Multimodal Deep Learning for Cardiovascular and Diabetes Risk Prediction Using Retinal Fundus Images and Clinical Data

Shahina Begum[1], Rubina Begum[2]
[1]*Faculty & Research Scholar, Department of Computer Science and Engineering, SECAB Institute of Engineering and Technology, Vijayapura, Karnataka, India*
[2]*Research Scholar, Department of Computer Science and Engineering, SECAB Institute of Engineering and Technology, Vijayapura, Karnataka, India*

***Abstract*—cardiovascular disease and diabetes are major health conditions for which early risk assessment can support timely clinical evaluation. Retinal fundus images provide non-invasive information about retinal microvascular characteristics, while structured clinical variables provide complementary physiological information. This work presents a multimodal deep learning framework that combines retinal fundus images with clinical data for simultaneous cardiovascular and diabetes risk prediction. Retinal images are resized to 224×224 pixels and processed using an ImageNet-pretrained EfficientNet-B3 backbone. A two-layer Transformer Encoder with eight attention heads is applied to the extracted feature sequence to model global contextual relationships. Clinical variables comprising blood pressure, body mass index, glucose, cholesterol, age, and diabetes status are processed through a fully connected network. The resulting image and clinical representations are combined using an attention-based fusion mechanism, followed by two task-specific output heads. Grad-CAM is incorporated to provide visual explanations of image-based predictions. The reported evaluation on 4,128 samples gives an overall accuracy of 98.18%, with 98.06% for cardiovascular-risk prediction and 98.30% for diabetes-risk prediction. However, the clinical variables in the present implementation are generated from retinal-disease severity information using medically informed distributions; therefore, the reported results should be interpreted as an experimental proof of concept and require validation with real patient-level clinical measurements before clinical use.**

***Index Terms*—multimodal learning, deep learning, retinal fundus images, EfficientNet-B3, Transformer Encoder, cardiovascular risk, diabetes risk, Grad-CAM, explainable AI**

I. INTRODUCTION

Cardiovascular diseases (CVDs) and diabetes mellitus are major causes of morbidity and mortality worldwide. Early risk assessment is important because timely preventive measures, monitoring, and clinical evaluation can reduce the likelihood of severe complications. Conventional assessment may require laboratory investigations, imaging procedures, and expert interpretation, which can be difficult to scale in resource-constrained settings.

Retinal fundus imaging provides a non-invasive view of the retinal microvasculature. Retinal vascular and structural characteristics have been investigated as potential indicators of systemic health. Public retinal datasets, including the Indian Diabetic Retinopathy Image Dataset (IDRiD), have also supported the development of automated retinal-image analysis systems.

Deep learning has substantially improved medical image analysis. Convolutional Neural Networks (CNNs) can learn hierarchical visual representations directly from images, while EfficientNet provides an effective balance between representation capacity and computational efficiency. Transformer-based attention mechanisms can complement CNNs by modelling long-range relationships among spatial feature representations.

Image information alone, however, does not fully represent an individual's health profile. Clinical

attributes such as blood pressure, body mass index, glucose, cholesterol, age, and diabetes status can provide complementary information. This motivates multimodal learning, in which heterogeneous image and structured data are processed and fused within a common prediction framework.

Interpretability is also important in medical AI. Grad-CAM can highlight image regions that contribute to a model's output and therefore provides a visual aid for examining whether the learned representation is associated with relevant retinal regions.

The objective of this study is to develop an experimental multimodal framework for simultaneous cardiovascular and diabetes risk prediction using retinal fundus images and structured clinical features. The main contributions are: (i) a CNN-Transformer image representation, (ii) a dedicated clinical-data encoder, (iii) attention-based multimodal fusion with two prediction heads, and (iv) Grad-CAM-based visual explanation. Because the clinical variables used in the current implementation are generated rather than collected as matched patient measurements, the work is presented as a proof-of-concept rather than a clinically validated diagnostic system.

