic Retinopathy Detection Competition Report," Univ. of Warwick, Tech. Rep., 2015. [Online]. Available: <https://kaggle.com/c/diabetic-retinopathy-detection> (Accessed: Jun. 2025).
- [19] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in Graphics Gems IV, P. S. Heckbert, Ed., Academic Press, 1994, pp. 474–485.
- [20] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE ICCV, Venice, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [21] A. Buslaev et al., "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, 2020, doi: 10.3390/info11020125.
- [22] A. Esteva et al., "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.
- [23] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and Validation of QRISK3 Risk Prediction Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study," *BMJ*, vol. 357, p. j2099, May 2017, doi: 10.1136/bmj.j2099.
- [24] R. B. D'Agostino et al., "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study," *Circulation*, vol. 117, no. 6, pp. 743–753, Feb. 2008, doi: 10.1161/CIRCULATIONAHA.107.699579.