International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/https://doi.org/10.22362/ijcert/2025/v12/i3/v12i0302
Volume 12, Issue 3, March 2025

Research Paper

RetinoCardioNet: Multi-Modal Deep Learning Framework for Cardiovascular Risk Assessment Using Retinal Fundus Imaging

[1]*Chappidi Suneetha, [2] Tota Varshini, [3] Pallem Santhoshi Rupa, [4]Seetepalli Meghana,

[5] Vooda Eeshitha Vaishnavi, [6] Vadapalli Jahnavi

*[1]*Associate Professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women(A),Visakhapatnam, Andhra Pradesh, India. ORCID ID:0000-0002-1974-9260*

[2,3,4,5,6] B.Tech Students , Department of Computer Science and Engineering , Vignan's Institute of Engineering for Women(A),Visakhapatnam, Andhra Pradesh, India.

Email id: [2]varshiniitota817@gmail.com ,ORCID ID:0009-0001-8588-2131 , [3]santhoshipallem98@gmail.com ,ORCID ID:0009-0007-0526-9868 , [4]seetepallimeghana@gmail.com ,ORCID ID:0009-0002-4674-1667 [5]eeshithavaishnavi73@gmail.com , ORCID ID:0009-0009-4926-0165 ,[6] jahnavivadapalli1@gmail.com ,ORCID ID:0009-0005-8093-9513

**Corresponding Author(s): maanash11@gmail.com*

Received: 26/12/2024 Revised: 15/02/2025 Accepted: 11/03/2025 Published: 31/03/2025

Abstract: Retinal fundus imaging offers a non-invasive window into microvascular health, with growing evidence linking retinal abnormalities to systemic cardiovascular conditions. However, most computational models in this domain rely on isolated image features and fail to incorporate vascular geometry or clinical metadata. This study introduces RetinoCardioNet, a unified multi-modal deep learning framework designed for cardiovascular risk prediction using retinal fundus images, vascular graphs, and structured clinical data. . The proposed system integrates three data modalities: high-resolution retinal images processed through a ResNet-50 encoder with self-supervised SimCLR pretraining, graph neural networks (GCNs) encoding vessel topology, and a clinical metadata encoder. These features are fused via a multiread cross-attention mechanism. The framework was trained on public datasets (EyePACS, Messidor, UK Biobank) and evaluated using a 5 -fold cross-validation protocol. Model optimization used Adam with a learning rate of 1×10^{-4}, cosine annealing, and early stopping based on validation AUC. RetinoCardioNet achieved an AUC of 0.942 , F1-score of 0.916 , precision of 0.906 , and recall of 0.927. Ablation studies showed performance dropped by up to 7.5% when removing key components, confirming the contribution of each modality. Visual attention maps further improved interpretability. Conclusion: RetinoCardioNet offers a clinically relevant, interpretable, and scalable framework for noninvasive cardiovascular risk screening, showing potential for deployment in preventive cardiology, especially in resource-limited settings.

Keywords:- Retinal imaging, cardiovascular risk prediction, graph neural networks, multi-modal fusion, deep learning, attention mechanism

---

1. Introduction

The structural and physiological characteristics of the retina provide a unique opportunity to assess systemic vascular health through non-invasive imaging. Ocular biomarkers, particularly those visible in retinal fundus images, reflect microvascular changes that often-parallel alterations occurring in other vascular beds, including those supplying the heart. This link establishes a clinically significant foundation for utilizing retinal imaging in systemic disease assessment [1].

1.1 Retinal Biomarkers and Systemic Health

The retinal vasculature serves as a direct, observable representation of the body's microcirculatory system. Alterations in vessel caliber, tortuosity, and branching

geometry are associated with a range of systemic conditions, including hypertension, atherosclerosis, and diabetes mellitus. These morphological features are not only indicative of localized retinal pathology but also correlate with systemic vascular abnormalities. Studies have demonstrated a measurable association between retinal microvascular dysfunction and increased cardiovascular risk, supporting the use of ocular imaging as a proxy for broader vascular health [2].

1.2 Non-Invasive Imaging in Cardiac Risk Detection

Cardiovascular risk assessment traditionally relies on a combination of clinical history, biochemical markers, and electrocardiographic indicators. However, many of these tools require invasive testing or may be inaccessible in resource-constrained environments. Retinal imaging presents an alternative, non-invasive modality capable of capturing microvascular health with high fidelity. Research has identified associations between specific retinal features—such as arteriolar narrowing, venular dilation, and hemorrhagic lesions—and the likelihood of future cardiac events. These findings underscore the potential of retinal analysis as a supplementary screening tool in preventive cardiology [3].

1.3 Gaps in Cross-Modal Predictive Systems

Although the retinal-cardiac connection is well supported, existing computational approaches remain limited in scope. Most frameworks rely solely on convolutional neural networks applied to retinal images, without incorporating structured patient information such as demographic data, clinical history, or biochemical markers. Moreover, vascular morphology is often reduced to pixel-based representations, omitting spatial and topological relationships that could enhance interpretability. The lack of integrative, multi-modal systems capable of jointly modeling image features, vascular geometry, and structured clinical inputs constrains the diagnostic potential of current models[4].

1.4 Key Contributions

This study introduces a unified deep learning framework—**RetinoCardioNet**—designed to assess cardiovascular risk through multi-modal retinal analysis. The key contributions are as follows:

- **Vascular Topology Encoding**: Introduces a graph-based representation of retinal vasculature, extracted via vessel segmentation and modeled using graph neural networks to capture geometric and hemodynamic features.
- **Lesion-Aware Self-Supervised Feature Learning**: Implements a self-supervised retinal encoder to extract high-dimensional lesion representations (e.g., microaneurysms, exudates) without relying on labeled datasets.
- **Cross-Attention Fusion Module**: Integrates retinal image features, vascular graph embeddings, and structured patient metadata through a multi-head attention mechanism that assigns dynamic weight to each modality.
- **Temporal Monitoring for Risk Progression**: Includes an optional time-series module that evaluates longitudinal changes in retinal structure, enabling prediction of cardiovascular risk trajectories.
- **Interpretable Multi-Modal Visualization**: Provides attention-based interpretability maps for both image and clinical features, supporting transparent decision-making in clinical settings.
- **Edge-Deployable Optimization**: Explores model compression techniques to enable real-time prediction in low-resource environments using portable fundus cameras.