II. RELATED WORK

Earlier retinal-analysis systems relied on handcrafted descriptors such as texture, colour, vessel morphology, thresholding, and morphological operations, followed by classifiers such as SVM, k-NN, and Random Forest. These methods can be sensitive to image quality and may not capture high-level representations.

CNN-based systems subsequently enabled end-to-end feature learning. Gulshan et al. demonstrated deep learning for diabetic-retinopathy detection, while Poplin et al. investigated prediction of cardiovascular risk factors from retinal fundus photographs. EfficientNet introduced compound scaling to balance network depth, width, and image resolution.

Attention and Transformer architectures provide mechanisms for modelling interactions among distant image regions. The Transformer architecture and later Vision Transformer work established self-attention as an effective approach for learning contextual representations. In the present work, a Transformer Encoder is applied to EfficientNet-B3 feature tokens; this is intentionally described as a CNN-Transformer architecture rather than a separate Vision Transformer branch.

Multimodal medical AI combines complementary sources such as images, laboratory measurements, demographics, and electronic health records. Attention-based fusion is attractive because it can learn feature-wise weights rather than treating all modalities as equally informative.

Explainable AI techniques such as Grad-CAM can provide visual evidence for image-based predictions. Such explanations are useful for model inspection, but a heatmap alone should not be interpreted as proof of clinical causality.

A. Research Gap

Existing retinal prediction studies commonly focus on a single disease, a single modality, or a single task. The proposed framework addresses these limitations experimentally by combining retinal-image representations with structured clinical features and producing two risk predictions from a shared fused representation. The study also incorporates visual explanation. The principal remaining gap is external validation with real, patient-matched clinical data, which is outside the evidence available in the current implementation.

TABLE I. SELECTED LITERATURE SURVEY

Author / Study	Year	Approach	Reported metric / limitation
Gulshan et al.	2016	Deep CNN for diabetic retinopathy	High diagnostic performance; retinal screening
Poplin et al.	2018	Deep learning on fundus photographs	Prediction of cardiovascular risk factors
Tan and Le	2019	EfficientNet	Efficient compound scaling
Vaswani et al.	2017	Transformer	Self-attention and long-range relationships
Dosovitskiy et al.	2021	Vision Transformer	Transformer-based image

			representation
Henge et al.	2025	Inception-ResNet blended model	Reported 89.20% in source table
Salluri et al.	2025	Stacked ensemble	Reported 91.60% in source table
Addanki and Sumathi	2025	CVD-Net	Cardiovascular-risk prediction from retinal vasculature

Note: Reported metrics are retained only where they are explicitly available in the supplied project document; they should be checked against the original publications before journal submission.

III. PROPOSED METHODOLOGY

3.1 Dataset and Clinical-Data Construction

The image component uses the Indian Diabetic Retinopathy Image Dataset (IDRiD), which provides retinal fundus images and diabetic-retinopathy severity information. The implementation described in the supplied paper uses 4,128 samples after dataset preparation and an 80/10/10 train/validation/test split. Importantly, the clinical variables are generated from retinal-disease severity information using medically informed distributions rather than being documented as real patient-level clinical measurements. This distinction is critical: the current results therefore demonstrate model behaviour under the implemented experimental data-generation procedure and do not establish clinical validity.

3.2 Image Preprocessing

Retinal images are resized to 224×224×3 and normalized using ImageNet mean and standard deviation values. Training augmentation includes random horizontal flipping, random rotation up to ±15°, colour jitter, and random affine transformations. Augmentation is applied only to the training set. The implementation section also mentions CLAHE and noise reduction; these operations should be reconciled with the actual training code before submission so that the manuscript describes exactly the preprocessing used for the reported experiment.

3.3 CNN-Transformer Image Encoder

EfficientNet-B3 pretrained on ImageNet is used as the image backbone. The first three feature blocks are frozen, while subsequent layers are fine-tuned. The backbone produces a 1536-channel feature representation. The feature map is reshaped into a sequence of spatial tokens and passed through a two-layer Transformer Encoder with eight attention heads. Mean pooling over the sequence produces a 1536-dimensional image representation. This architecture should be referred to consistently throughout the paper as EfficientNet-B3 + Transformer Encoder; the earlier project-report references to a separate Vision Transformer branch are removed here because they conflict with the detailed model configuration.

