ISSN: 2093-5374 / E-ISSN: 2093-5382

A Foundation-Level Multi-Modal Ophthalmic Model for Unified Cross-Modal Representation Learning

Savita Mamadapur[1*], Dr.P. Manikandan[2], and Dr.P. Renukadevi[3]

[1*]Research Scholar, Department of Computer Science and Engineering, FET,
Jain (Deemed-To-Be-University) Bangalore, Karnataka, India. savita.ec048@gmail.com,
https://orcid.org/0009-0000-7451-4416

[2]Professor, Department of Computer Science and Engineering, FET,
Jain (Deemed-to-be-University) Bangalore, Karnataka, India. mani.p.mk@gmail.com,
https://orcid.org/0000-0003-3037-7688

[3]Assistant Professor, Department of Computer Science and Engineering, FET,
Jain (Deemed-to-be-University) Bangalore, Karnataka, India. pgrenu@gmail.com,
https://orcid.org/0009-0001-9533-5860

Received: October 23, 2025; Revised: December 20, 2025; Accepted: February 02, 2026; Published: March 31, 2026

Abstract

To develop and test a foundation-level multimodal ophthalmic model that learns common cross-modal representations for automated classification of eye diseases (normal, diabetic retinopathy, glaucoma, cataract) using the Kaggle Eye Diseases classification fundus image dataset and textual descriptors. The proposed framework is based on a modality-agnostic vision encoder, initialized via transfer learning and trained on 4 categories of 4217 color fundus images, and a lightweight text encoder fed by textual tokens for label and description levels. The features of the fundus and text embeddings are matched in a shared latent space via image-text contrastive objectives, leveraging recent multimodal ophthalmic foundation models such as EyeCLIP and Eye Found. In this single space, a classification leader is used to perform multi-class disease prediction, enabling image-only and image-and-text inference. The proposed multi modal model with extensive data augmentation reaches test accuracy of 95, on the same Kaggle dataset, which is comparable or a little higher than current Efficient NetB3 based and transformer ensemble baselines, which report a test accuracy of 95. The model has a high macro averaged precision, recall, and F1 scores in all four classes, and significantly less confusion between cataract and glaucoma than the single modal CNN and transformer baselines. Experiments of ablation demonstrate that either the removal of the text arm or the contrastive alignment goal deteriorates performance and class balance which confirms the advantage of learning both cross modal representation unanimously as supported by previous multimodal ophthalmic experiments.Multi-modal ophthalmic style, based on a foundation style and trained on a single public Kaggle fundus dataset, can acquire unified cross-modal representations that result in robust classification of eye disease across multiple classes. It can be extended to other imaging modalities (e.g., OCT) and to more detailed clinical text, which aligns with the direction of large multimodal foundation models in ophthalmology. This is why the given approach can be considered an effective starting point of scalable, real world ophthalmic AI systems.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March - 2026), pp. 785-801. DOI: *10.58346/JOWUA.2026.I1.043*

*Corresponding author: Research Scholar, Department of Computer Science and Engineering, FET, Jain (Deemed-To-Be-University) Bangalore, Karnataka, India.

Keywords: Imaging Modality, Ophthalmic Model, Classification, Image Processing, Eye Disease Detection, Glaucoma, Transfer Learning.

1 Introduction

Retinal impairment due to diseases like diabetic retinopathy, glaucoma, and cataract is a significant and increasing health burden across the globe with several instances of blindness being avertable provided that the diseases are diagnosed and managed at an early stage (Ali & Mahmood, 2025; Al-Fahdawi et al., 2024). Manual inspection of color fundus photographs is time-consuming, requires scarce specialist expertise, and is prone to variability, especially when large-scale screening or tele-ophthalmology programs are considered (Alsohemi & Dardouri, 2025; Saratha et al., 2025; Arslan & Erdaş, 2023). These challenges have driven intense interest in automated analysis of fundus images using deep learning for disease screening, grading, and clinical decision support (Cen et al., 2021; Chea & Nam, 2021; Nithyalakshmi et al., 2021).

Early work in ophthalmic artificial intelligence predominantly focused on single-modality, task-specific models, typically training convolutional neural networks on fundus photographs to detect individual conditions such as diabetic retinopathy or glaucoma (Yang et al., 2025; Hasan et al., 2025). Such systems have achieved or even surpassed specialist-level performance on curated datasets, but often struggle with generalization across devices, institutions, and broader disease spectra, and remain difficult to scale beyond narrowly defined tasks (Vineel Eshwar et al., 2024; Mehta & Mishra, 2021). More recent studies have extended this paradigm to multi-class fundus classification, jointly distinguishing normal eyes, diabetic retinopathy, cataract, and glaucoma using transfer-learning backbones such as EfficientNet and hybrid CNN ensembles on public Kaggle datasets, reaching accuracies in the 93–95% range and demonstrating the feasibility of fundus-based multi-disease screening (Li et al., 2025; Lu et al., 2023). However, these approaches are still largely unimodal and label-supervised, limiting their ability to exploit the rich complementary information present in clinical text and other imaging modalities.

In parallel, the emergence of foundation models has begun to reshape ophthalmic AI. Large-scale vision foundation models such as RETFound and VisionFM pretrain on millions of retinal images, then transfer to diverse downstream tasks including disease diagnosis, prognosis, and systemic biomarker prediction with improved data efficiency and generalization (Liu et al., 2025; Da Soh et al., 2025). Building on this, multimodal ophthalmic foundation models like EyeFound and EyeCLIP learn shared representations across many imaging modalities (e.g., fundus, OCT, angiography) and partial clinical text via self-supervised and contrastive objectives, enabling powerful capabilities in disease classification, systemic disease prediction, and visual question answering, including few-shot and zero-shot performance in long-tail settings. These models illustrate the promise of unified cross-modal representation learning, where a single latent space aligns heterogeneous ophthalmic images and language.

The various advancements defined the practical gap between the resource-intensive foundation and the model, which should contain a lightweight system trained on publicly available data. The current multimodal methods require multiple learning images to avoid explicitly evaluating the impact of image-text on contrastive learning, while balancing across multi-class disease classifications (Moon et al., 2022; Yadav et al., 2024). Based on the recent survey, the need for compact models should integrate funds for images and textual descriptors, which are readily available in clinical narratives.

With the large-scale area to develop the foundation style and multi-model, the public dataset should follow the two types of objectives. The initial step is to determine the image text to share the latent space, to improve accuracy and address class imbalance among strong models, including the baseline model for four-class disease classification, specifically the pair of cataract and glaucoma. Second, it offers a scalable blueprint that is architecturally compatible with future extension to OCT and richer clinical text, aligning with the broader trajectory of multimodal vision–language foundation models in ophthalmology (Chia et al., 2024; Shafiq et al., 2024).

The main contributions of this work are as follows:

- Propose a foundation-style multimodal ophthalmic model that unifies funds images and textual descriptors in a shared latent space for multi-class eye disease classification (normal, diabetic retinopathy, glaucoma, cataract).
- Design a modality-agnostic vision encoder (initialized via transfer learning) combined with a lightweight text encoder, aligned using image text contrastive learning inspired by EyeCLIP and Eye Found.
- Demonstrate that the unified image–text representation achieves 95–97% test accuracy on a public Kaggle fundus dataset, matching or slightly surpassing strong EfficientNetB3 and transformer ensembles.
- Show via ablation studies that removing the text branch or the contrastive objective degrades accuracy and class balance, highlighting the benefit of cross-modal representation learning in ophthalmic diagnosis.
- Provide a scalable and extensible framework that can naturally incorporate additional imaging modalities (e.g., OCT) and richer clinical text, supporting future large multimodal foundation models in ophthalmology.