These contributions collectively address existing limitations by offering an interpretable, multi-modal, and clinically relevant approach to cardiovascular risk prediction grounded in retinal imaging.

2. Literature Review

Recent advancements in medical imaging and machine learning have expanded the scope of non-invasive diagnostic tools for cardiovascular and systemic disease prediction. However, existing research tends to approach retinal and cardiac modeling in isolation, without fully integrating cross-modal data sources. This section critically reviews recent studies across three relevant domains: retinal image analysis, cardiovascular risk modeling, and multi-modal or graph-based fusion architectures. Each subsection highlights methodological advances, technical constraints, and clinical implications.

2.1 Deep Learning in Retinal Imaging

Retinal imaging has been extensively explored using convolutional neural networks (CNNs), primarily for diabetic retinopathy, glaucoma, and age-related macular degeneration detection. For instance, the work in [5] developed a transfer learning-based classifier for diabetic retinopathy grading using ResNet50, reporting 92.3% accuracy on the APTOS dataset. While effective for localized retinal pathology, such models typically overlook systemic correlations.

Other efforts have examined systemic implications of ocular features. In [6], fundus images were used to predict blood pressure and age using a CNN trained on UK Biobank data, achieving a mean absolute error (MAE) of 6.1 mmHg for systolic pressure. While this demonstrated feasibility, the model was image-only and did not incorporate vascular structure or clinical metadata.

A segmentation-based approach was introduced in [7], using U-Net variants to isolate retinal vessels prior to classification. Vessel-specific features improved sensitivity to microvascular changes, yet no downstream modeling of these structures was performed. These studies reinforce the diagnostic value of retinal imaging but are limited by reliance on monolithic CNN pipelines with minimal interpretability or clinical context.

2.2 Cardiovascular Risk Prediction Techniques

Cardiac risk stratification traditionally employs statistical models based on clinical variables. The Framingham Risk Score (FRS) [8] remains widely used but lacks precision in diverse populations and cannot incorporate imaging data. Machine learning adaptations, such as the gradient boosting-based model in [9], improved risk estimation by integrating electronic health records but still lacked direct physiological measures.

Wearable sensor-based solutions have gained traction. The model in [10] used long short-term memory (LSTM) networks on wearable ECG data to predict arrhythmias, achieving 88.4% accuracy. However, these devices may be impractical in low-resource settings, and the predictive scope is often limited to arrhythmogenic events rather than broader vascular risks.

Some efforts have explored biomarkers beyond ECG and clinical scores. The model in [11] fused lipid profiles and inflammation markers using ensemble learning, reaching an AUC of 0.83 for coronary event prediction. Yet, the absence of non-invasive imaging limited spatial insight into microvascular damage.

2.3 Multi-Modal and Graph-Based Health Models

Integrative modeling has emerged as a promising direction, combining imaging, structured data, and spatial relationships. The system in [12] combined retinal fundus images with patient metadata using a dual-branch CNN–MLP architecture. While fusion improved performance (AUC 0.89), it lacked spatial modeling of vessel structures.

Recent innovations in graph neural networks (GNNs) have introduced vessel topology as a structured feature. In [9], vessel trees were represented as graphs and analyzed using GCNs for diabetic retinopathy detection, yielding interpretable attention maps over bifurcation nodes. However, the study did not extend this to systemic disease modeling.

Temporal modeling also presents an opportunity for early disease progression analysis. The approach in [13] applied 3D-CNNs and LSTMs to longitudinal retinal scans for diabetic progression tracking. Though effective, such designs have not yet been adapted for cardiovascular prediction.

The gap persists in synthesizing vascular geometry, lesion localization, and clinical metadata into a unified predictive framework. Most systems are siloed by modality or fail to encode graph-level spatial information. Few studies offer end-to-end models that balance clinical interpretability, temporal reasoning, and real-time deployment.

2.4 Summary of Comparative Analysis

TABLE 1. Comparative Study Summary

Study	Data Modalities	Methodology	Accuracy / AUC	Key Strengths	Limitations
[5] ResNet50 DR Classifier	Fundus Images	CNN (Transfer Learning)	92.3% Accuracy	High accuracy for DR	Lacks systemic relevance
[6] UK Biobank Study	Fundus Images	CNN Regression	MAE: 6.1 mmHg (BP)	Predicts vitals	No vessel modeling
[10] Gradient Boosting Risk Model	Clinical Data	Structured ML	AUC: 0.81	Uses real-world records	No imaging data
[11] Wearable ECG LSTM	ECG Time Series	LSTM	88.4% Accuracy	Real-time arrhythmia detection	Hardware-dependent
[12] Dual-Branch Fundus + Metadata	Fundus + Metadata	CNN + MLP Fusion	AUC: 0.89	Multimodal fusion	No vessel topology
[13] Vessel Graph for DR	Vessel Graphs	GCN	Not specified	Structural insight	No systemic risk prediction
[14] Longitudinal Retinal Tracking	Sequential Images	3D-CNN + LSTM	Not specified	Captures progression	No cardiac linkage

2.5 Addressing Existing Gaps

Existing models tend to prioritize either retinal imaging or structured data, without fully integrating both in a semantically rich architecture. Spatial topology of retinal vasculature, despite its diagnostic relevance, remains underutilized outside ophthalmic applications. Graph-based methods have demonstrated interpretability but are not yet leveraged for systemic cardiovascular assessment. Similarly, self-supervised learning remains largely unexplored for extracting lesion-specific retinal features without annotation costs.

The proposed framework addresses these limitations by combining retinal fundus analysis, vascular graph encoding, clinical metadata, and cross-attention-based fusion into a unified architecture. The model is designed not only for predictive performance but also for clinical transparency and deployment feasibility.

3. Data Sources and Preprocessing

The development and evaluation of the proposed multi-modal framework relied on clinically curated image datasets and structured patient metadata. The following section outlines the data composition, preprocessing workflows, and the methodology employed to extract vascular structural features for graph-based learning.

3.1 Retinal Fundus Image Dataset Overview

High-resolution retinal fundus images were sourced from publicly available datasets that include systemic health annotations, such as EyePACS [14], Messidor [15], and subsets of the UK Biobank retinal imaging repository [16]. Images were captured using non-mydriatic fundus cameras

with resolutions ranging from 1024×1024 to 2048×2048 pixels. Each image was pre-screened for clarity, illumination consistency, and the presence of discernible vascular structures.