3.4 Clinical Feature Encoder

The structured feature vector contains six variables: blood pressure, body mass index, glucose level, cholesterol level, age, and diabetes status. Numerical variables are normalized and missing values are handled as described in the implementation. A fully connected network maps the six-dimensional input to 64 hidden units and then to a 128-dimensional clinical representation using ReLU activations.

3.5 Attention-Based Multimodal Fusion

The 1536-dimensional image representation and 128-dimensional clinical representation are concatenated to obtain a 1664-dimensional vector. An attention network uses a 1664→128 linear projection, Tanh activation, a 128→1664 projection, and Sigmoid activation to obtain adaptive feature weights. The weighted representation is then passed through a 1664→256 fusion layer with ReLU and dropout (0.4).

3.6 Multi-Task Prediction

Two task-specific output heads are attached to the fused representation. Each head maps the 256-dimensional fused vector to one logit for binary prediction. During training, BCEWithLogitsLoss is used for the cardiovascular and diabetes tasks. The two outputs provide separate probability estimates for the two target conditions.

3.7 Explainability with Grad-CAM

Grad-CAM is computed from the convolutional feature maps of the image backbone. The resulting heatmap is overlaid on the retinal image to visualize regions contributing to the prediction. The explanation is intended for model inspection and interpretability; it should not be interpreted as a clinical diagnosis.

TABLE II. MODEL CONFIGURATION

Component	Configuration
Image backbone	EfficientNet-B3, ImageNet pretrained
Image feature dimension	1536
Frozen layers	First three feature blocks
Context model	2-layer Transformer Encoder, 8 attention heads
Clinical encoder	Linear(6→64) → ReLU → Linear(64→128) → ReLU
Fusion dimension	1536 + 128 = 1664
Attention network	Linear(1664→128) → Tanh → Linear(128→1664) → Sigmoid
Fusion network	Linear(1664→256) → ReLU → Dropout(0.4)
Output heads	2 × Linear(256→1)

IV. EXPERIMENTAL SETUP

The reported implementation uses the following training configuration:

4.1 Data Split and Leakage Control

The supplied project document reports an 80/10/10 split with 3,304 training, 412 validation, and 412 test samples. For a journal submission, the exact split procedure must be documented in the final experimental record, including whether the split was performed at patient level, whether duplicate or augmented images could cross partitions, and whether any clinical variables were generated or normalized using information from the held-out sets. The test set must remain untouched until final evaluation.

4.2 Evaluation Metrics

Accuracy, precision, recall (sensitivity), and F1-score are reported. For a medical-risk prediction study, specificity, negative predictive value, Matthews correlation coefficient, ROC-AUC, and confidence intervals are also recommended. These additional metrics should be calculated from the final held-out test predictions before submission.

V. RESULTS AND DISCUSSION

A. Dataset Distribution

TABLE IV. DATASET DISTRIBUTION

Set	Diabetes Normal	Diabetes Positive	Heart Normal	Heart Positive	Total
Train	1,072	2,232	2,186	1,118	3,304
Validation	136	276	277	135	412
Test	136	276	261	151	412
Total	1,344	2,784	2,724	1,404	4,128

B. Training Behaviour

The best checkpoint was reported at epoch 4, with validation accuracy of 98.28%. Training accuracy continued to improve after epoch 4 while validation accuracy stabilized around 97–98%, indicating a small degree of overfitting. Selecting the epoch-4 checkpoint is therefore preferable to using the final epoch solely because it has higher training accuracy.

