JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 6 No. 3 (2025) 429-442
https://publisher.uthm.edu.my/ojs/index.php/jscdm

EffNetEye: A Multimodal Fusion Model for Multiclass Classification of Retinal Diseases

Zameer Fatima[1]*, Parneeta Dhaliwal[1], Deepak Gupta[2]

[1] *Department of Computer Science and Technology, School of Engineering, Manav Rachna University, Sector 43 Aravalli Hills, Faridabad, Haryana, 121004, INDIA*

[2] *Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, Sector 22, Rohini, Delhi, 110085, INDIA*

*Corresponding Author: zameer.fatima@gmail.com
DOI: https://doi.org/10.30880/jscdm.2025.06.03.027

Article Info

Received: 3 September 2025
Accepted: 26 November 2025
Available online: 30 December 2025

Keywords

Multimodal imaging, deep learning, machine learning, eye disease, OCT, fundus

Abstract

The Eye is the most sensitive human organ; if affected by any disease, it can hinder the individual's quality of life. Some Retinal diseases, such as Epiretinal Membrane (ERM), Age-related Macular Degeneration (AMD), glaucoma, and Retinal Vein Occlusion (RVO), are major contributors to eyesight loss. Timely detection of such diseases is essential for successful treatment. A novel model called 'EffNetEye has been proposed for the classification and diagnosis of retinal diseases by combining two modalities, fundus and OCT. The suggested model provides a simple but effective dual-modality feature-level fusion approach. It uses two EfficientNetB0 backbone networks to extract features from each modality and classifies four retinal diseases: AMD, ERM, RVO, and Normal. None of the studies on multimodal approaches included ERM disease, which distinguishes this model from the existing multimodal approaches. A total of 5,484 image datasets were constructed from three publicly available datasets: OIA-ODIR, RFMid, and OCTDL. Different preprocessing steps are applied to each modality image to address the domain differences among the three datasets. OCT images were preprocessed with Wiener filtering to reduce speckle noise and to improve local contrast in fundus images; the CLAHE technique was applied. Additionally, data augmentation was applied to address class imbalance in the dataset. The model was trained on a combined training and validation dataset and evaluated using 5-fold stratified cross-validation to ensure consistency and eliminate bias. The use of Grad-CAM demonstrated the model's ability to highlight clinically relevant features in both fundus and OCT scans during prediction. Finally, the model was tested on an independent test set, which showed strong classification performance, achieving an accuracy of 94.2% and a high AUC of 99.99%.

1. Introduction

Vision plays an important role in human life. Timely diagnosis of eye diseases is necessary for accurate treatment. Diseases such as diabetic retinopathy, glaucoma, and AMD can damage the retina of the eye extensively. There are several imaging modalities, such as fundus, OCT, OCTA, etc., to detect these diseases. To diagnose retinal diseases, fundus or OCT scans are mostly suggested by an ophthalmologist to capture different aspects of retinal functionality. Use of multiple modalities like OCT and Fundus imaging for the detection of a disease has become widely used by doctors [1]. An ophthalmologist first suggests a fundus scan; if the disease is

not identified by a fundus scan, they suggest an OCT scan. Fundus scan offers a broad 2D image of the retina and surrounding tissues, while OCT provides a comprehensive cross-sectional view of retinal layers. Combining both modalities enables more efficient and accurate disease-related changes, which can improve diagnostic performance and support optimal treatment [2, 3].

The process of combining information from several scans of the same disease, from different imaging modalities, is called image fusion. The fused image preserves all relevant and complementary information from the input images, which makes it more helpful for disease analysis by doctors. Image fusion integrates the critical details extracted from two or more images. The region-based method involves partitioning input images into distinct regions using various segmentation techniques. The features are derived from the specified regions, and appropriate integration rules were employed to achieve the combined results [4]. CIRF combines reconstruction and fusion with deep learning to better integrate multi-modality medical images, improving both image quality and detail preservation [5]. In 2025, a fusion method based on interval gradients and CNNs was proposed to decompose and recombine image information, preserving structure and texture more effectively than traditional multiscale fusion approaches [6].

Retinal fluid segmentation is a common sign used by medical experts in diagnosing diseases like diabetic macular edema, AMD, and RVO. Deep learning-based retinal fluid segmentation can improve the accuracy of semantic segmentation [7]. The authors [8] summarized various deep learning paradigms, such as CNN, FCN, and U-Net, utilized for the segmentation of retinal fluid in OCT images. A recent study published in 2025 reported that foundation models such as the Segment Anything Model (SAM), can be adapted for medical image segmentation and are capable of achieving strong, consistent results across multiple anatomical regions [9]. The growing trend of segmentation models shows a shift in medical imaging, moving from networks designed for a single task to flexible models that perform well across different datasets.

Although the latest developments in deep learning have demonstrated impressive results in analyzing medical images, especially for the automated detection and classification of retinal diseases using single imaging modalities. A multimodal strategy that combines information from fundus and OCT images can provide a more comprehensive understanding of retinal conditions, improving diagnostic accuracy compared to using either modality alone. The purpose of this research is to demonstrate the effectiveness of a multimodal framework for accurate diagnosis of retinal disorders, which will support clinicians in the timely and accurate identification of diseases. Although several multimodal approaches have been proposed, many are complex or primarily focus on diseases such as DME and AMD, leaving ERM largely unexplored. This shows a research gap in diagnosing ERM disease by deep learning models.

We developed EffNetEye, a lightweight multimodal deep learning architecture that combines features from Fundus and OCT images to classify four retinal diseases, including ERM. Features from each modality were extracted using a pre-trained EfficientNetB0 architecture and then combined and passed to a classification layer. This dual-modality approach captures complementary information while retaining an effective architecture that is simple to train. The primary contributions of this work are as follows:

- We developed a lightweight multimodal fusion model, 'EffNetEye', that combines features from fundus and OCT scans using a dual EfficientNetB0 backbone network.
- We combined three publicly available datasets (OIA-ODIR, RFMid, and OCTDL), and only four disease categories - AMD**,** RVO**,** ERM**,** and Normal were chosen for the research. Including ERM disease in our study distinguishes it from previous multimodal approaches.
- Extensive evaluation of the model was conducted on independent test data and achieved high accuracy (94.2%) and AUC (99.99%).
- For visualizing clinically relevant retinal structures, we applied Grad-CAM on the OCT and fundus images, which can help the ophthalmologist in disease diagnosis.