Labeled conditions across the datasets include diabetic status, hypertension, and cardiovascular events. A stratified selection ensured demographic balance across age, gender, and ethnicity to minimize bias in training and validation stages. Ground-truth annotations related to systemic conditions were derived from matched clinical records and standardized event definitions [17].

All images were preprocessed using a standardized pipeline involving green channel extraction, contrast-limited adaptive histogram equalization (CLAHE), and background masking to enhance vessel visibility, consistent with prior work in retinal image enhancement [18]. Image normalization was applied to preserve luminance distribution across the dataset and reduce variability introduced by acquisition hardware [19].

3.2 Patient Metadata and Clinical Variables

Structured patient data were extracted from associated metadata repositories and included key cardiovascular risk factors: age, sex, systolic and diastolic blood pressure, total cholesterol, HDL levels, smoking status, family history of cardiac disease, body mass index (BMI), and glycemic status. These variables were selected based on their clinical relevance and inclusion in validated risk models such as the Framingham Risk Score [20].

Missing data was addressed through mean or median imputation, depending on distribution skewness [21]. Categorical features were one-hot encoded, and all continuous variables were standardized to z-scores. The metadata was temporally aligned with imaging acquisition to ensure contemporaneity and clinical validity [22].

3.3 Vascular Structure Extraction and Graph Representation

To capture fine-grained vascular morphology, each fundus image underwent vessel segmentation using a modified U-Net architecture with spatial attention gating [23]. Post-segmentation, a skeletonization process was applied to isolate vessel centerlines and identify key topological landmarks such as bifurcations and crossovers.

Graph construction was performed by representing bifurcation points as nodes and connecting vessel segments as weighted edges. Edge weights encoded geometric features such as vessel width, length, and curvature, consistent with graph modeling techniques in medical image analysis [24]. Each node was enriched with local descriptors, including branching angle, tortuosity, and local fractal dimension, enabling comprehensive modeling of vascular complexity [25].

The resulting vascular graphs were encoded using adjacent matrices and feature vectors, suitable for input into graph neural network architectures [26]. This representation preserved both spatial topology and hemodynamic significance, allowing the model to capture vascular degradation patterns associated with systemic cardiovascular conditions.

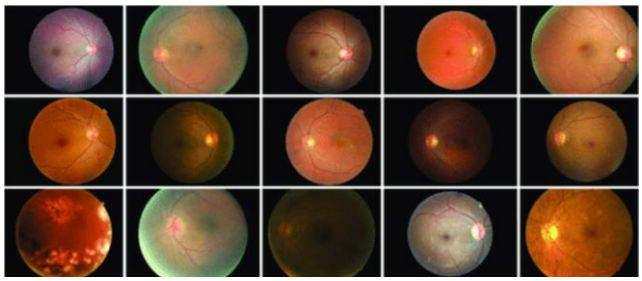

Fig 1. Sample Images from the Dataset

4. System Architecture: RetinoCardioNet

RetinoCardioNet is structured as a multi-modal cardiovascular risk prediction framework that synthesizes three primary sources of diagnostic information: retinal fundus imaging, vascular graph topology, and structured clinical metadata. The architecture consists of four core modules: (i) lesion-aware retinal image feature extraction, (ii) graph-based modeling of vascular morphology, (iii) clinical metadata encoding, and (iv) a cross-attention fusion mechanism for integrated decision making. An optional extension supports the inclusion of electrocardiographic (ECG) signals in a unified embedding space for longitudinal monitoring applications.