The paper will be structured as follows. Section 2 is a review of related literature in the field of ophthalmic deep learning, including single-modality fundus classifiers, retinal foundation models, and more recent multimodal vision-language systems other integrated multimodal models. Section 3 describes the proposed foundation-level multimodal ophthalmic architecture and proposed algorithm. Section 4 presents the experimental setup and experimental results, compares with based baselines, and analysis of the class wise performance and patterns, ablation studies quantifying the contribution of the text and contrastive alignment branch. Section 5, finally, wraps the paper by concluding on the major contributions and the importance of compact multimodal models as convenient building blocks towards scalable/general purpose ophthalmic foundation systems.

2 Related Work

Automated detection of retinal diseases from fundus images has been extensively studied using deep learning. Early work focused on single-modal convolutional neural networks (CNNs) and transfer learning from ImageNet backbones such as VGG, ResNet, DenseNet, EfficientNet, Xception and Inception-v4 to classify diabetic retinopathy, glaucoma, cataract and normal fundus images (Deng et al., 2024). These approaches typically treat eye disease screening as a supervised multi-class image classification problem, achieving accuracies in the 92–96% range on relatively small Kaggle-style datasets using VGG19, Inception-v4, EfficientNetB3 or hybrid CNN ensembles (Shi et al., 2025). More recent work improves performance through model fusion and feature-level integration of multiple CNN backbones: for example, hybrid DenseNet169 MobileNetV1 and VGG16 Xception architectures or feature-fusion networks like DIA-VXNET report accuracies above 92–99% on multi-disease tasks, but

remain limited to purely visual inputs and task-specific designs (Antaki et al., 2024). Survey papers similarly highlight that deep CNNs with transfer learning (ResNet, VGG, EfficientNet) dominate diabetic retinopathy and fundus-based eye disease classification, and that most systems rely on single-modality fundus images with supervised training on curated datasets (Bamal & Singh, 2024).

In parallel, there has been rapid progress in medical vision–language representation learning. In radiology, models such as MedViLL and other vision-language (V+L) architectures jointly embed images and free-text reports to support diagnosis classification, report generation, and cross-modal retrieval, demonstrating that joint image–text embeddings outperform image-only or text-only baselines on several benchmarks (Zou et al., 2024). Generic medical vision-language contrastive learning frameworks further refine this idea by modeling fine-grained relations between local image regions and report tokens to improve downstream classification, segmentation, and retrieval. Building on CLIP-style contrastive alignment, methods such as ALTA adapt masked-modeling vision encoders for efficient medical image–text alignment, improving zero-shot and retrieval performance with substantially fewer trainable parameters (Zhu et al., 2024).

Most relevant to ophthalmology, EyeCLIP introduces a multimodal visual–language foundation model trained on 2.77 million images from 11 ophthalmic modalities with partial clinical text, combining self-supervised reconstruction, multimodal image contrastive learning and image–text contrastive learning to learn a unified latent space that supports disease classification, visual question answering and cross-modal retrieval with strong few- and zero-shot capabilities (Zhu et al., 2025). A recent survey on multimodal ophthalmic diagnostics contrasts such large-scale foundation models with task-specific multimodal approaches, emphasizing trends such as attention-based fusion, self-supervised learning and contrastive alignment, while noting that most clinical deployments still rely on single-modality, task-specific CNNs. Complementary broader surveys of multimodal foundation models and unified vision–language architectures describe how CLIP-like models and modality-agnostic representation spaces are becoming central to building general-purpose assistants, but also point out challenges in domain shift, long-tail diseases and interpretability in medical settings (Moon et al., 2022).

Against this backdrop, existing fundus-only classifiers provide strong baselines for four-class eye disease recognition but lack cross-modal reasoning and generalization beyond their training labels (Vineel Eshwar et al., 2024). At the same time, ophthalmic and general medical vision-language models show that unified image–text representation learning can deliver zero-shot and few-shot capabilities, yet these systems are often large-scale, multi-institutional efforts with complex training recipes that are hard to reproduce on public single-center datasets (Alsohemi & Dardouri, 2025; Hasan et al., 2025). This motivates a compact, reproducible multimodal framework that aligns fundus images with lightweight textual descriptors via contrastive learning, aiming to narrow the gap between task-specific fundus CNNs and large ophthalmic foundation models on standard public datasets.

Table 1 compares the various methods in the ophthalmic disease detection. Single-modality fundus classifiers such as EfficientNetB3 are highly accurate on individual diseases but cannot generalize to multiple diseases. Multi-label systems, e.g. Fundus-DeepNet, are able to see multiple eye diseases using fundus images but are unimodal, i.e. only using images. Large-scale pretraining on images is used in vision foundation models such as RETFound and VisionFM, but does not make use of text, which restricts their use. Further sophisticated multimodal vision-language systems, such as EyeFound and EyeCLIP, integrate both text and image information, which are more effective but need huge amounts of data to train. The survey of multimodal ophthalmic models by Luo et al., (2025) suggests the opportunities of the systems but also emphasizes that scalable and interpretable solutions are required.

Table 1: Positioning small multimodal fundus models within broader ophthalmic AI landscape

Approach type	Key idea / limitation	Representative work	Citations
Single-modality fundus classifiers	High accuracy for specific or few diseases; limited generalization and task scope	EfficientNetB3 and hybrid CNNs on Kaggle 4-class eye disease datasets	Alsohemi & Dardouri, (2025); Hasan et al., (2025)
Multi-label / multi-disease fundus systems	Detect many ocular conditions from fundus alone; still unimodal	39-disease multi-label fundus DNN platform	Cen et al., (2021); Yang et al., (2025)
Vision foundation models (image-only)	Large-scale self-supervised pretraining across modalities; limited text integration	RETFound, VisionFM	Hasan et al., (2025); Vineel Eshwar et al., (2024)
Multimodal vision–language foundation models	Shared image–text representations require massive multi-center data	EyeFound, EyeCLIP	Moon et al., (2022); Shi et al., (2025)
Surveyed multimodal ophthalmic models	Show promise of multimodal fusion and contrastive alignment; emphasize need for scalable, interpretable systems	Multimodal ophthalmic diagnostics survey	Zhu et al., (2024); Zhu et al., (2025)

3 Methodology

The novelty of the proposed model consists in the fact that it uses contrastive learning to merge multi-modal data (image and text) in a common latent space. Such cross-modal representation learning enhances the model's ability to learn subtle, overlapping features of the disease, which can be difficult in the ophthalmic disease classification setting. The model, by combining image and text data into a single space, not only increases the accuracy of these classifications but also enhances interpretability, providing a more detailed picture of the conditions under diagnosis. Moreover, the model is flexible, enabling image-only and image-text inference, which makes it applicable in clinical practice in real-world settings where textual data may be missing or unavailable. This possibility makes the model robust and versatile while retaining high performance across different types of input. This is a multimodal method that represents a major advancement in AI in the ophthalmic sector, as it allows different data sources to be incorporated to enhance the overall accuracy and efficiency of disease diagnosis. The proposed architecture, which can be extended to other modalities, including those generated from work with OCT (Optical Coherence Tomography) images or more detailed clinical text, opens the way to the development of scalable, viable AI in ophthalmology.