2. Related Work

Retinal disease classification using multimodal imaging data presents many challenges. One of the main challenges is the diversity in image quality, which can result from patient-specific factors, variations among imaging equipment, and the skill set of the operator [3]. For both doctors and automated systems, poor-quality images, such as those with blur, uneven illumination, or color distortions, can drastically reduce the accuracy of diagnosis [7].

To classify a variety of diseases, a strategy was proposed by [10] that utilizes **a** multilevel glowworm swarm optimization CNN. Two phases were suggested for the proposed system: preprocessing and categorization. Following preprocessing, the pictures were input into the MGSCNN classifier for the purpose of categorizing an image as either normal or abnormal (encompassing 39 various classifications of ailments). In the convolutional neural network classifier, utilizing the Glowworm Swarm optimization technique, this methodology attains a remarkable accuracy of 95.09% through diverse metrics.

The authors of [11] generated a multi-model CNN for categorizing AMD into four categories: normal, wet AMD, dry AMD, and PVC by using two imaging modalities, fundus and OCT. To facilitate the visual understanding of the contributions made by distinct modalities to the ultimate prediction, a technique called class activation mapping was adapted for application in a multi-modal context. Two methods for data augmentation were used: GAN-based image synthesis and loose pairing. The loose pairing technique associates a fundus picture with an OCT picture based on their respective classes instead of the identities of the eyes. Experiments were conducted on 1,094 fundus images and 1,289 OCT images.

In order to effectively represent the modality-specific characteristics inherent to various modalities, [12] introduced a new "modality-specific attention network" designed for the classification of multimodal ocular diseases. Their experimental findings derived from a medically recorded "multi-modal" ocular picture dataset indicate that "MSAN" exhibits superior performance compared to other established single-modals.

A retinal "fundus" and OCT-based automated "Macular Edema" and healthy eye classification methodology were suggested by [13]. Their deep ensemble learning system recognizes and processes input fundus and OCT scans using a deep CNN. After processing, the second layer retrieves features from both images. The retrieved features are combined and fed to a hybrid model using ANN, SVM, and Naïve Bayes. The proposed framework diagnosed "macular edema" and healthy participants with 94.33% accuracy.

Project Macula postmortem OCT and fundus imaging data were utilized by [14] to diagnose AMD. These models used pretrained VGG-19 and transfer learning utilizing a random forest. Fundus-OCT combination boosted evaluation with 0.969 AUC of and an accuracy of 90.5%. The "Delong Test" showed that combining OCT and fundus data performed better than using either OCT alone or fundus alone. The “multimodal random forest" model outperformed both “constrained Boltzmann machines” and "deep belief network" methods.

The authors of [15] developed a new "multi-view deep learning" methodology for the diagnosis of "cervical dysplasia (CDD)," combining multi-views of acetic images and iodine images, obtained from colposcopy. This feature-level fusion (FLF) method effectively records the intricate relationship between the iodine and acetic image views and adequately employs insights from both perspectives. The "FLF" method utilizes "attention mechanisms, enabling one view to support another or facilitating mutual assistance between both views to enhance feature learning.

A novel "feature-level fusion" technique was used by [16], to integrate the OCT and OCTA modes in the multimodal model for the evaluation of "choroidal neovascularization (CNV)" in neovascular AMD. The outcomes were contrasted with an ophthalmologist's identification. To demonstrate its potential for clinical application, the top model was assessed on two additional data sets. On the internal data set, the best model demonstrated, similar to retinal specialists, achieved an AUC of 0.9796 and 95.5% accuracy on multi-modal image inputs.

The authors of [17] employed multimodal imaging techniques to establish a "deep learning model" aimed at predicting retinal vascular disorder, such as "diabetic macular edema", "neovascular age-related macular degeneration", "myopic choroidal neovascularization", "branch retinal vein occlusion", and "central retinal vein occlusion", while also assessing the necessity for "anti-VEGF treatment". The model demonstrated average AUC values of 0.987 for the prediction of "retinal vascular" diseases and 0.969 for the prediction of diseases that require treatment, respectively. The heat map illustrates the capability of the model to discern disease characteristics via various retinal imaging techniques, including "fundus images, "OCT, and "Fluorescein Angiography.

The authors of [18] demonstrated that the integration of OCT, infrared imaging, and diabetes-related data for retinal vein occlusion (RVO) resulted in an accuracy of 95.20%. Advanced fusion methodologies, such as vertical plane, spatially invariant, and attention-based networks, demonstrated enhanced performance when compared to single-modal baselines or human experts. The author of [19] integrated optical coherence tomography with infrared imaging via vertical plane feature fusion, resulting in an accuracy of 96.08% and AUC values exceeding 0.96 across various subtypes of age-related macular degeneration, demonstrating performance on par with experienced ophthalmologists.

3. Materials and Methodology

3.1 Dataset Overview

Two types of imaging modalities were used in the research: fundus and OCT. Three publicly available datasets were used specifically: OCTDL [20], OIA-ODIR [21], and RFMid [22]. OIA-ODIR and RFMid datasets were used for the fundus images, and the OCTDL dataset for the OCT images. The ODIR dataset consists of 5000 images of 8 disease categories: "normal", "diabetes", "glaucoma", "cataract", "AMD", "hypertension (H)", "pathological myopia", and other diseases. The other disease category contains thirty-two retinal diseases. These thirty-two diseases were converted from multilabel to multiple classes and again labelled by their disease category; therefore, the entire OIA-ODIR data set was categorized into 39 disease classes. Out of these 39 classes, only four

classes were chosen for the research: AMD, RVO, ERM, and Normal. RVO disease in the ODIR dataset contains only 50 images, which was very few for the experiments; therefore, RVO disease fundus images of the RfMid dataset were added to the RVO disease in the ODIR dataset. The RFMid dataset contains 3200 original color fundus photos with 45 different disease categories.

There are more than 2000 OCT images in the OCTDL [20] dataset, which is arranged by disease class and ocular disorder. The seven disease categories make up the dataset. Out of seven disease categories from OCTDL, only four were chosen for the experiment: AMD, RVO, ERM, and NO. Table 1 shows the quantity of OCT and fundus images within each dataset.