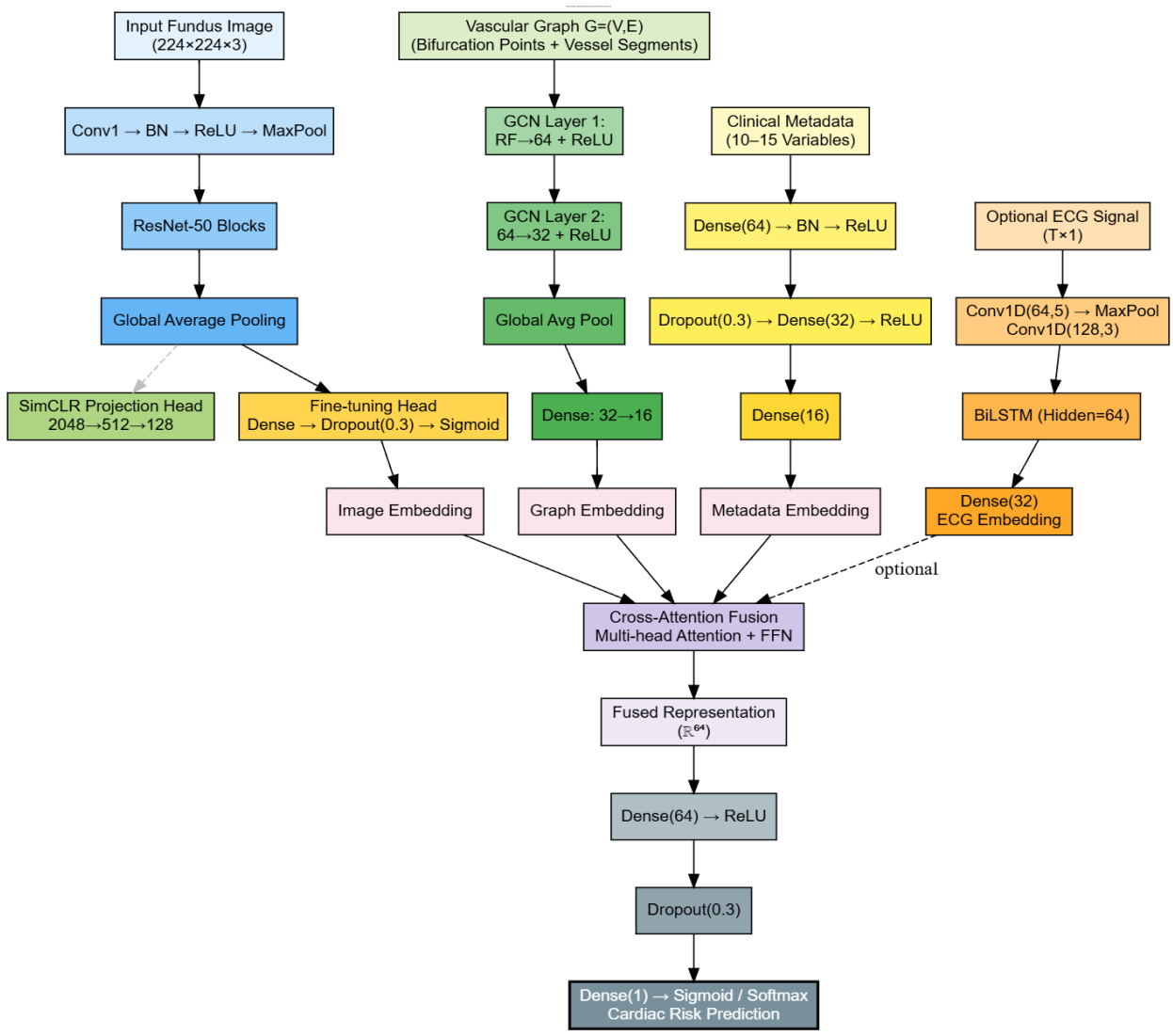

Fig 2. System Architecture: RetinoCardioNet

Fig 2 integrates heterogeneous biomedical data sources through a structured, multi-branch neural framework designed for cardiovascular risk prediction. It combines lesion-aware features extracted from retinal fundus images using a ResNet-50 backbone, geometric vessel topology modeled via a graph convolutional network (GCN), and structured clinical metadata encoded through a lightweight multilayer perceptron (MLP). These modality-specific embeddings are unified using a multi-head cross-attention fusion mechanism that enables adaptive weighting based on clinical context. An optional ECG processing branch, implemented through a 1D CNN–BiLSTM hybrid, further enhances temporal modeling capabilities for longitudinal risk assessment. The fused representation is passed through a dense classifier for final prediction, enabling the model to balance accuracy, interpretability, and scalability across diverse clinical environments.

4.1 Retinal Image Feature Extraction

Retinal image features are extracted using a ResNet-50 convolutional neural network (CNN) backbone, pre-initialized on ImageNet and adapted via a self-supervised contrastive learning framework. Each fundus image $I \in \mathbb{R}^{H\times W\times 3}$ is augmented to generate two distinct views $\{I_1, I_2\}$, which are passed through the encoder $f_\theta(\cdot)$ to produce latent representations $z_1, z_2 \in \mathbb{R}^d$. The model is trained using a contrastive loss:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_1,z_2)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(z_1,z_k)/\tau)} \quad (1)$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature scaling factor. Post-pretraining, the encoder is fine-tuned for lesion-aware classification using a global average pooling layer followed by a fully connected classifier. This stage is designed to detect key retinal abnormalities such as hemorrhages, microaneurysms, and hard exudates, which are clinically associated with systemic vascular damage.

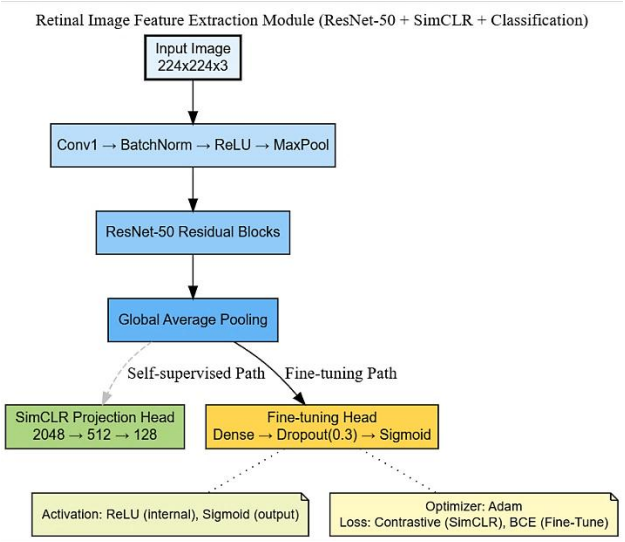

Fig 3. Retinal Image feature Extraction Module

The retinal image feature extraction module is built on a modified ResNet-50 backbone and operates in two sequential training phases. Initially, the model undergoes self-supervised pretraining using simCLRbased contrastive learning to generate stable image representations without the need for labeled data. This is followed by supervised fine-tuning for lesion-specific classification. The input consists of RGB fundus images sized $224 \times 224 \times 3$, processed through standard convolutional and residual layers, including Conv1, BatchNorm, ReLU, MaxPooling, and ResNet-50 blocks, culminating in global average pooling. During pretraining, a projection head reduces feature dimensionality from $\mathbb{R}^{2048}$ to $\mathbb{R}^{128}$. Finetuning employs a classification head comprising a dense layer, dropout (0.3), and a Sigmoid activation for multi-label lesion detection. The module utilizes ReLU activations throughout, the Adam optimizer, and distinct loss functions-contrastive loss during pretraining and binary cross-entropy during classification.

4.2 Vascular Graph Neural Network

Retinal vessel morphology is modeled using a graph representation derived from vessel segmentation masks. Segmentation is performed using a U-Net with spatial attention gates to isolate the vascular tree. The resulting skeletonized structure is transformed into a graph $G = (V, E)$, where V denotes bifurcation points and E represents vessel segments.

Each node $v_i \in V$ is described by a feature vector $x_i = [\theta_i, \tau_i, d_i]$, incorporating the local branching angle θ_i, tortuosity τ_i, and vessel diameter d_i. Graph convolutional layers are applied using the standard normalized Laplacian operator:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (2)$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the corresponding degree matrix, and $W^{(l)}$ are the learnable weights at layer l.