C. Test Performance

TABLE V. TEST PERFORMANCE

Metric	Cardiovascular risk	Diabetes risk	Overall
Accuracy	0.9806	0.9830	0.9818
Precision	0.9799	0.9927	0.9863
Recall / Sensitivity	0.9669	0.9819	0.9744
F1-score	0.9733	0.9872	0.9803
ROC-AUC	0.9985	0.9962	—

The reported test results are strong, with an overall accuracy of 98.18%. Cardiovascular-risk prediction reaches 98.06% accuracy, while diabetes-risk prediction reaches 98.30%. The corresponding precision, recall, and F1-score values are also high. The reported ROC-AUC values of 0.9985 and 0.9962 indicate excellent discrimination on the held-out test

set. Because these values are unusually high for a medical prediction problem, independent verification of the test split, label construction, synthetic clinical-data generation, and absence of data leakage is essential before publication.

D. Confusion-Matrix Results

TABLE VI. CONFUSION-MATRIX RESULTS

Task	TN	FP	FN	TP
Cardiovascular risk	258	3	5	146
Diabetes risk	134	2	5	271

From the reported confusion matrices, the cardiovascular task contains 258 true negatives, 3 false positives, 5 false negatives, and 146 true positives, while the diabetes task contains 134 true negatives, 2 false positives, 5 false negatives, and 271 true positives. These counts should be used to regenerate every reported metric so that the manuscript remains internally consistent.

E. Ablation and Baseline Evaluation

The current project report does not provide a complete controlled ablation study. To substantiate the contribution of multimodal fusion, the final manuscript should evaluate at least: (1) clinical data only, (2) EfficientNet-B3 only, (3) EfficientNet-B3 + Transformer Encoder, (4) image features + clinical features without attention fusion, and (5) the complete proposed model. The same held-out test protocol should be used for every baseline. The resulting table should be populated from actual experiments and not estimated.

Required before submission: replace all 'To be measured' entries with results from actual controlled experiments.

VI. DISCUSSION

The results suggest that combining image-derived representations and structured features can provide a strong experimental predictor under the implemented data-generation and evaluation procedure. EfficientNet-B3 supplies hierarchical spatial features, while the Transformer Encoder provides contextual modelling over the feature sequence. The attention mechanism then learns feature-wise weights over the combined representation.

The very high reported AUC values require cautious interpretation. In particular, the current implementation generates clinical variables from retinal-disease severity information. If a clinical feature is statistically or deterministically related to the target label, the multimodal model may obtain information that would not be available in an independent real-world clinical setting. Therefore, the model should not be described as clinically validated.

Grad-CAM provides a useful qualitative explanation mechanism, but visual correspondence between a heatmap and a retinal structure does not by itself establish clinical relevance. Future work should include clinician-reviewed explanation quality and external validation on independent datasets.

VII. LIMITATIONS

The present study is based primarily on the IDRiD retinal dataset and therefore may not represent the diversity of imaging devices, populations, and clinical settings encountered in practice.

The clinical variables used in the implementation are generated using medically informed distributions rather than documented real patient-level measurements. This limits the clinical interpretation of the multimodal results.

The reported evaluation is retrospective and does not include prospective clinical validation or external validation on an independent clinical cohort.

The manuscript requires an explicit patient-level split and leakage audit before the reported performance can be considered robust.

The current project report does not contain a complete controlled ablation study; such experiments are necessary to quantify the contribution of the Transformer Encoder, clinical branch, and attention fusion.

VIII. DATA ETHICS AND CLINICAL USE

The image data are derived from a publicly available retinal dataset. No claim of prospective clinical use is made in this experimental study. The generated clinical variables are synthetic/derived for the present implementation and should not be interpreted as real patient measurements. Any future study using real patient-level clinical data should obtain the appropriate institutional ethics approval, informed-

consent or waiver documentation, and data-protection safeguards as required by the study setting.

IX. CONCLUSION AND FUTURE WORK

This study presents a multimodal deep learning framework that combines retinal fundus images and structured clinical features for simultaneous cardiovascular and diabetes risk prediction. The proposed implementation uses EfficientNet-B3 for image feature extraction, a Transformer Encoder for contextual modelling, an attention-based fusion network for multimodal integration, and two task-specific prediction heads. Grad-CAM is used to provide visual explanations of image-based predictions.