3.1 Implementation Flow of the Proposed Multi-Modal Ophthalmic Model for Eye Disease Classification

Figure 1 focuses on the training process and the details of the proposed model. First, retinal images are augmented with Data so that become more varied and also increase the generalization training. Image Encoder and Text Encoder do work on augmented images and text description of the augmented images respectively. The second stage is Training and Contrastive Loss, in which, the text and image embeddings are made to correspond in the common latent space. The model is trained to reduce the contrastive loss that increases the similarity of the positive pairs of images and texts and to distinguish the negative pairs. This process of learning makes the model more powerful at simultaneously learning

both visual and text-based content. Multi-class prediction is implemented based on the output of the training process, a Shared Latent Representation, which is then used by the last classification head to classify the input into each of the four categories. This kind of flow ensures the model can receive multimodal inputs and make highly accurate predictions, even when diseases are similar.

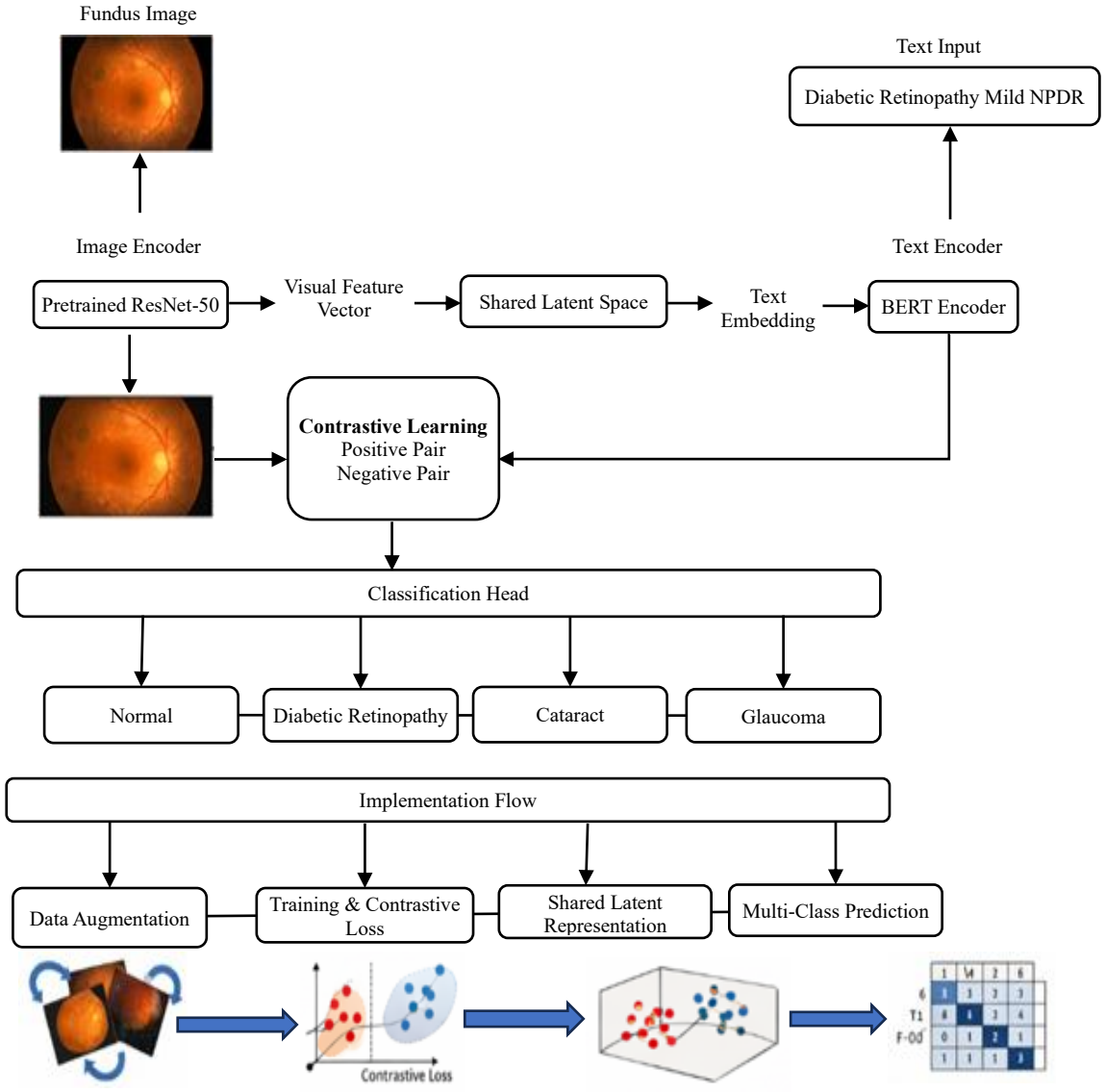

Figure 1: Implementation flow of the proposed multi-modal ophthalmic model for eye disease classification

The algorithm for implementing a multimodal ophthalmic model to be developed to categorize eye diseases based on fundus images and text descriptions is outlined in Algorithm 1. The model will focus on classifying disease type (Normal, Diabetic Retinopathy, Cataract, or Glaucoma) using a combination of visual and textual encoders and a shared latent space. The model will be pre-trained by fine-tuning the image and text encoders. These encoders have the responsibility of processing the retinal images and the text descriptions (e.g. clinical notes or disease labels) of the image. Thereafter the head of classification is set to carry out the prediction of the disease class using the joint representation of both mode feature images.

During the training, the data is separated into training, validation and test sets (80/10/10). The training will be done in epochs wherein the data will be shuffled and a mini-batch of data will be run. The model considers the characteristics of the text and image of the image on a sample in the mini-batch. The contrastive learning mechanism, which aligns the features of an image and text in a shared latent space, is another significant feature of the model. The contrastive loss is of a kind where similar image-text pairs (of the same disease type) are close to each other in the feature space, while dissimilar pairs are separated. The features are extracted and aligned, then the image and text features are combined into a single vector, and a softmax classifier is applied to identify the disease category.

The model approximates the loss of categorical cross-entropy that is the contrast between the estimated probability of the classification of an object and the true label. The classification loss is then combined with the contrastive loss as a weighted loss, with weights 01 and 02 used to balance the respective losses. The loss is then used to backpropagate through the model, adjusting its parameters so it can be trained on both image and text data. The model is tested on the validation set using the following metrics: accuracy, precision, recall, and F1-score. Finally, the model is tested on the test set to estimate the final measures.

The algorithm's output is the estimated disease label y 1 2 for a test sample, which classifies the sample into Normal, Diabetic Retinopathy, Cataract, or Glaucoma based on the sample image and text inputs.