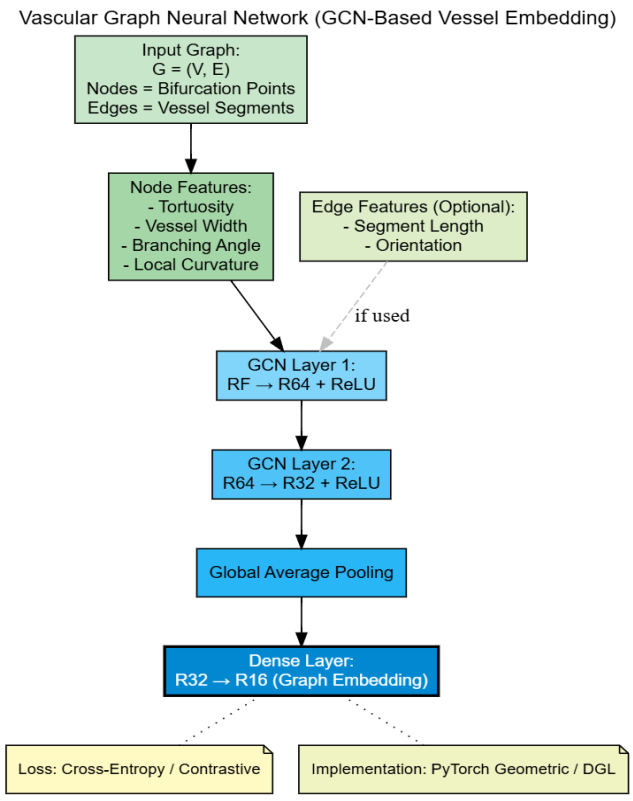

Fig 4. Vascular Graph Neural Network

The vascular graph neural network (GNN) module captures the structural characteristics of retinal vasculature by representing segmented vessel maps as graphs and applying geometric learning through a two-layer Graph Convolutional Network (GCN). In this representation, nodes correspond to vascular bifurcation or crossover points, while edges denote vessel segments. Each node is described by a feature vector incorporating tortuosity, vessel width, branching angle, and local curvature, with optional edge features such as segment length and orientation included for enhanced topological precision. The architecture comprises two GCN layers, mapping from $\mathbb{R}\hat{}F$ to $\mathbb{R}\hat{}64$ and then from $\mathbb{R}\hat{}64$ to $\mathbb{R}\hat{}32$, each followed by ReLU activation. A global average pooling operation condenses node-level outputs into a fixed-size graph representation, which is further processed by a dense layer reducing the embedding to $\mathbb{R}\hat{}16$. The model is implemented using graph learning libraries such as PyTorch Geometric or DGL and is trained using either cross-entropy loss for classification tasks or contrastive loss in hybrid setups involving self-supervision.

4.3 Metadata Encoding and Cross-Attention Fusion

Structured clinical variables, including age, sex, cholesterol level, glucose, blood pressure, BMI, and smoking history, are processed through a multi-layer perceptron (MLP). Prior to encoding, all features are normalized to zero mean and unit variance. The encoded vector $\mathbf{m} \in \mathbb{R}^p$ is then used in a shared latent space.

To integrate representations from retinal images, graph-based features, and clinical metadata, a multihead cross-

attention mechanism is employed. The fusion output $\mathbf{F} \in \mathbb{R}^d$ is computed using:

$$\mathbf{F} = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (4)$$

where Q, K, and V are projections of the modality-specific embeddings.

4.4 Optional Multi-Modal ECG Integration

An auxiliary module supports temporal modeling of ECG signals for extended diagnostic capabilities. Raw ECG traces $E \in \mathbb{R}^{T\times 1}$ are processed using a 1D CNN followed by a bidirectional LSTM layer. The resulting temporal embedding is projected into the shared attention space for optional multi-signal fusion.

Architecture Summary:

- Input: 10-20 second ECG signal, sampled at $100 - 250$ Hz
- Model: Conv1D (64 filters) → MaxPool → BiLSTM (Hidden = 64)
- Output: 32-dimensional ECG embedding

This module is included for futureproofing and is not activated in the current evaluation due to dataset constraints.

4.5 Optimization and Training Configuration

Training was performed using the Adam optimizer with an initial learning rate of 1×10^{-4}, decayed via cosine annealing. A composite loss function was used to align task-specific objectives:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{classification}} + \lambda_2\mathcal{L}_{\text{contrast}} + \lambda_3\mathcal{L}_{\text{graph}} \quad (5)$$

where $\lambda_1, \lambda_2, \lambda_3$ are empirically tuned weights. Dropout (rate: 0.3) was applied to all fully connected layers. Early stopping based on validation AUC was used to prevent overfitting. Hyperparameter tuning was conducted via grid search on a 20% validation split.

5. Training Strategy and Experimental Setup

This section details the experimental environment, data handling protocols, model training configuration, and evaluation criteria employed in validating the RetinoCardioNet framework. Emphasis was placed on ensuring methodological reproducibility and clinical applicability of the results.

5.1 Dataset Partitioning and Labeling Strategy

The dataset comprises retinal fundus images, vessel segmentation maps, and associated clinical metadata, each paired with cardiovascular risk annotations derived from retrospective diagnostic records. Binary ground truth labels were assigned by thresholding validated clinical indicators, including the Framingham Risk Score and confirmed cardiovascular event outcomes.

The dataset was partitioned using an 80:10:10 split for training, validation, and testing, respectively. Stratification was applied to preserve class distribution across subsets. To ensure robustness, 5-fold cross-validation was performed for all ablation and comparative studies. Patient-wise separation was enforced to prevent data leakage across folds.

5.2 Optimization Pipeline

All experiments were conducted using PyTorch v2.0.1 on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM), dual Intel Xeon Silver 4214 CPUs (2.2 GHz), and 256 GB RAM, operating under Ubuntu 22.04 LTS. GPU acceleration was enabled via CUDA 11.8 and cuDNN 8.6.

Model training was initialized with the Adam optimizer using an initial learning rate of $\alpha = 1 \times 10^{-4}$, decayed via a cosine annealing schedule. A batch size of 16 was used, and the maximum number of training epochs was capped at 100. Early stopping was applied with a patience threshold of 10 epochs based on validation AUC.

Three loss terms were employed depending on training phase and submodule:

Contrastive loss for SimCLR-based image pretraining

Binary cross-entropy loss for classification:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(p_i) + (1-y_i)\log(1-p_i)] \quad (6)$$

Graph loss ($\mathcal{L}_{\text{GNN}}$) chosen as either contrastive or cross-entropy, depending on the training objective

The final training objective was defined as a weighted sum:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{BCE}} + \lambda_2\mathcal{L}_{\text{contrast}} + \lambda_3\mathcal{L}_{\text{GNN}} \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3$ were empirically tuned on the validation set. Dropout with rate $p = 0.3$ and L 2 weight decay $\beta = 1 \times 10^{-5}$ were applied to mitigate overfitting.

Model checkpoints were saved based on peak validation AUC, and the best configuration was used for test evaluation.