Table 1 *Number of OCT and Fundus images in each dataset*

Disease	OCTDL (OCT)	RFMiD (Fundus)	OIA-ODIR (Fundus)
AMD	1231	-	280
RVO	101	86	50
ERM	155	-	375
Normal	332	-	2874

3.2 Data Pre-processing and Augmentation

This step prepared raw images for training by enhancing their quality and expanding the dataset size to improve model generalization and handle potential imbalances. Initially, the image loading step includes a check to skip any corrupted image files. The preprocess_image function takes an image file path and its corresponding modality, which may be either Fundus or OCT. The function includes opening the image, converting it to RGB format, and resizing it to 224 × 224 pixels. For Fundus images, the Contrast Limited Adaptive Histogram Equalization (CLAHE) approach is used to improve contrast and highlight fine retinal details, especially in areas with uneven illumination. To make retinal structures more visible in OCT images, a Wiener filter is used to reduce speckle noise while preserving edge details. Fig. 1 shows the examples of images before and after preprocessing.

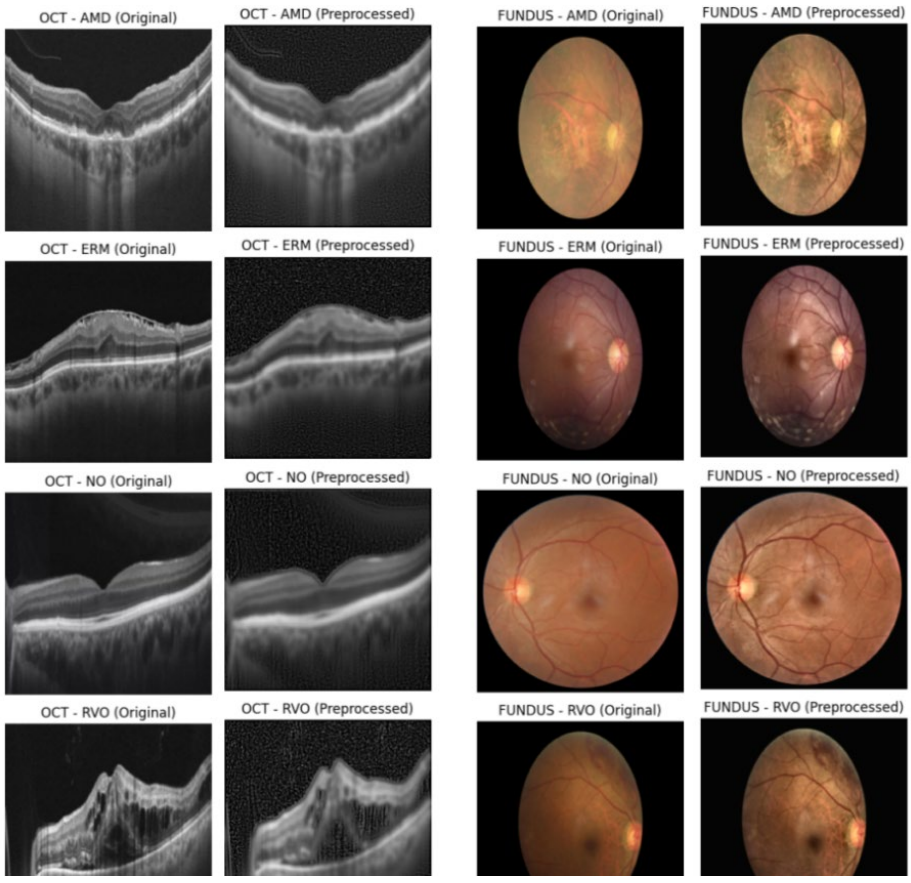

Fig. 1 *OCT and Fundus images before and after image pre-processing*

When a class has fewer than 800 images, the ImageDataGenerator function was used for data augmentation until the desired count was reached. The process of data augmentation includes transformations such as zooming, horizontal flipping, adjustments to width and height, and random rotations. For classes with more than 800 images, 800 samples were randomly selected to ensure a consistent dataset size. Finally, the dataset was partitioned into three sets: training, validation, and testing, with a 70%, 15%, and 15% split, respectively.

3.3 EffNetEye Model Architecture

The EffNetEye model is a multimodal fusion approach for the diagnosis and classification of multiple retinal diseases. By combining features from two different modalities: OCT and Fundus, the model can better categorize different retinal disorders.

Let $X_F \in R^{H\times W\times 3}$ denote a Fundus image, and $XO \in RH \times W \times 3$ denote an OCT image, where H is the height, and W is the width of the images. Each modality is passed through a pre-trained EfficientNetB0 model without the top classification layer to extract feature vectors:

$$F_F = f_{\{EffNetB0\}(X_F)} \in \mathbb{R}^{\{1280\}} \quad (1)$$

$$F_O = f_{\{EffNetB0\}(X_O)} \in \mathbb{R}^{\{1280\}} \quad (2)$$

where F_F and F_O are the 1280-dimensional feature vectors for Fundus and OCT images, respectively, the extracted feature vectors are concatenated to form a combined multimodal fusion representation, as shown in equation 3.

$$F_{\{fusion\}} = [F_F \| F_O] \in \mathbb{R}^{\{2560\}} \quad (3)$$

Where $[\mathbf{F_F}\|\mathbf{F_O}]$ denotes vector concatenation.

The fused feature vector is passed through a series of dense layers with ReLU activation and dropout for regularization:

$$H_1 = ReLU\left(W_1 F_{\{fusion\}} + b_1\right) \quad (4)$$

$$H_2 = Dropout(H_1, p = 0.5) \quad (5)$$

$$H_3 = ReLU(W_2 H_2 + b_2) \quad (6)$$

$$\hat{y} = Softmax(W_3 H_3 + b_3) \quad (7)$$

Where:

- $W_1 \in R^{256\times 2560}$, $b_1 \in R^{256}$
- $W_2 \in R^{128\times 256}$, $b_2 \in R^{128}$
- $W_3 \in R^{4\times 128}$, $b_3 \in R^4$
- $\hat{y} \in R^4$ is the predicted probability vector over four classes (AMD, RVO, ERM, Normal)

The network was trained using the sparse categorical cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(\hat{y}_{i,yi}) \quad (8)$$

Where N is the number of training samples, y_i is the true label for sample i, and $\hat{y}_{i,yi}$ is the predicted probability of the correct class.