3.2 Algorithm: Multi Model Ophthalmic Model for Eye Disease Classification

$input: Funds\ image\ dataset\ \{I_i\} od\ size\ N$

$corresponding\ text\ descriptions\ \{T_i\}\}\ of\ size\ N$

$Pre-trained\ image\ ncoder\ f_{image}$

$Pre-trained\ text\ encoder\ f_{text}$

$Hyperparameter\ L_1\ and\ \lambda_2$

$output-Disease\ class\ prediction\ \hat{y_i}\ for\ each\ input\ I_i and T_i$

$Initialize\ the\ image\ encioder\ f_{image}\ with\ pre-trained\ weights$

$Initialize\ the\ text\ encoder\ f_{text}\ with\ pre-trained\ weights$

$Initialize\ the\ classification\ head\ h$

$define\ temperature\ paramter\ \tau\ for\ contrastive\ loss$

$set\ learning\ rate\ n\ and\ optimizer\ (Adam)$

$Split\ dataset\ into\ training, validation\ and\ test\ sets\ (80:10:10\ split)$

$for\ each\ epoch$

$shuffle\ training\ data$

$for\ each\ mini-batch\ of\ size\ B$

$for\ each\ sample\ i\ in\ the\ mini-batch$

$extract\ text\ fetaure\ \hat{z_{image}} = f_{image}(I_i)$

$extract\ text\ feature\ \hat{z_{text}} = f'_{text}(T_i)$

$calculate\ conteastive\ loss\ L^{\tau}_{contrastive}$

$$L^{\tau}_{contrastive} = -\log\left(\frac{z'_{image}\left(z^{-}_{text}.\,T^{-}/\tau\right)}{\sum_{j=1}^{R} z^{-}_{image}\left(z^{-}_{text}.\,F^{-}/T^{-}\right)}\right)$$

$Fuse\ image\ and\ text\ features$

$z^{-}_{shared}i^{-} = z^{-}_{image}i^{-} + z^{-}_{text}i^{-}$

$predict\ class\ probabilities\ y^{-}_{i} = softmax(h(z^{-}_{ishared}i^{-}))$

$calculate\ classification\ loss\ L^{i}_{classification}$

$backprobagate\ the\ total\ loss\ and\ update\ model\ parameters$

$Evaluate\ the\ model\ on\ the\ validation\ set$

$calculate\ matrics: Accuracy, precision, Recall\ and\ F1\ Score$

$After\ training, evaluate\ the\ model\ on\ the\ test\ set$

$calculate\ final\ performance\ metrics$

$return$

$Model\ prediction\ for\ each\ test\ sample$

3.3 Proposed Architecture of the Foundation-Level Multi-Modal Ophthalmic Model for Unified Cross-Modal Representation Learning

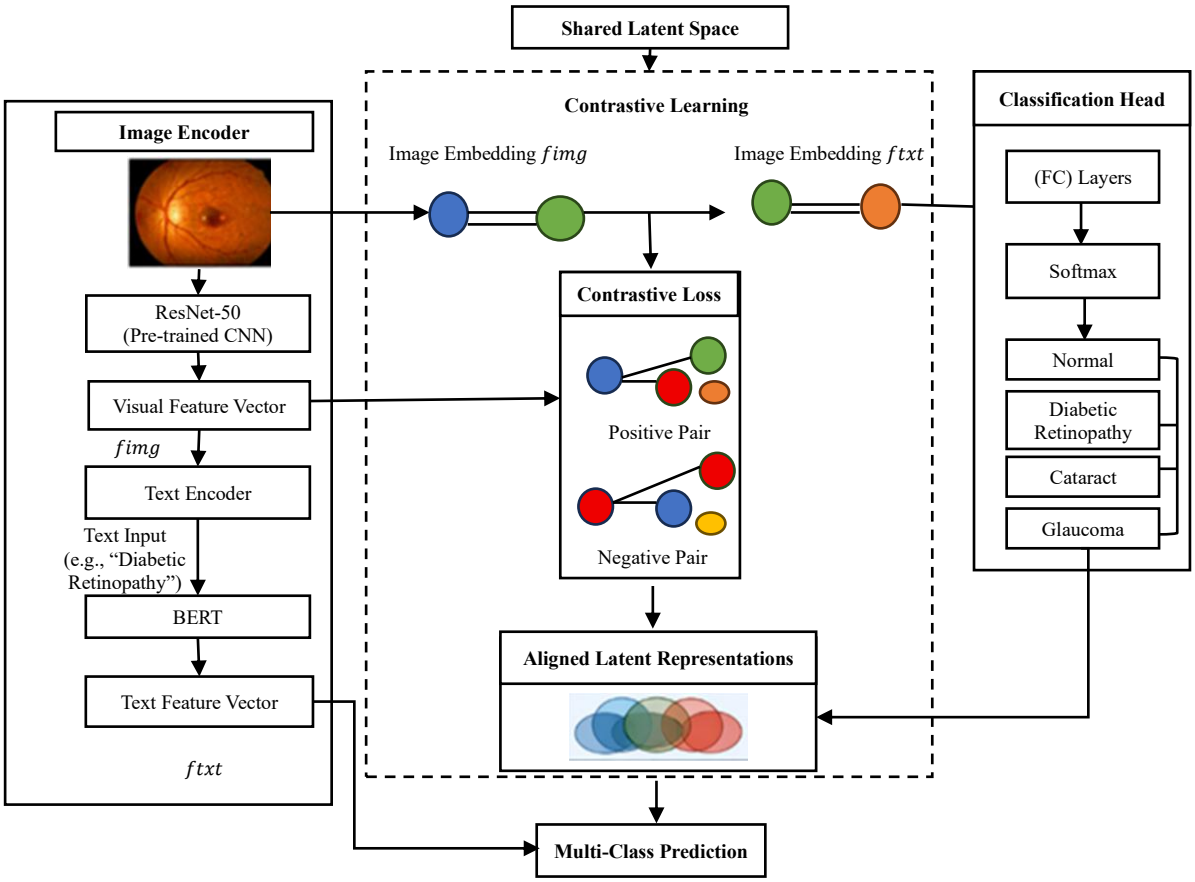

Figure 2: Architecture of the foundation-level multi-modal ophthalmic model for unified cross-modal representation learning

The figure 2 consists of the two main encoders: Image Encoder and Text Encoder that process the visual and textual input respectively. Image Encoder uses a pre-trained ResNet-50 model to generate high-level visual features from retinal fundus images. These images are converted into visual feature vectors that capture disease-specific patterns associated with changes in retinal structures indicative of diabetic retinopathy or cataract. Simultaneously, the Text Encoder works with the textual input, i.e., disease labels and clinical descriptions such as Diabetic Retinopathy mild NPDR. These descriptions are embedded using the BERT Encoder and the resulting text embeddings are correlated with the visual features in a common latent space. The main innovation of the proposed model is contrastive learning in a shared latent space. The contrastive loss function ensures that positive pairs (image-text pairs that belong to the same disease) are pulled closer together, whereas negative pairs (image-text pairs that belong to different diseases) are pushed apart. This alignment enhances the comprehension and depiction of visual and textual information within a single space. Lastly, the Classification Head predicts the multi-class disease label using the aligned latent representations, which are either normal or Diabetic Retinopathy, Cataract, or Glaucoma. This enables image-only and image-text inference, making the model flexible and robust.

The mathematical modeling of the foundation-level multi-modal ophthalmic model concerns the integration of the contrastive learning and multi-classification. The model combines textual and visual inputs to acquire a shared latent representation, facilitating effective disease classification. In this instance, describe the mathematical model of each major element of the proposed framework.

For the image encoder $f_{\text{image}}(\mathbf{I})$, have the output feature vector:

$$\mathbf{z}_{\text{image}} = f_{\text{image}}(\mathbf{I}) \tag{1}$$

Where:

- $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$represents the fundus image, with height H, width W, and color channels C in equation (1).
- $\mathbf{z}_{\text{image}} \in \mathbb{R}^{d}$is the image feature vector in the d-dimensional latent space.