On the reported held-out test set, the model achieves 98.18% overall accuracy, with 98.06% cardiovascular-risk accuracy and 98.30% diabetes-risk accuracy. The reported ROC-AUC values are 0.9985 and 0.9962, respectively. These results are promising as an experimental proof of concept; however, they should not be interpreted as evidence of clinical diagnostic performance because the clinical variables are generated and external validation is not provided.

Future work should focus on real patient-matched clinical data, patient-level leakage-controlled splits, external validation across independent datasets and imaging devices, controlled ablation studies, confidence intervals, calibration analysis, clinician assessment of Grad-CAM explanations, and prospective clinical evaluation.

REFERENCES

[1] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, 2016.

[2] R. Poplin et al., "Prediction of cardiovascular risk factors from retinal fundus photographs using deep learning," Nature Biomedical Engineering, 2018.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Proc. ICML, 2019.

[4] A. Vaswani et al., "Attention Is All You Need," Proc. NeurIPS, 2017.

[5] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proc. ICCV, 2017.

[6] P. Porwal et al., "Indian Diabetic Retinopathy Image Dataset (IDRiD): A database for diabetic retinopathy screening research," Data in Brief / related IDRiD publication, 2018.

[7] J. Son et al., "Predicting high coronary artery calcium score from retinal fundus images," Nature Machine Intelligence, 2020.

[8] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," Proc. ICLR, 2021.

[9] S. Woo et al., "CBAM: Convolutional Block Attention Module," Proc. ECCV, 2018.

[10] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Proc. ICLR, 2019.

[11] J. Hu et al., "Squeeze-and-Excitation Networks," Proc. CVPR, 2018.

[12] P. Rajpurkar et al., "AI in health and medicine," Nature Medicine, 2022.

[13] S. K. Henge et al., "Detection of Diabetic Retinopathy Using a Multi-Decision Inception-ResNet-Blended Hybrid Model," IEEE Access, vol. 13, pp. 8988–9005, 2025, doi: 10.1109/ACCESS.2024.3525154.

[14] D. K. Salluri et al., "A Synergistic Stacked Ensemble Deep Learning Model for Predicting Diabetic Retinopathy Severity," IEEE Access, vol. 13, pp. 202454–202466, 2025, doi: 10.1109/ACCESS.2025.3637311.

[15] K. Ahnaf Alavee et al., "Enhancing Early Detection of Diabetic Retinopathy Through the Integration of Deep Learning Models and Explainable Artificial Intelligence," IEEE Access, vol. 12, pp. 73950–73969, 2024, doi: 10.1109/ACCESS.2024.3405570.

[16] T. Shahzad et al., "Developing a Transparent Diagnosis Model for Diabetic Retinopathy Using Explainable AI," IEEE Access, vol. 12, pp. 149700–149709, 2024, doi: 10.1109/ACCESS.2024.3475550.

[17] W. Nazih et al., "Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images," IEEE Access, vol. 11, pp. 117546–117561, 2023, doi: 10.1109/ACCESS.2023.3326528.

[18] C. Choudhary, A. Mathur, and R. Gupta, "Diabetic Retinopathy Detection: A Transfer

Learning Based Approach for Accurate Diagnosis," Proc. ICCCIS, 2023, pp. 280–285, doi: 10.1109/ICCCIS60361.2023.10425349.

[19] S. Addanki and D. Sumathi, "Prediction of Cardiovascular Disease Risk from Retinal Vasculature Using a Quantitative Diagnostic Approach With CVD-Net in DR and HR Patients," IEEE Access, vol. 13, pp. 171406–171421, 2025, doi: 10.1109/ACCESS.2025.3610424. Manuscript prepared in a compact IJIRT-compatible research-paper structure. Author details and final experimental validation must be completed before submission.