The model employs a simple feature concatenation strategy to integrate information from different inputs. Although several studies have explored complex fusion methods such as attention mechanisms and tensor-based fusion, our experiments showed that simple concatenation was both effective and efficient. The model is able to capture useful patterns without the need for additional fusion layers because the retrieved features from each imaging modality worked well together. This lightweight design preserves high levels of accuracy and generalization while reducing computing complexity. The model's architecture is shown in Fig. 2.

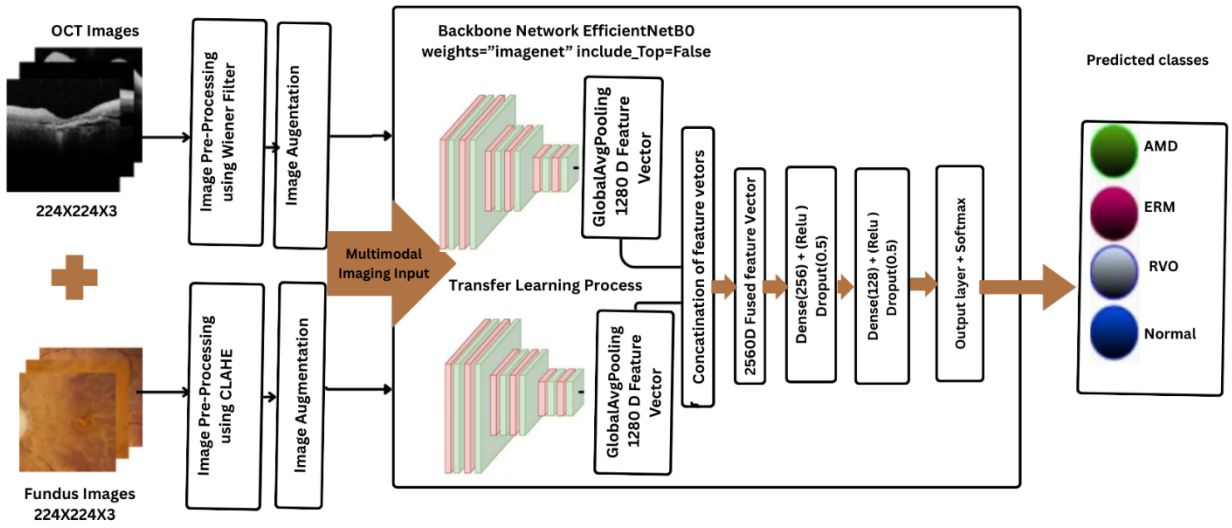

Fig. 2 *Model architecture*

The EfficientNetB0 model serves as a foundational framework for "feature extraction, which is initialized with weights from the "ImageNet" dataset. The uppermost classification layer of EfficientNetB0 is eliminated by keeping "include_top = False". A 2D GAP layer is added to the output of the EfficientNetB0 base, resulting in a 1280-dimensional "feature vector" for each input image fundus and OCT. Using a Concatenate layer, these "feature vectors" are combined to produce a composite feature vector of shape 2560. The fused feature vector was subsequently processed through a sequence of "dense layers":

- A "dense layer" with 256 units and "ReLU activation".
- A "dropout layer" with a regularization rate of 0.5.
- A "dense layer" with 128 units and "ReLU activation".
- A final output layer consisted of a "dense layer" utilizing "softmax activation", which facilitated the generation of probabilities for the four categories.

The fusion model was compiled with the "Adam optimizer." The "sparse_categorical_crossentropy loss function" was used for multiclass classification with integer labels. The model was trained for 50 epochs.

4. Result

To show the performance of the suggested EffNetEye model, the training and testing procedures were carefully organized. TensorFlow and Keras were used to build the model, and GPUs were used to train it. The process includes two steps: K-fold cross-validation, and then training and evaluating the model.

4.1 K-Fold Cross-Validation

To evaluate the model's consistency over splits, a stratified 5-fold cross-validation was performed on the combined training and validation data. Stratified split guarantees that the distribution of each disease class is preserved across all folds, which is important for imbalanced datasets. To prevent overfitting the training dataset and to observe the validation loss, early stopping was used. If the validation loss does not improve over a period of five epochs, the training process is terminated prematurely, and the model weights are reset to those from the epoch with the lowest validation loss. After training for each fold, the model's effectiveness was evaluated on the validation set for that fold. For each fold, performance was assessed using metrics such as precision, recall, F1-score, accuracy, and AUC. Cross-validation metrics corresponding to every fold are shown in Table 2.

Table 2 displays the cross-validation metrics and results for each fold, indicating the model's performance over five-fold cross-validation. Mean and standard deviation were also calculated. The model achieved an average accuracy of 94.15%, precision of 94.23%, recall of 94.15%, and F1-score of 94.14%. Low standard deviations for these metrics (Accuracy: 0.0104, Precision: 0.0105, Recall: 0.0104, F1-score: 0.0104) indicate that the model performs consistently across different data partitions, demonstrating high stability and generalization ability.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 429-442

Table 2 *Cross-validation metrics across each fold*

Fold	Accuracy	Precision	Recall	F1
1	92.63	92.65	92.63	92.6
2	95.31	95.30	95.31	95.29
3	94.86	95.0	94.85	94.85
4	93.75	93.83	93.78	93.78
5	94.19	94.34	94.17	94.17

The AUC-ROC curve was also computed using the one-hot-encoded true labels and the predicted probabilities. A graph of the AUC score across all of the folds is shown in Fig. 3.

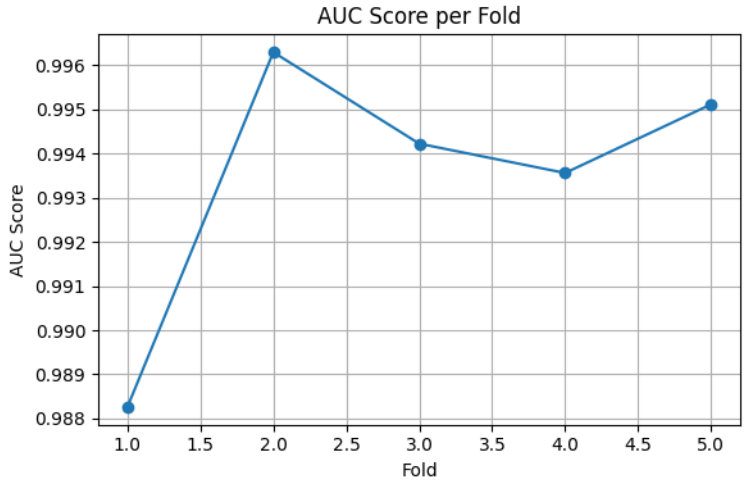

Fig. 3 *AUC score across folds*

For each fold, Grad-CAM visualizations were produced for several randomly selected samples from the validation set, which is shown in Fig. 4. This approach facilitates the visualization of the areas within the input image that hold significant relevance for the estimate made by the model, thereby offering significant findings about the aspects the model is focusing on.