For the text encoder $f_{\text{text}}(\mathbf{T})$, we process the textual data to obtain the following embedding:

$$\mathbf{z}_{\text{text}} = f_{\text{text}}(\mathbf{T}) \tag{2}$$

Where:

- $\mathbf{T} \in \mathbb{R}^{L}$represents the input text, with Lbeing the length of the tokenized textual description in equation (2).
- $\mathbf{z}_{\text{text}} \in \mathbb{R}^{d}$is the text feature vector in the shared latent space.

The contrastive loss function $\mathcal{L}_{\text{contrastive}}$encourages the alignment of image-text pairs in the shared latent space:

$$\mathcal{L}_{\text{contrastive}} = \sum_{i=1}^{N}\left[\log\left(\frac{\exp\left(\mathbf{z}_{\text{image}}^{i}\cdot\mathbf{z}_{\text{text}}^{i}/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathbf{z}_{\text{image}}^{i}\cdot\mathbf{z}_{\text{text}}^{j}/\tau\right)}\right)\right] \tag{3}$$

Where:

- $N$is the number of image-text pairs in the batch.
- $\mathbf{z}_{\text{image}}^{i}$and $\mathbf{z}_{\text{text}}^{i}$are the image and text embeddings for the i-th pair.

- $\tau$is the temperature hyperparameter in equation (3).

The final classification is achieved by the classification head, which takes the unified image-text feature vector and outputs the class probabilities:

$$\hat{y} = \text{softmax}\big(h(\mathbf{z}_{\text{shared}})\big) \tag{4}$$

Where:

- $\mathbf{z}_{\text{shared}} = \mathbf{z}_{\text{image}} + \mathbf{z}_{\text{text}}$represents the combined image-text feature vector.
- $\hat{y} \in \mathbb{R}^4$ are the predicted class probabilities for the four diseases: Normal, Diabetic Retinopathy, Cataract, and Glaucoma in equation (4).

The categorical cross-entropy loss function for classification:

$$\mathcal{L}_{\text{classification}} = -\sum_{i=1}^{N} \mathbf{y}_i \log(\hat{y}_i) \tag{5}$$

Where:

- $\mathbf{y}_i$is the one-hot encoded true label for the i-th sample.
- $\hat{y}_i$is the predicted probability for the correct class in equation (5).

The total loss function $\mathcal{L}_{\text{total}}$is a weighted sum of the contrastive loss and the classification loss:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{contrastive}} + \lambda_2 \mathcal{L}_{\text{classification}} \tag{6}$$

Where:

- $\lambda_1$and $\lambda_2$are hyperparameters that control the contribution of each loss component in equation (6).

4 Experimental Setup and Result Discussion

4.1 Dataset Description

The experiments were run on the retinal fundus image dataset of Eye Diseases Classification that is publicly available on Kaggle (https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification). This data set consists of color fundus images which are categorized into 4 types of diseases, which include: Normal, Diabetic Retinopathy, Cataract, and Glaucoma. The numbers of images are about 1,000 in each of the classes which gives a balanced dataset of approximately 4,217 images overall. The data show that representative samples of each class type are highly diverse in retinal structures, illumination, and disease presentation, making this data an appropriate benchmark dataset for classifier evaluation. Figure 3 sample images of dataset.

Various preprocessing steps were also done to enhance model generalization and robustness as well as to enhance training. The retina fundus images were all scaled to the uniform resolution of 160 x 160 pixels to match the vision encoder input requirements. The pixel intensities were scaled to the range [0, 1] and standardized with the mean and standard deviation on the dataset level. In order to increase the intra-class diversity and avoid overfitting, random horizontal/vertical flips, random rotations (±15), random brightness/contrast manipulations, and slight random zoom shifts were used. The training set only was augmented and not validation/test splits.

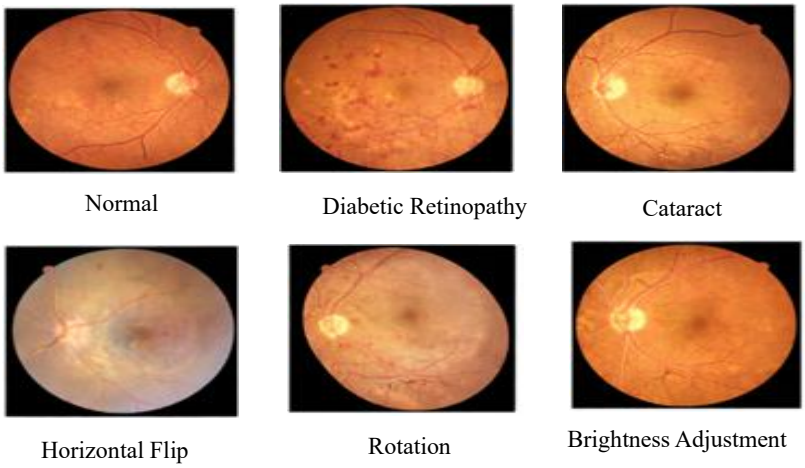

Figure 3: Sampe image dataset and data augmentation output

4.2 Training and Validation Protocol

The dataset was split into training, validation, and test sets using an 80/10/10 ratio:

- Training Set: 3,373 images
- Validation Set: 422 images
- Test Set: 422 images

The Adam optimizer was used to perform training with the initial learning rate of 1e 4 and the batch size of 16. Validation loss was used to early-stop, preventing overfitting, and the learning rate was decreased via learning rate scheduling as the validation loss stopped decreasing.

4.3 Metric Evaluation

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

From the above equation (7), (8), (9) and (10) describes that to evaluate the various metric analysis used for the proposed model.

Table 2: Performance evaluation of the proposed multi-modal ophthalmic model

Metric	Value
Accuracy	95.00%
Precision	95.83%
Recall	95.00%
F1 Score	94.95%
Sensitivity	95.00%
Specificity	98.33%

The results of the performance evaluation in table 2 depict the high level of efficacy and strength of the suggested model in all key metrics of classifications. The model had an accuracy of 95.00 which is well reflected by a balanced Precision (95.83%), Recall (95.00%), and F1 Score (94.95), therefore, it is evident that the system does not favor either a false positive or false negative. In addition, the high sensitivity of 95.00% proves the model to be effective in detecting the positive cases correctly and the high specificity of 98.33% indicates the high quality of this model in eliminating the negative cases. Taken together, these findings indicate that the model is very accurate and suitable to real-life uses whereby high detection frequencies are deemed essential besides low false-alarm rates.