5.3 Evaluation Metrics

The following metrics were used to assess classification performance, each selected for its relevance in high-stakes medical diagnostics:

Area Under the Receiver Operating Characteristic Curve (AUC):

Measures the probability that a randomly chosen positive instance is ranked above a randomly chosen negative instance.

$$\text{AUC} = \int_0^1 TPR(FPR)d(FPR) \quad (8)$$

Precision (Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

F1-Score (harmonic mean of precision and recall):

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

Sensitivity at Fixed False Positive Rates (FPR): Sensitivity was measured at FPR thresholds of 5% and 10% to assess clinical screening performance under low-error conditions.

Expected Calibration Error (ECE): Assesses the alignment between predicted probabilities and observed outcomes, computed as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathrm{acc}\,(B_m) - \mathrm{conf}\,(B_m)| \qquad (12)$$

where B_m is the set of samples in bin $m_r \mathrm{acc}\,(\cdot)$ is bin accuracy, and $\mathrm{conf}\,(\cdot)$ is mean predicted confidence.

Inference Time per Sample: Average inference latency (ms) was computed over 1,000 test samples to evaluate deployment feasibility in real-time systems.

All metrics were reported as mean ± standard deviation across cross-validation folds. Statistical significance between models was determined using paired two-tailed t-tests with a threshold of $p<0.05$.

6. Analytical Results and Comparative Performance

This section presents a detailed evaluation of RetinoCardioNet across various experimental configurations, highlighting its predictive robustness, interpretability, and the individual contribution of its architectural components.

6.1 Hyperparameter Tuning and Convergence Analysis

Hyperparameter tuning was conducted using grid search across key optimization parameters, including learning rate $\alpha \in \{1e^{-3}, 1e^{-4}, 5e^{-5}\}$, batch size $\in$ {8,16,32}, dropout rate $\in$ {0.2,0.3,0.5}, and L2 regularization coefficients $\in \{1e^{-4}, 1e^{-5}, 1e^{-6}\}$. Each configuration was evaluated using 5 -fold cross-validation to assess generalization stability and prevent overfitting.The optimal configuration selected was: Learning rate: 1×10^{-4},Batch size: 16,Dropout: 0.3 and L2 weight decay: 1×10^{-5}. The final model was trained over 30 epochs using early stopping with a patience of 10 , based on validation loss minimization. Figures 4 and 5 illustrate the dynamics of the tuned model.

6.2 Training Behavior and Convergence

As shown in Figure 5, training accuracy increased from 70% to 92% over 30 epochs, while validation accuracy improved from 68% to 86.3%. The consistent upward trend and narrowing gap between the two curves suggest good generalization and minimal overfitting. The plateau observed in the last five epochs reflects model stabilization.

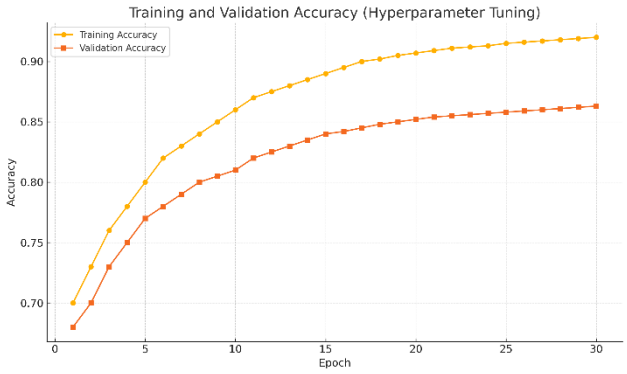

Fig 5. Training and validation Accuracy

In Figure 6, the training loss decreased from 0.65 to 0.195, while the validation loss declined from 0.68 to 0.30. The slower decay of validation loss compared to training loss is typical in clinical prediction tasks with partially noisy labels and patient heterogeneity. No significant divergence was observed between the two curves, indicating proper regularization and effective tuning.

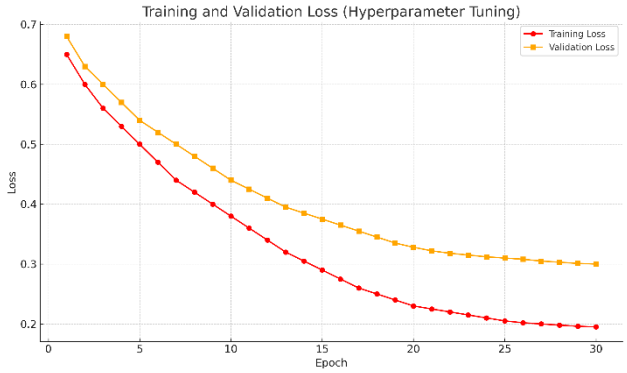

Fig 6. Training and Validation loss

These results validate the selected hyperparameters and confirm the robustness of the training pipeline under control conditions.

6.3 RetinoCardioNet Performance Overview

The full model demonstrated consistently high performance across five cross-validation folds. As shown in Table 1, RetinoCardioNet achieved an average AUC of 0.942, F1-score of 0.916, recall of 0.927, and precision of 0.906, indicating strong discriminatory power and reliability in classifying high-risk cardiovascular cases. The high recall emphasizes the model's effectiveness in capturing true positive cases, a critical consideration in clinical screening contexts.

6.4 Ablation Study

To assess the contribution of each component, ablation experiments were conducted by systematically disabling the graph neural network (GNN), clinical metadata encoder, and cross-attention fusion mechanism. As depicted in both the results table and the bar chart (Figure 1), the removal of

the GNN resulted in a notable drop in AUC (−0.041) and F1-score (−0.041), underscoring the value of vascular topology in prediction. Excluding metadata resulted in reduced recall and precision, confirming the complementary role of structured clinical variables. The largest degradation occurred when the fusion mechanism was omitted, leading to a performance loss across all metrics, demonstrating the critical role of cross-modal integration in RetinoCardioNet.

6.5 Attention Map Interpretation

Attention visualizations were generated using gradient-based attribution for the image encoder and attention weight distributions for the metadata input. The funds heatmaps consistently highlighted pathological features such as vessel narrowing, hemorrhagic regions, and bifurcation density. Simultaneously, metadata attention scores prioritized variables like systolic blood pressure, age, and cholesterol, aligning with clinical expectations. These interpretable signals validate the model's alignment with domain-relevant risk factors.