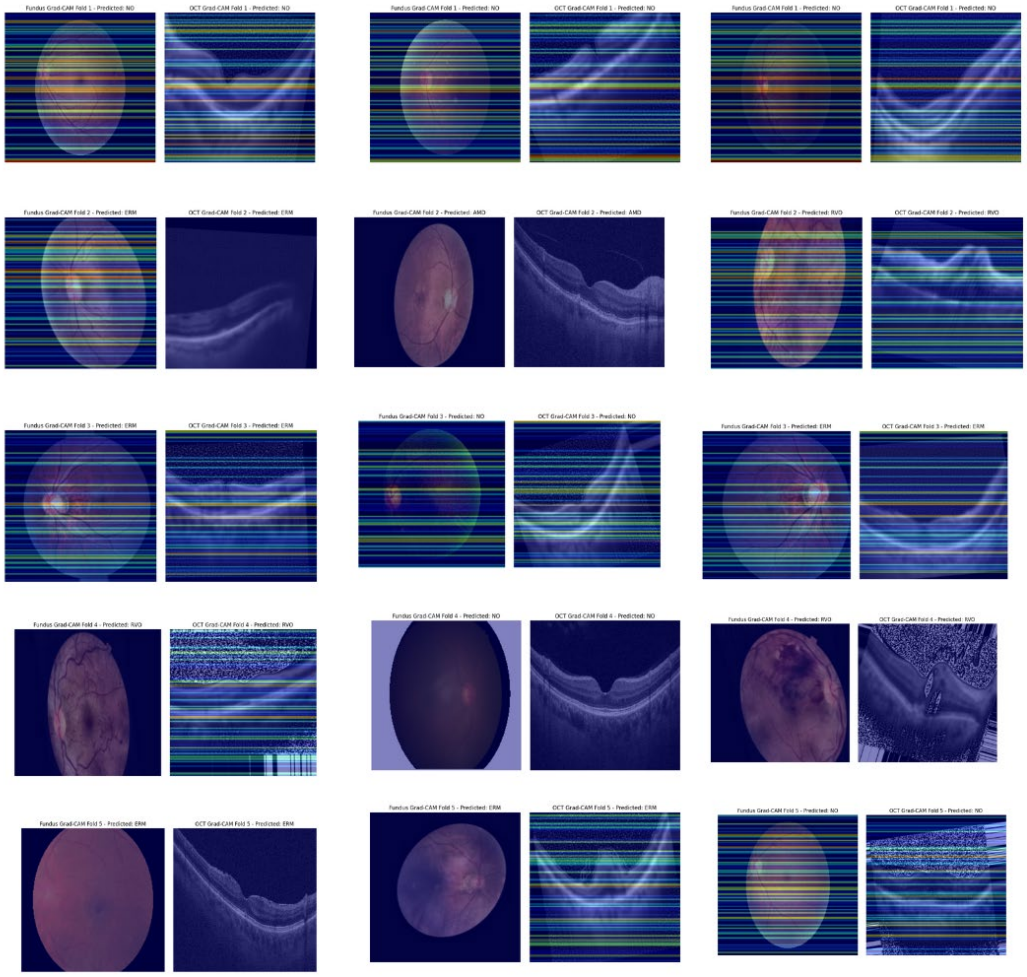

Fig. 4 *Grad-CAM for randomly selected samples*

4.2 Training

This phase comprises training the final model on an expanded dataset, followed by an assessment of its performance on a test set that has not been seen before. The datasets utilized for training and validation during the splitting phase were amalgamated to form an expanded training set for the final model. The images from the integrated training set were systematically loaded and subjected to pre-processing procedures. Features were derived from these images utilizing the EfficientNetB0 feature extractor. A novel instance of the fusion model has been constructed and compiled. The model was developed utilizing the characteristics and classifications derived from the integrated training dataset. In a manner similar to the K-Fold stage, early stopping was implemented.

A plot of training loss and accuracy across epochs is shown in Fig. 5 to visualize training progress. The training history, as depicted in the training curve, shows a steady increase in accuracy and a decrease in loss over epochs, indicating effective learning. The final model achieved 99% accuracy and a near-perfect AUC of 100% on the combined training and validation data.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 429-442

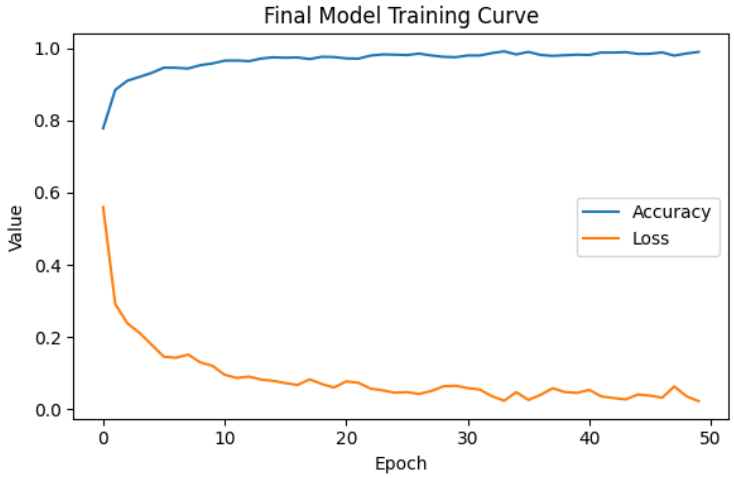

Fig. 5 *Model training and loss curve*

The final model classification report and confusion matrix are shown in Table 3 and in Fig. 6, respectively.