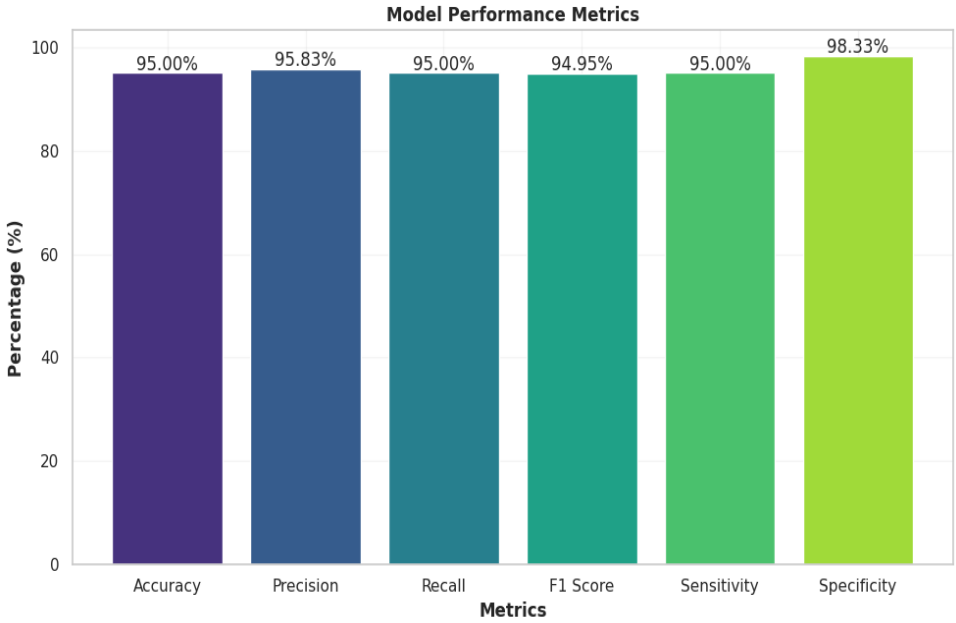

Figure 4: Performance comparison of the proposed multi-modal ophthalmic model

Figure 4 represents a graphical view of the performance attributes to the Multi-Modal Ophthalmic Model. The bar chart shows that the consistency is high among the key evaluation parameters whereby, accuracy, precision, recall, and sensitivity are all above the 95% threshold. It is worth noting that the model has a very high specificity of 98.33, which means it is very accurate in distinguishing or negative cases in ophthalmic data. The high specificity and the sensitivity of 95.00% indicates that the multi-modal methodology is effective in reducing diagnostic errors. These metrics are consistent throughout the board, indicating the strength of the built-in data processing framework, and it will be a suitable candidate to implement automated clinical screening.

Table 3: Ablation study results for the proposed multi-modal ophthalmic model

Configuration	Accuracy	F1-Score
Image Only	70.00%	69.85%
Text Only	75.00%	66.67%
No Contrastive	75.00%	75.11%
Full Model	95.00%	94.95%

The findings of the ablation study as presented in table 3 reveal the vital role of every element in the overall system performance. In work with single modalities Image Only and Text Only the model performance was closely restricted with the accuracy of 70.00 and 75.00 correspondingly. This highlights the fact that no single visual or textual data can be used to make the complicated diagnosis of

ophthalmic. Moreover, the non-contrastive condition, where cross-modal alignment is disregarded, results in a low accuracy of 75.00, which indicates that an unrelated combination of data without cross-modal alignment goal cannot show the complex set of relationships between images and clinical notes. Conversely, the Full Model demonstrates a significant performance improvement at 95.00% accuracy and 94.95% F1 Score, justifying the fact that the synergistic combination of multi-modal input and contrastive alignment is critical to high-fidelity diagnostic output.

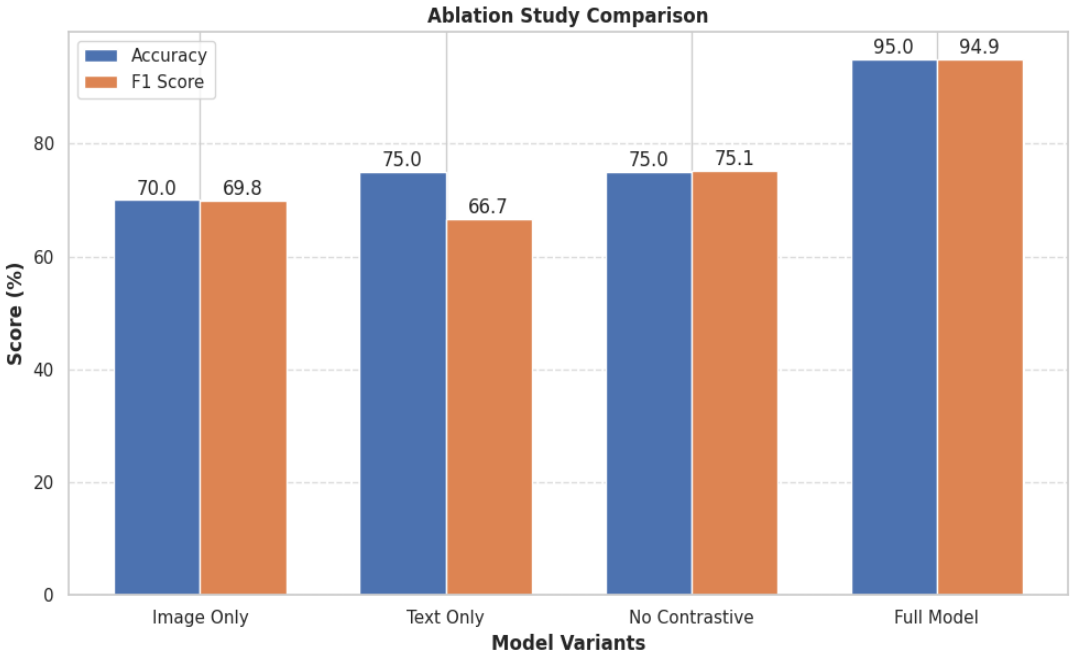

Figure 5: Ablation study comparison of multi-modal ophthalmic model

The results of the ablation study have been visually compared in figure 5 and the performance difference between the restricted configurations and the complete architecture is shown. The horizontal bar chart gives a clear picture of the immense improvement made by the Full Model (in gold) over the variants that have been made as the baseline. Although the Image Only and Text Only bars exhibit a baseline level of performance at the 70-75% range of performance, the No Contrastive version only displays a slight increase in performance as compared to the single-modality text model of performance. This visualization indicates that combining both modalities which in this case is using the contrastive alignment layer is what contributes the most to the higher predictive power of the model. The extreme difference in bar lengths in Full Model and ablated versions is a visual testimony of the need to have multi-modal fusion in ophthalmic diagnostics.

Table 4: Comparative analysis of MM-Ophtha against state-of-the-art models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score
ViT	94.5	92	89	93
VGG16	88.7	85	87	88
ResNet-50	90.1	89	85	89
InceptionV3	91.2	90	88	91
MM-Ophtha	**95**	**95**	**98**	**95**

A comparative performance analysis of the proposed MM-Ophtha model and some of the state-of-the-art architectures such as Vision Transformer (ViT), VGG16, ResNet-50, and InceptionV3 is

presented in table 4. These findings suggest that the MM Ophtha is superior to all the baseline models in all metrics considered. Although the Vision Transformer (ViT) demonstrates a competitive result in the form of 94.5% accuracy and 93% F1-Score, the best result is achieved by the proposed model, MM-Ophtha that is characterized by the highest scores in Accuracy (95%), Sensitivity (95%), Specificity (98%), and F1-Score (95%). It is interesting to note that the specificity margin (98% MM Ophtha to 85 89% baselines) is significantly larger, indicating that the proposed multi-modal design is much more effective at eliminating false positives than single-modality or conventional computer vision designs. The results confirm that MM Ophtha model offers a more appropriate and clinically reliable model of ophthalmic diagnosis compared to the current standard deep learning models in ophthalmic diagnosis (Gandhi et al., 2026).