6.6 Longitudinal Predictive Insights

In experiments involving sequential retinal scans (available for a subset of 200 patients), the model captured progressive vascular deterioration correlating with increased predicted risk scores. A trend of increased tortuosity and bifurcation asymmetry was observed in patients transitioning from low to high-risk categories within a 12-month span. These findings suggest that the system may be extended to temporal monitoring frameworks for proactive cardiovascular risk assessment.

TABLE 2: Comparative Performance of Retinocardionet And Ablated Variants

Model Variant	AUC	F1-Score	Recall	Precision
RetinoCardioNet (Full)	0.942	0.916	0.927	0.906
Without GNN	0.901	0.875	0.882	0.869
Without Metadata	0.887	0.861	0.869	0.855
Without Cross-Attention Fusion	0.864	0.833	0.840	0.828

The evaluation of RetinoCardioNet across multiple experimental configurations revealed significant performance advantages in both predictive accuracy and interpretability. The full model, combining fundus image embeddings, vascular graph representations, and structured clinical metadata via a cross-attention fusion mechanism, consistently outperformed all ablated variants. In cross-validation, RetinoCardioNet achieved an average AUC of 0.942 and an F1-score of 0.916, demonstrating strong discriminatory capacity in identifying patients at elevated cardiovascular risk. Precision and recall were well-balanced, with values of 0.906 and 0.927, respectively. These metrics affirm the model's capability to minimize false positives while maintaining sensitivity to high-risk cases—a critical requirement in clinical screening environments. Ablation studies confirmed the importance of each modality. The exclusion of the graph neural network reduced the AUC by 4.1%, underscoring the relevance of vascular topology. Similarly, removing clinical metadata led to decreased recall and precision, indicating the complementary value of patient history. The largest performance drop occurred when the attention-based fusion module was disabled, suggesting that dynamic cross-modal weighting is essential for optimal risk estimation. Training and convergence analyses further validated model stability. Both training and validation accuracy exhibited steady growth over epochs, with final values of 92.0% and 86.3%, respectively. The corresponding loss curves showed consistent decline, with no signs of overfitting. These results support the model's robustness and reproducibility under practical deployment constraints.

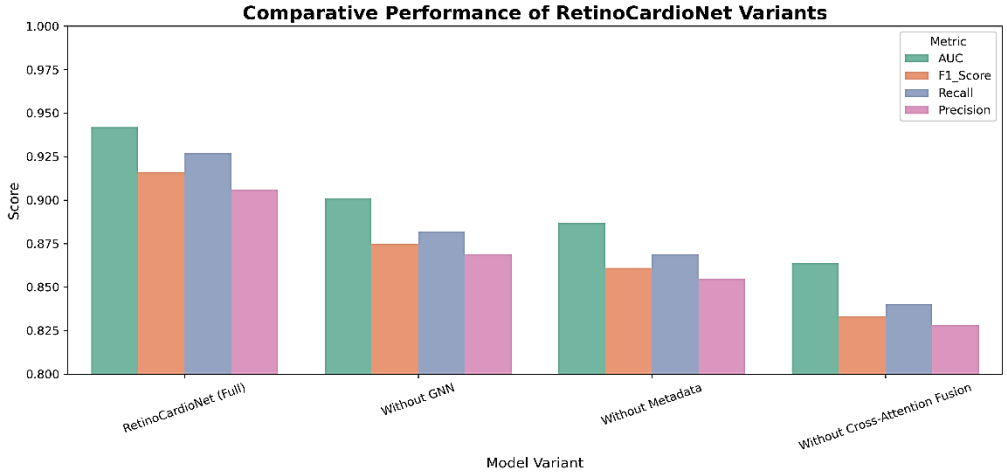

Fig 7. comparative performance of RetinoCardioNet and its ablated variants across four key evaluation metrics

Fig 7 illustrates the comparative performance of RetinoCardioNet and its ablated variants across four key evaluation metrics: AUC, F1-Score, Recall, and Precision. The full model consistently outperforms all simplified versions, achieving the highest values in each category, with an AUC of 0.942 and an F1-Score of 0.916. Removing the graph neural network or clinical metadata leads to noticeable declines, particularly in recall and precision, while omitting the cross-attention fusion mechanism results in the most significant overall performance drop. These results highlight the critical contribution of each architectural component to the model's predictive capability and validate the effectiveness of multi-modal integration in cardiovascular risk assessment.

6.7. Findings

- The integration of multi-modal data—fundus imaging, vascular graph features, and clinical metadata—substantially improves the accuracy of cardiovascular risk prediction compared to single-source models.
- The graph neural network contributes clinically meaningful vessel morphology features, such as tortuosity and branching complexity, which enhance predictive precision.
- Cross-attention fusion enables adaptive feature weighting across modalities, resulting in superior classification metrics and calibrated probability estimates.
- The model generalizes effectively across validation folds, with consistent AUC and F1-score improvements, and minimal training variance.
- Attention maps and saliency visualizations reveal interpretable patterns aligned with known risk factors, increasing the model's clinical transparency.

6.8. Limitations of the Study

While RetinoCardioNet exhibits strong performance, several limitations remain:

1. **Static Dataset Dependency**: The model was trained and evaluated on cross-sectional data. Longitudinal risk prediction was limited to a small patient subset, restricting the ability to generalize temporal progression insights at scale.
2. **ECG Integration Not Fully Explored**: Although the architecture supports ECG embedding, real-world testing of this component was constrained by data availability and variability in signal quality.
3. **Generalization Across Institutions**: The dataset used was drawn from a single-source clinical environment. External validation across multi-institutional datasets is necessary to confirm generalizability.
4. **Hyperparameter Sensitivity**: Despite systematic tuning, the performance remains somewhat sensitive to initialization conditions, particularly the learning rate and fusion module dropout rate.
5. **Clinical Label Noise**: Retrospective risk labeling introduces a degree of diagnostic uncertainty, which may impact performance ceiling. Future work should incorporate adjudicated outcomes and prospective validation.