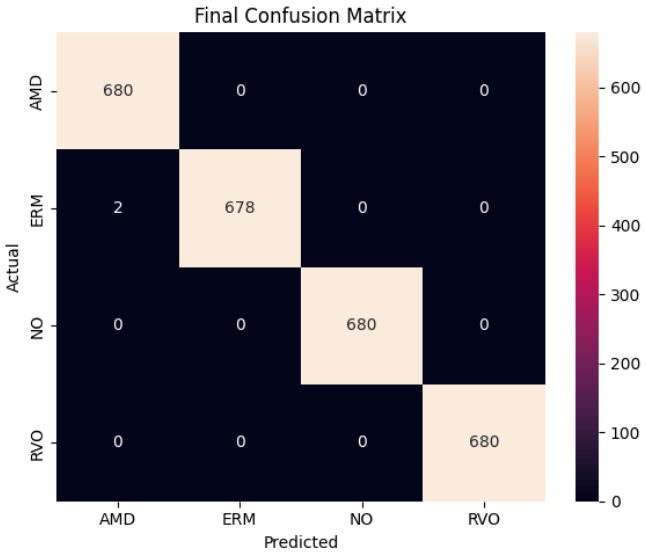

Fig. 6 *Model training confusion matrix*

Table 3 shows that the model performs well across all four disease types, with precision, recall, and F1-score values in the range 0.98-1.00. Both AMD and ERM have a precision and recall rate of 0.99, and this suggests their high reliability and few false positives and false negatives. The precision of the Normal class is the highest (1.00), and the recall is slightly lower (0.98). In comparison, RVO scores highest across all parameters, and detection accuracy is 100% on the training set. The total accuracy of 0.99, the macro and weighted averages, 0.99, shows that the model has excellent, balanced performance, regardless of the class distribution. The results indicate the effectiveness and precision of the classifier when it comes to distinguishing different retinal diseases.

Table 3 *Model training classification report*

Disease	Precision	Recall	f1-score	Support
AMD	0.99	1	0.99	680
ERM	0.99	0.99	0.99	680
Normal	1	0.98	0.99	680
RVO	1	1	1	680
accuracy	0.99	0.99	0.99	0.99
macro avg	0.99	0.99	0.99	2720
weighted avg	0.99	0.99	0.99	2720

4.3 Testing

The EffNetEye model was tested on an independent test set. The confusion matrix illustrates the relationship between the predicted output and the real labels on the test set, highlighting areas of strong performance and specific misclassification. The test confusion matrix (see Fig. 7) and the classification report (Table 4) exhibit the model's effectiveness on unknown data. The overall test accuracy is 0.93, with performance varying among individual classes as detailed in the classification report.

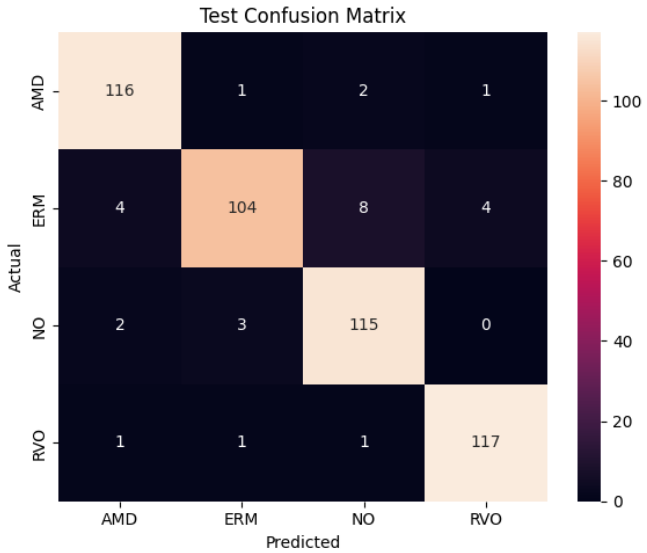

Fig. 7 *Confusion matrix on test set*

Table 4 shows that the model is very strong in generalizing its performance in the test dataset, with a general accuracy of 0.94 in all four disease categories. The F1-scores, precision, and recall are all consistently high, with a range of 0.87- 0.97, which means that it is a reliable detector with minimum false detection. AMD is also performing well with a precision of 0.94 and a recall of 0.97, and ERM has a comparatively lower recall of 0.87, indicating that it is more difficult to detect all the instances of ERM than it is with the other classes. The normal class has balanced performance with an F1-score of 0.93, and RVO has the best test measures with an F1-score of 0.97, which indicates a high level of predictive power. The classifier is robust and effective, as sometimes, there are imbalanced conditions and other times balanced conditions, which are supported by the macro and weighted averages of 0.94 to confirm that the classifier has been effective in real-world retinal disease classification.

Table 4 *Test classification report*

Disease	Precision	Recall	f1-score	Support
AMD	0.94	0.97	0.95	120
ERM	0.95	0.87	0.91	120
Normal	0.91	0.96	0.93	120
RVO	0.96	0.97	0.97	120
accuracy			0.94	480
macro avg	0.94	0.94	0.94	480
weighted avg	0.94	0.94	0.94	480

4.4 Class-Specific Model Performance

To assess the discriminatory capability of the multimodal fusion model for each disease category, ROC curves were generated for the independent test set using a one-vs.-rest methodology. The ROC curves and corresponding AUC values for the classes of AMD, ERM, Normal, and RVO are shown in Fig. 8. The ROC curves for all classes are above the diagonal line, approaching the top-left region of the plot, indicating a robust classification performance. The computed AUC values were 1.00 for AMD, 0.99 for ERM, 0.99 for Normal, and 1.00 for RVO. The elevated AUC values indicate the model's superior capacity to differentiate among each specific disease category and the other classes, thereby reinforcing its efficacy in the classification of multiple retinal diseases.