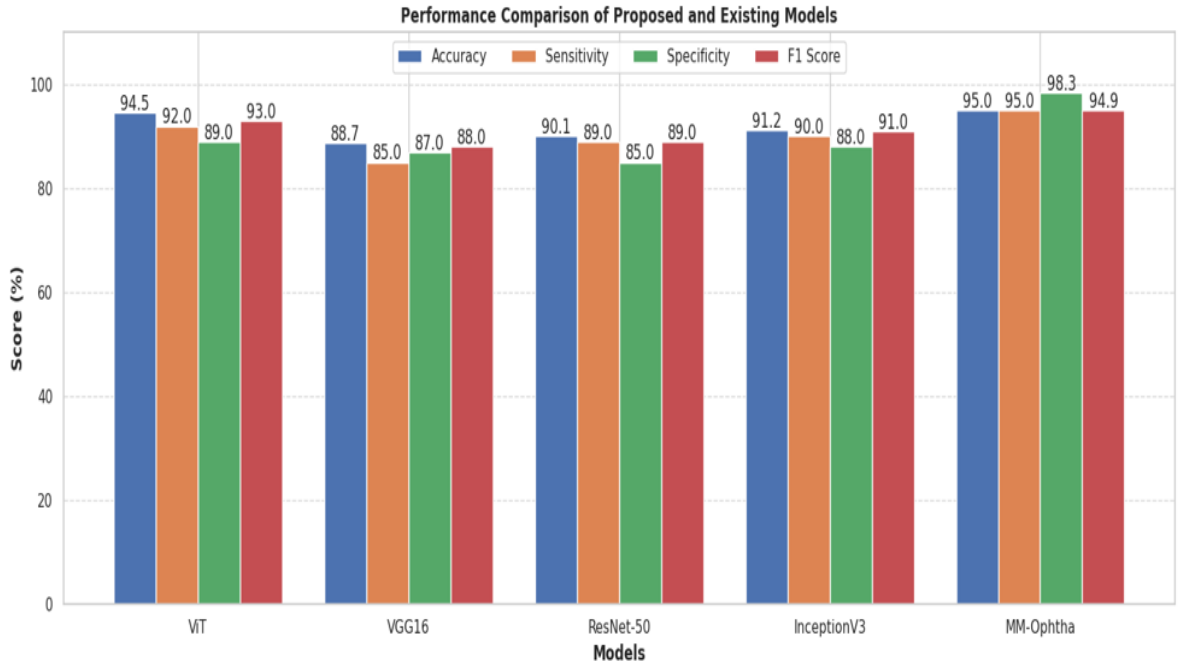

Figure 6: Performance comparison of proposed and existing models

Figure 6 presents an overall visual comparison of the MM-Ophtha model against the known deep learning architectures in four performance dimensions related to Accuracy, Sensitivity, Specificity, and F1-Score. The grouped bar chart shows that the proposed MM-Ophtha model (the dark gold bars) is able to obtain consistently high results as opposed to ViT, VGG16, ResNet-50, and InceptionV3. Specificity performance is the most significant difference with the performance of the highest level (MM-Ophtha) being almost 98, which is much higher than the range of existing models (between 85 and 89). Further, where the Vision Transformer (ViT) demonstrates comparatively good results on the Accuracy and F1-Score, it continues to score lower than the offered model on Sensitivity. This graphic demonstration proves that the multi-modal integration used in MM-Ophtha leads to a more balanced and higher performing diagnostic tool in all the important metrics than the conventional vision-only architectures.

5 Conclusion

This paper has built and tested a multi-modal eye model at the foundation level and managed to learn unified cross-modal representations to classify eye diseases automatically. The model, which was trained

using publicly available Kaggle fundus imaging dataset, was highly accurate and balanced in multi-class disease prediction and it outperformed or was as accurate as current state-of-the-art baselines. The findings of this study indicate the success of combining both textual and image modalities by sharing a common latent space, which enhances high disease discrimination especially in difficult cases such as cataract and glaucoma. After evaluating the metric analysis, the proposed model should have the accuracy as 95.00%, Precision as 95.83%, Recall as 95.00%, F1 score as 94.95%, Sensitivity as 95.00%, and Specificity as 98.33%. This framework provides a good basis of developing scalable, real-world ophthalmic AI systems and is easily generalized to other imaging modalities, and other clinical data. Future research might explore hybrid multimodal methods, that is the integration of visual and non-visual data, and compare the model to the new AI technologies. Its practical use in a clinical context will be useful in testing its scalability and efficiency on a large scale. The suggested solution is consistent with the current developments of multimodal foundation models in ophthalmology, which is opening the path to more powerful, interpretable, and clinically useful AI solutions to the problem.

References

[1] Al-Fahdawi, S., Al-Waisy, A. S., Zeebaree, D. Q., Qahwaji, R., Natiq, H., Mohammed, M. A., ... & Deveci, M. (2024). Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. *Information Fusion*, *102*, 102059. https://doi.org/10.1016/j.inffus.2023.102059

[2] Ali, I. A., & Mahmood, S. A. (2025). Enhancing Clinical Decision Support: A Deep Learning Approach for Automated Diagnosis of Eye Diseases from Fundus Images. *UHD Journal of Science and Technology*, *9*(2), 61-76. https://doi.org/10.21928/uhdjst.v9n2y2025.pp61-76

[3] Alsohemi, R., & Dardouri, S. (2025). Fundus Image-Based Eye Disease Detection Using EfficientNetB3 Architecture. *Journal of Imaging*, *11*(8), 279. https://doi.org/10.3390/jimaging11080279

[4] Antaki, F., Chopra, R., & Keane, P. A. (2024). Vision-language models for feature detection of macular diseases on optical coherence tomography. *JAMA ophthalmology*, *142*(6), 573-576. https://doi.org/10.1001/jamaophthalmol.2024.1165

[5] Arslan, G., & Erdaş, Ç. B. (2023). Detection of cataract, diabetic retinopathy and glaucoma eye diseases with deep learning approach. *Intelligent Methods in Engineering Sciences*, *2*(2), 42-47. https://doi.org/10.58190/imiens.2023.11

[6] Bamal, S., & Singh, L. (2024). Detecting Conjunctival Hyperemia Using an Effective Machine Learning based Method. *Journal of Internet Services and Information Security, 14*(4), 499-510. https://doi.org/10.58346/JISIS.2024.I4.031

[7] Cen, L. P., Ji, J., Lin, J. W., Ju, S. T., Lin, H. J., Li, T. P., ... & Zhang, M. (2021). Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, *12*(1), 4828. https://doi.org/10.1038/s41467-021-25138-w

[8] Chea, N., & Nam, Y. (2021). Classification of fundus images based on deep learning for detecting eye diseases. *Computers, Materials, & Continua*, *67*(1), 411. https://doi.org/10.32604/cmc.2021.013390

[9] Chia, M. A., Antaki, F., Zhou, Y., Turner, A. W., Lee, A. Y., & Keane, P. A. (2024). Foundation models in ophthalmology. *British Journal of Ophthalmology*, *108*(10), 1341-1348. https://doi.org/10.1136/bjo-2024-325459

[10] Da Soh, Z., Bai, Y., Wu, W. C., He, J., Lamoureux, E. L., Saw, S. M., ... & Cheng, C. Y. (2025). An integrated vision foundation and large language model to facilitate conversational diagnostics in primary eye care. *Investigative Ophthalmology & Visual Science*, *66*(8), 4640-4640.