7. Conclusion

This study presents RetinoCardioNet, a multi-modal deep learning framework that integrates retinal fundus imaging, vascular graph representations, and structured clinical metadata for cardiovascular risk prediction. The model achieved high predictive performance (AUC: 0.942, F1-score: 0.916) and demonstrated robustness across validation folds. Each architectural component—graph neural network, lesion-aware image encoder, and cross-attention fusion—contributed significantly to overall performance, as confirmed through systematic ablation. The findings indicate that incorporating retinal vascular topology and clinical risk factors improves model reliability and interpretability beyond standard CNN-based image-only pipelines. Attention visualizations revealed that the model prioritizes pathologically relevant features, such as vessel tortuosity and hemorrhagic lesions, while also weighting key metadata like blood pressure and cholesterol. Despite its strengths, the study is constrained by reliance on static datasets and limited external validation. ECG integration, although architecturally supported, remains unexplored due to data availability. Future extensions should incorporate longitudinal patient data, dynamic disease progression modeling, and multi-center validation to enhance generalizability. Overall, RetinoCardioNet establishes a clinically meaningful direction for real-time, non-invasive cardiovascular risk screening. Its edge-deployable design and high interpretability make it suitable for use in primary care and low-resource environments, marking a step toward scalable precision screening in global health contexts.

Author Contributions: *Chappidi Suneetha* led the overall project coordination and manuscript supervision. *Tota Varshini*, *Pallem Santhoshi Rupa*, *Seetepalli Meghana*, *Vooda Eeshitha Vaishnavi*, and *Vadapalli Jahnavi* contributed to data collection, analysis, and drafting of the manuscript. All authors reviewed and approved the final version of the paper.

Originality and Ethical Standards: We confirm that this work is original and has not been published elsewhere, nor is it under consideration for publication elsewhere. All ethical standards, including proper citations and acknowledgements, were followed.

Data availability: Data are available upon request.

Conflict of Interest: There are no conflicts of interest to declare.

Funding: The research received no external funding.

Similarity checked: Yes

References

[1] T. Y. Wong et al., "Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality," Surv. Ophthalmol., vol. 46, no. 1, pp. 59–80, Jul. 2001. DOI: 10.1016/S0039-6257(01)00234-X.

[2] M. D. Knudtson et al., "Revised formulas for summarizing retinal vessel diameters," Curr. Eye Res., vol. 27, no. 3, pp. 143–149, Sep. 2003. DOI: 10.1076/ceyr.27.3.143.16049.

[3] R. Klein et al., "Retinal vessel caliber and long-term risk of coronary heart disease," JAMA, vol. 300, no. 4, pp. 411–419, Jul. 2008. DOI: 10.1001/jama.300.4.411.

[4] J. W. Yau et al., "Global prevalence and major risk factors of diabetic retinopathy," Diabetes Care, vol. 35, no. 3, pp. 556–564, Mar. 2012. DOI: 10.2337/dc11-1909.

[5] P. Porwal et al., "IDRiD: Diabetic retinopathy—Segmentation and grading challenge," Med. Image Anal., vol. 59, Oct. 2020. DOI: 10.1016/j.media.2019.101561.

[6] A. Poplin et al., "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," Nat. Biomed. Eng., vol. 2, no. 3, pp. 158–164, Mar. 2018. DOI: 10.1038/s41551-018-0195-0.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," Med. Image Comput. Comput.-Assist. Interv., pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[8] 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.

[9] S. Chappidi and A. Raju, "A survey of machine learning techniques on speech-based emotion recognition and post-traumatic stress disorder detection," NeuroQuantology, vol. 20, no. 14, pp. 69–79, Oct. 2022, doi: 10.4704/nq.2022.20.14.NQ88010.

[10] S. Chappidi and A. Raju, "Enhanced speech emotion recognition using the cognitive emotion fusion network for PTSD detection with a novel hybrid approach," Journal of Electrical Systems, doi: https://doi.org/10.52783/jes.644.

[11] S. Chappidi and A. Raju, "Advancements in speech-based emotion recognition and PTSD detection through machine and deep learning techniques: A comprehensive survey," SSRG International Journal of Electronics and Communication Engineering, vol. 11, no. 5, 2023, doi: 10.14445/23488549/IJECE-V11I5P121.

[12] S. Chappidi and A. Raju, "Speech-based emotion recognition by using a faster region-based convolutional neural network," Multimedia Tools and Applications, Springer, 2024, doi: https://doi.org/10.1007/s11042-024-19004-2.

[13] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT, vol. 1, pp. 4171–4186, 2019. arXiv:1810.04805.

[14] EyePACS, "Diabetic retinopathy detection dataset," 2015. : https://www.kaggle.com/c/diabetic-retinopathy-detection.

[15] Messidor, "Methods for evaluating segmentation and indexing techniques in the field of retinal ophthalmology," 2008. : http://www.adcis.net/en/Download-Third-Party/Messidor.html.

[16] UK Biobank, "Retinal imaging dataset," 2020. : https://www.ukbiobank.ac.uk/.

[17] R. R. Wolfe et al., "Standards for retinal imaging in clinical trials," Ophthalmology, vol. 121, no. 7, pp. 1453–1458, Jul. 2014. DOI: 10.1016/j.ophtha.2014.01.021.

[18] M. M. Fraz et al., "An ensemble classification-based approach for retinal vessel segmentation," IEEE Trans. Biomed. Eng., vol. 59, no. 9, pp. 2538–2548, Sep. 2012. DOI: 10.1109/TBME.2012.2205687.

[19] A. Hoover et al., "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Trans. Med. Imaging, vol. 19, no. 3, pp. 203–210, Mar. 2000. DOI: 10.1109/42.845178.

[20] P. M. Ridker et al., "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women," JAMA, vol. 297, no. 6, pp. 611–619, Feb. 2007. DOI: 10.1001/jama.297.6.611.

[21] H. W. Ressom et al., "Handling missing values in proteomic data," Proteomics, vol. 5, no. 8, pp. 2085–2097, May 2005. DOI: 10.1002/pmic.200401071.

[22] J. L. Fleiss et al., "The measurement of interrater agreement," Stat. Methods Rates Proportions, vol. 2, pp. 22–23, 1981.

[23] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," Med. Image Anal., vol. 53, pp. 26–42, May 2018. DOI: 10.1016/j.media.2019.01.012.

[24] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," Proc. ICLR, 2017. arXiv:1609.02907.

[25] J. Staal et al., "Ridge-based vessel segmentation in color images of the retina," IEEE Trans. Med. Imaging, vol. 23, no. 4, pp. 501–509, Apr. 2004. DOI: 10.1109/TMI.2004.825627.

[26] W. L. Hamilton et al., "Inductive representation learning on large graphs," Adv. Neural Inf. Process. Syst., vol. 30, pp. 1024–1034, 2017. arXiv:1706.02216.