J. of Soft Computing and Data Mining Vol. 6 No. 3 (2025) p. 429-442

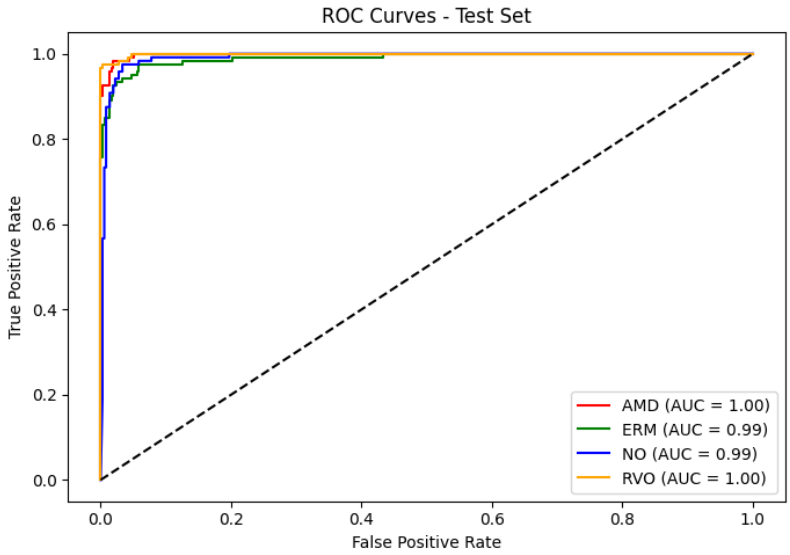

Fig. 8 *ROC curves on the test set*

Grad-CAM visualizations were produced to elucidate how the model makes decisions for representative samples from the test set, as shown in Fig. 9. The region of input images that the model concentrated on when generating its predictions is shown by these heatmaps superimposed on the original photo The Grad-CAM results indicate that the model is attending to relevant anatomical structures and pathological features within the fundus and OCT images.

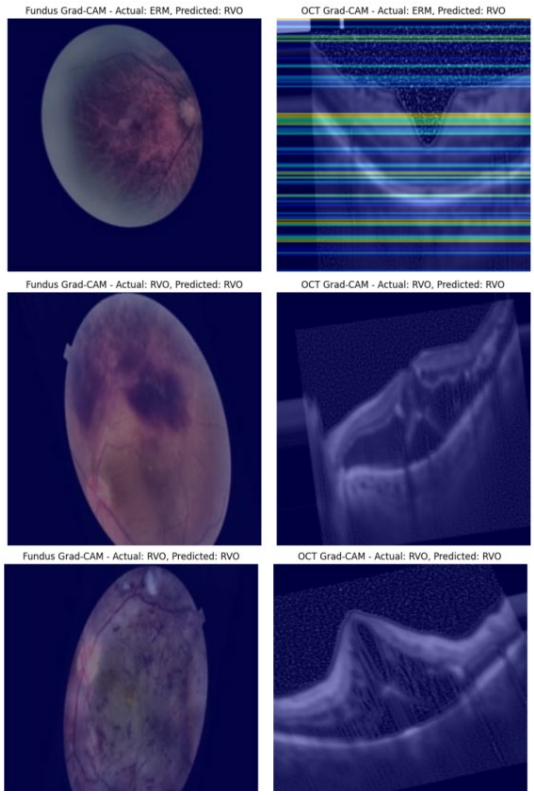

Fig. 9 *Grad-cam samples of test data*

4.5 Comparison with Existing Approaches

Table 5 shows an extensive comparison between the proposed EffNetEye model and existing leading approaches. As shown, EffNetEye achieved the highest AUC while uniquely incorporating ERM into the classification task.

Table 5 *Comparative performance of the proposed EffNetEye model with existing models*

Author	Modality	Disease classify	AUC	Accuracy
[3]	Pair of Fundus Images(Left and Right Eye)	AMD, Cataract DME, Glaucoma	99.76	92.41
[16]	Fundus, OCT	FA/ICGA DME, nAMD, mCNV, BRVO, CRVO	98.7	93
[13]	OCT and fundus	AMD	98.1	94.6
[14]	Infrared and OCT	AMD, Normal, Dry, Wet AMD	96.08	91.59
[17]	Infrared and OCT	Diabeties RV0, DME, Normal	99.5	95
EffNetEye model	OCT and fundus	AMD, RVO, Normal ERM	99.99	Training 99 Test 94.2

Compared with existing methods, the proposed EffNetEye model offers clear advantages. Fundus-DeepNet [3] achieved high performance using fundus images, but its accuracy of 92.41% and AUC 99.76% remained slightly lower than EffNetEye's perfect AUC. Authors [16] employed multimodal inputs (fundus, OCT, and FA/ICGA) and achieved reasonable accuracy (93%) and an AUC of 98.7%, but required complex, resource-intensive imaging. In contrast, EffNetEye achieved superior performance with simpler modalities. The authors of [13] focused only on AMD detection using OCT and fundus images, reporting an AUC of 98.1% and accuracy of 94.6%; while strong in single-disease classification, it lacked broader coverage of retinal disorders. Authors of [14] relied on infrared and OCT data to classify AMD subtypes, but their accuracy (91.59%) and AUC (96.08%) were lower, and the disease scope was narrower. [17] integrated diabetes-related information and infrared imaging to classify RVO, DME, and normal cases with strong performance (AUC 99.5%, accuracy 95%), but again excluded ERM. In contrast, EffNetEye not only achieved the highest AUC (100%) but also uniquely incorporated ERM into the classification task, offering broader and more clinically relevant diagnostic capabilities.

5. Conclusion

A multimodal fusion model, EffNetEye, for multiclass retinal disease classification is proposed. The model takes multimodal scans: fundus images and OCT of retinal diseases, and classifies them into four disease classes: AMD, ERM, RVO, and Normal. The model achieved consistent performance on both the validation folds and the independent test set. The high AUC scores of 1.0 for most classes in the training set and 0.99-1.00 in the test set indicate excellent discriminative power, meaning the model is very good at distinguishing between different disease classes and healthy cases. The training accuracy and loss curves demonstrate that the model converged well during training, achieving excellent accuracy and minimal loss on the training data. The "K-fold cross-validation" provides an accurate assessment of the model's functionality and helps in hyperparameter tuning and model selection. The final model was trained on a larger dataset to build a production-ready model, and evaluation on the test set provides an impartial assessment of its generalization performance. Earlier studies tend to concentrate mainly on AMD, DME, or RVO. On the other hand, EffNetEye provides a wider, more clinically useful diagnosis by combining ERM with other retinal illnesses. The experimental results show that the EffNetEye multimodal fusion model is useful for detecting multiple retinal diseases accurately. Furthermore, EffNetEye demonstrated the ability to predict the correct disease category and achieved the highest AUC (99.99%) among all examined approaches. In general, the EffNetEye model is more clinically advantageous than other models due to its superior performance and broader disease coverage. Future study could focus on real-time deployment and clinical validation of the models

Acknowledgement

I would like to convey my heartfelt gratitude to my guide, Professor Parneeta Dhaliwal, for her guidance, encouragement, and continuous support throughout this research. I am also sincerely thankful to my co-guide, Dr. Deepak Gupta, for his continuous assistance, motivation, and technical insights. I extend my sincere appreciation to everyone who, in any way, supported me during this research journey. Your contributions, whether direct or indirect, are truly appreciated.