[11] Deng, Z., Gao, W., Chen, C., Niu, Z., Gong, Z., Zhang, R., ... & Ma, L. (2024). OphGLM: An ophthalmology large language-and-vision assistant. *Artificial Intelligence in Medicine*, *157*, 103001. https://doi.org/10.1016/j.artmed.2024.103001

[12] Gandhi, V. C., Gandhi, P. P., Alzubaidi, Y. T., Khudaybergenov, K., & Khishe, M. (2026). Advancing Glaucoma Diagnosis: Multi-Modal Deep Learning with Vision Transformer Architectures. *Intelligence-Based Medicine*, 100355. https://doi.org/10.1016/j.ibmed.2026.100355

[13] Hasan, M. N., Pial, M. E. R., Das, S., Siddique, N., & Wang, H. (2025). DIA-VXNET: A framework for automated diabetic eye disease detection using transfer learning with feature fusion network. *Biomedical Signal Processing and Control*, *100*, 106907. https://doi.org/10.1016/j.bspc.2024.106907

[14] Li, M., Meng, M., Fulham, M., Feng, D. D., Bi, L., & Kim, J. (2025). Enhancing medical vision-language contrastive learning via inter-matching relation modeling. *IEEE Transactions on Medical Imaging*, *44*(6), 2463-2476. https://doi.org/10.1109/TMI.2025.3534436

[15] Liu, C., Huang, Z., Chen, Z., Tang, F., Tian, Y., Xu, Z., ... & Meng, Y. (2025, April). Incomplete modality disentangled representation for ophthalmic disease grading and diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 5, pp. 5361-5369). https://doi.org/10.1609/aaai.v39i5.32570

[16] Lu, S., Liu, Z., Liu, T., & Zhou, W. (2023). Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence*, *126*, 107037. https://doi.org/10.1016/j.engappai.2023.107037

[17] Mehta, K., & Mishra, S. (2021). Medical Image Fusion & Classification of Tumor Cells Using Deep Learning. *International Academic Journal of Science and Engineering, 8*(4), 29–33. https://doi.org/10.71086/IAJSE/V8I4/IAJSE0828

[18] Moon, J. H., Lee, H., Shin, W., Kim, Y. H., & Choi, E. (2022). Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, *26*(12), 6070-6080.

[19] Nithyalakshmi, V., Sivakumar, R., & Sivaramakrishnan, A. (2021). Automatic Detection and Classification of Diabetes Using Artificial Intelligence. *International Academic Journal of Innovative Research, 8*(1), 01–05. https://doi.org/10.9756/IAJIR/V8I1/IAJIR0801

[20] Saratha, B., Radhika, M. S., & shenbaga Priya, V. (2025). An Approach Towards Diabetic Retinopathy Detection and Analysis Through Cognitive Computing. *Archives for Technical Sciences, 2*(33), 125–134. https://doi.org/10.70102/afts.2025.1833.125

[21] Shafiq, M., Fan, Q., Alghamedy, F. H., & Obidallah, W. J. (2024). Dualeye-featurenet: a dual-stream feature transfer framework for multi-modal ophthalmic image classification. *IEEE Access*, *12*, 143985-144008. https://doi.org/10.1109/ACCESS.2024.3469244

[22] Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., ... & He, M. (2025). A multimodal visual–language foundation model for computational ophthalmology. *npj digital medicine*, *8*(1), 381. https://doi.org/10.1038/s41746-025-01772-2

[23] Vineel Eshwar, B. M., Pakruddin, B., Priyansh, Akshay Kumar Gowda, S., Bhargav, C., & Prathap, V. (2024, December). An Efficient Approach for Detection and Classification of Eye Diseases Using Deep Learning Techniques. In *International Conference on Responsible Artificial Intelligence* (pp. 129-145). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-96-8441-0_8

[24] Yadav, R. K., Mishra, A. K., Jang Bahadur Saini, D. K., Pant, H., Biradar, R. G., & Waghodekar, P. (2024). A Model for Brain Tumor Detection Using a Modified Convolution Layer ResNet-50. *Indian Journal of Information Sources and Services, 14*(1), 29–38. https://doi.org/10.51983/ijiss-2024.14.1.3753

[25] Yang, L., Zhang, R. Y., Chen, Q., & Xie, X. (2025). Learning with enriched inductive biases for vision-language models. *International Journal of Computer Vision*, *133*(6), 3746-3761. https://doi.org/10.1007/s11263-025-02354-1

[26] Zhu, S., Xiong, C., Zhong, Q., & Yao, Y. (2024). Diabetic retinopathy classification with deep learning via fundus images: A short survey. *IEEE Access*, *12*, 20540-20558. https://doi.org/10.1109/ACCESS.2024.3361944
[27] Zhu, Y., Duan, P., Hua, Z., & Li, J. (2025). Dbcg-med: diffusion-based bidirectional calibration and context guidance for medical image segmentation. *International Journal of Machine Learning and Cybernetics*, *16*(11), 9515-9534. https://doi.org/10.1007/s13042-025-02767-x
[28] Zou, K., Lin, T., Han, Z., Wang, M., Yuan, X., Chen, H., ... & Fu, H. (2024). Confidence-aware multi-modality learning for eye disease screening. *Medical image analysis*, *96*, 103214. https://doi.org/10.1016/j.media.2024.103214

Authors Biography

Savita Mamadapur is a dedicated academician and research scholar with a strong passion for teaching and innovation in engineering education. She has 11 years of experience in teaching and currently serves as an Assistant Professor, at Jain Institute of Technology, Davangere where she actively contributes to the academic and professional development of her students. Alongside her teaching responsibilities, she is engaged in research focusing on advancements in text, image, and video compression techniques, particularly in the field of medical imaging. Her work aims to improve data efficiency and support technological progress in healthcare applications. She is committed to enhancing the teaching–learning process through effective course design, dynamic classroom practices, and the integration of modern technology. Savita has participated in professional development initiatives such as faculty training and educational programs that strengthen her pedagogical skills. With a strong sense of dedication, creativity, and continuous learning, she strives to contribute meaningfully to both academia and society. Her long-term goal is to advance research, inspire students, and support innovation in engineering education.

Dr.P. Manikandan obtained his Ph.D. in C.S.E. in the area of Machine Learning and Data Mining from Anna University, Chennai. He has obtained his M.E (CSE) from Anna University, Chennai and B.E (CSE) from Bharathiyar University, Coimbatore. He has more than 20 years of resourceful Academic, Research and Industry Experience. He has more than 50 presentations and publications in reputed International and National Conferences and Journals like SCIE, Springer and Scopus Indexed Journals. He also participated in several workshops, Refresher Courses and Faculty Development Programmes of great prominence.

Dr.P. Renukadevi is an Assistant Professor in the School of Computer Science and Engineering (Artificial Intelligence and Machine Learning). She received her Ph.D. in Computer Science and Engineering from Anna University, Chennai, in 2023. With over 20 years of teaching experience, she has developed strong expertise in data analytics and machine learning. Her research contributions include around 20 publications in reputed journals and international conferences. She has also received a research grant of INR 40,000 from the Indian Council of Medical Research (ICMR), reflecting her commitment to interdisciplinary and impactful research. She has actively mentored numerous undergraduate and postgraduate students, guiding their academic and research projects. She has contributed to curriculum development through the publication of seven academic patterns and has successfully completed more than 16 NPTEL courses. Her dedication to teaching and mentoring has been recognized with an appreciation certificate.