Conflict of Interest

The authors have no relevant financial or non-financial conflict of interest regarding the publication of the paper.

Author Contribution

The author Zameer Fatima confirms responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. Parneeta Dhaliwal and Deepak Gupta contribute to conceptualization, methodology, and project administration, supervision, manuscript writing, review, and editing.

References

[1] Huang, X., Wang, H., She, C., Feng, J., Liu, X., Hu, X., Chen, L., & Tao, Y. (2022). Artificial intelligence promotes the diagnosis and screening of diabetic retinopathy. *Frontiers in Endocrinology, 13*. https://doi.org/10.3389/fendo.2022.946915

[2] Odaibo, S. G., et al. (2019). Generative adversarial networks synthesize realistic OCT images of the retina. *arXiv preprint arXiv:1902.06676*.

[3] Al-Fahdawi, S., Al-Waisy, A. S., Zeebaree, D. Q., Qahwaji, R., Natiq, H., Mohammed, M. A., Nedoma, J., Martinek, R., & 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

[4] Meher, B., Agrawal, S., Panda, R., & Abraham, A. (2019). A survey on region based image fusion methods. *Information Fusion, 48*, 119–132.

[5] Zheng, J., Xiao, J., Wang, Y., & Zhang, X. (2024). CIRF: Coupled Image Reconstruction and Fusion Strategy for Deep Learning Based Multi-Modal Image Fusion. *Sensors*, *24*(11), 3545. https://doi.org/10.3390/s24113545

[6] Gu, X., Xia, Y., & Zhang, J. (2024). *Multimodal medical image fusion based on interval gradients and convolutional neural networks.* BMC Medical Imaging, 24, 232. https://doi.org/10.1186/s12880-024-01418-x

[7] Wan, C., Zhou, X., You, Q., Sun, J., Shen, J., Zhu, S., Jiang, Q., & Yang, W. (2022). Retinal image enhancement using cycle-constraint adversarial network. *Frontiers in Medicine, 8*, 793726.

[8] Lin, M., Bao, G., Sang, X., & Wu, Y. (2022). Recent advanced deep learning architectures for retinal fluid segmentation on optical coherence tomography images. *Sensors, 22*(8), 3055.

[9] Gu, H., Dong, H., Yang, J., & Mazurowski, M. A. (2025). *How to build the best medical image segmentation algorithm using foundation models: A comprehensive empirical study with Segment Anything Model*. *Machine Learning for Biomedical Imaging*, 3, 88–120. https://doi.org/10.59275/j.melba.2025-86a6

[10] Lee, D.-G., Jang, Y., & Seo, Y.-S. (2020). Intelligent image synthesis for accurate retinal diagnosis. *Electronics, 9*(5), 767.

[11] Wang, W., Li, X., Xu, Z., Yu, W., Zhao, J., Ding, D., & Chen, Y. (2022). Learning two-stream CNN for multi-modal age-related macular degeneration categorization. *IEEE Journal of Biomedical and Health Informatics, 26*(8), 4111–4122.

[12] He, X., Deng, Y., Fang, L., & Peng, Q. (2021). Multi-modal retinal image classification with modality-specific attention network. *IEEE Transactions on Medical Imaging, 40*(6), 1591–1602. https://doi.org/10.1109/TMI.2021.3059956

[13] Hassan, B., Hassan, T., Li, B., Ahmed, R., & Hassan, O. (2019). Deep ensemble learning based objective grading of macular edema by extracting clinically significant findings from fused retinal imaging modalities. *Sensors, 19*(13). https://doi.org/10.3390/s19132970

[14] Yoo, T. K., Choi, J. Y., Seo, J. G., Ramasubramanian, B., Selvaperumal, S., & Kim, D. W. (2019). The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: A preliminary experiment. *Medical & Biological Engineering & Computing, 57*, 677–687.

[15] Chen, T., Ma, X., Liu, X., Wang, W., Feng, R., Chen, J., Yuan, C., Lu, W., Chen, D. Z., & Wu, J. (2019). Multi-view learning with feature level fusion for cervical dysplasia diagnosis. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, & A. Khan (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (pp. 329–338). Springer, Cham. https://doi.org/10.1007/978-3-030-32239-7_37

[16] Jin, K., Yan, Y., Chen, M., Wang, J., Pan, X., Liu, X., Liu, M., Lou, L., Wang, Y., & Ye, J. (2022). Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration. *Acta Ophthalmologica, 100*(2), 512–520. https://doi.org/10.1111/aos.14928

[17] Kang, E. Y.-C., Yeung, L., Lee, Y.-L., Wu, C.-H., Peng, S.-Y., Chen, Y.-P., Gao, Q.-Z., Lin, C., Kuo, C.-F., & Lai, C.-C. (2021). A multimodal imaging–based deep learning model for detecting treatment-requiring retinal vascular diseases: Model development and validation study. *JMIR Medical Informatics, 9*(5), 28868. https://doi.org/10.2196/28868

[18] Barbosa, G., Carvalho, E., Guerra, A., Torres-Costa, S., Ramião, N., Parente, M. L., & Falcão, M. (2025). Deep learning to distinguish edema secondary to retinal vein occlusion and diabetic macular edema: A multimodal approach using OCT and infrared imaging. *Journal of Clinical Medicine, 14*(3), 1008.

[19] Chen, M., Jin, K., Yan, Y., Liu, X., Huang, X., Gao, Z., Wang, Y., Wang, S., & Ye, J. (2022). Automated diagnosis of age-related macular degeneration using multi-modal vertical plane feature fusion via deep learning. *Medical Physics, 49*(4), 2324–2333.

[20] Kulyabin, M., Zhdanov, A., Nikiforova, A., Stepichev, A., Kuznetsova, A., Ronkin, M., Borisov, V., Bogachev, A., Korotkich, S., Constable, P. A., & Maier, A. (2023). OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods. [Accessed 15 November 2024].

[21] nkicsl. (2024). GitHub - OIA-ODIR. GitHub. https://github.com/nkicsl/OIA-ODIR [Accessed 15 November 2024]

[22] RIADD. (2024). RIADD (ISBI-2021) - Grand Challenge. https://riadd.grand-challenge.org/download-all-classes/ [Accessed 15 November 2024]