AI

MDPI

Article

Comprehensive Analysis of Multimodal Fusion Techniques for Ocular Disease Detection

Veena K. M. [1], Pragya Gupta [1], Ruthvik Avadhanam [1], Rashmi Naveen Raj [1], Sulatha V. Bhandary [2], Varadraj Gurupur [3] and Veena Mayya [1,*]

[1] Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, Karnataka, India
[2] Department of Ophthalmology, Kasturba Medical College, Manipal Academy of Higher Education, Manipal 576104, Karnataka, India
[3] Center for Decision Support Systems and Informatics, School of Global Health Management and Informatics, University of Central Florida, Orlando, FL 32816, USA
* Correspondence: veena.mayya@manipal.edu

Abstract

Accurate and early identification of ocular diseases is essential to prevent vision impairment and enable timely medical intervention. In routine clinical practice, ophthalmologists rely on a structured diagnostic workflow that incorporates multiple imaging modalities to manually assess and diagnose ocular diseases. However, interpreting each modality requires significant clinical experience and can be time-consuming. These limitations can be effectively addressed through the application of AI (Artificial intelligence)-driven multimodal fusion techniques. In this study, we conducted an empirical investigation to assess the impact of different fusion strategies—including early, intermediate, and late fusion—on diagnostic performance, training requirements, and interpretability. The proposed methodology was evaluated using three publicly available datasets: FFA-Fundus (Fundus fluorescein angiography), GAMMA (Glaucoma Analysis and Multi-Modal Assessment), and OLIVES (Ophthalmic Labels to Investigate Visual Eye Semantics). Experimental results demonstrate that multimodal feature fusion improves disease detection performance. Although fused models typically required an increase in training parameters compared to single-modality models, they provided interpretability on par with that of individual single-modal networks. However, inference time increased by approximately 50% for multimodal architectures. These findings underscore the value of integrating diverse ophthalmic imaging modalities to enhance diagnostic accuracy in automated disease detection systems. At the same time, the results highlight that unimodal models containing highly discriminative features can also perform competitively, particularly when a single modality is sufficient for disease identification. Multimodal fusion provides the greatest benefit in scenarios where complementary information across modalities contributes distinct and non-redundant features. Furthermore, fusing all available modalities may not be optimal due to increased computational cost and reduced inference efficiency; thus, selective modality integration and lightweight fusion strategies are essential to balance accuracy, interpretability, and efficiency in clinical deployment.

Keywords: diabetes-related complications; fundus fluorescein angiography; fundus; optical coherence tomography; multimodal classification; fusion techniques; clinical decision support systems

Check for updates

Academic Editor: Olufemi A. Omitaomu

Received: 28 January 2026
Revised: 4 March 2026
Accepted: 13 March 2026
Published: 1 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Early and accurate diagnosis is crucial for effective intervention, particularly given that more than 2.2 billion people worldwide are affected by eye disorders. Notably, at least one billion of these cases are treatable or preventable; however, many individuals lack access to essential treatments [1]. Modern lifestyles characterized by excessive screen time, poor nutrition, and insufficient outdoor activity contribute significantly to the increasing prevalence of ocular disorders. The demand for screening for conditions such as Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), cataracts, and glaucoma continues to rise worldwide, driven by both population growth and increased longevity. In routine clinical practice, ophthalmologists follow a structured diagnostic workflow that incorporates various imaging modalities to manually assess and identify ocular pathologies. Multimodal ophthalmic imaging systems have been developed to enhance the contrast, resolution and functional capabilities of conventional techniques, thus improving diagnostic accuracy and informing treatment decisions [2]. However, the interpretation of each imaging modality, including fundus photography, Optical Coherence Tomography (OCT), Fluorescein Angiography (FA), or OCT Angiography (OCTA), requires substantial clinical expertise and a considerable investment of time. This manual process typically involves an image-by-image inspection, correlating findings across different modalities, and synthesizing these findings into a comprehensive diagnosis. Although these tools have significantly advanced diagnostic accuracy, their effectiveness is heavily dependent on the clinician's skill and the time dedicated to image analysis, elements that present challenges to scalability, especially in high-throughput or resource-constrained environments. These challenges can be effectively addressed, especially in developing countries, through the implementation of telemedicine, AI-driven diagnostics, and low-cost mobile clinics. Recent advances in automated retinal image analysis are revolutionizing the field of ophthalmology by improving the early detection and diagnosis of ocular pathologies. These AI-driven systems enable rapid and precise analysis of retinal images, facilitating timely interventions, and reducing the risk of visual impairment.

Integration of AI is essential for the comprehensive analysis of diseases and has attracted greater interest in recent studies [3–5]. Single-modal based AI techniques take data of single type or from single source as input, whereas multi-modal AI techniques integrate the data of different types from multiple sources. Multimodal disease diagnosis has shown strong performance in various recent research efforts on a range of diseases, including Alzheimer's [6–12], breast cancer [13–22], depression [23–29], and epilepsy [30–33], among others [34]. The use of multimodal disease detection enhances accuracy and robustness by integrating multiple data sources, reducing the risk of misdiagnosis and improving generalization in diverse patient populations [35,36]. Furthermore, multimodal analysis enhances disease recognition by providing additional insights that single-modal techniques often miss [12,37]. Multimodal approaches-based techniques have been applied effectively to the detection of various ocular diseases, such as Diabetic Macular Edema (DME) [38,39], glaucoma [40–43], DR [40,44,45], glaucomatous optic neuropathy [46] and optic neuropathy [47,48], yielding promising results.

The field of retinal imaging spans both local and global contexts. Relying solely on a single imaging modality for the identification of ocular diseases is no longer sufficient, as it often overlooks critical indicators. Researchers have highlighted the need for multimodal approaches to improve diagnostic accuracy, driven by advances in AI [49]. Compared to single-modality approaches, multimodal Deep Learning (DL) provides the significant advantage of integrating diverse data types, enhancing the model's capacity to learn a wider and more comprehensive set of features. By simultaneously processing different types of information during training, multimodal DL enables the effective fusion of complementary

modalities, which can potentially improve the diagnostic performance in ophthalmology. Many current multimodal approaches depend on paired datasets, wherein all imaging modalities (e.g., fundus photography, OCT, FFA) must correspond to the same patient. The construction of such datasets is logistically challenging and resource-intensive, especially at larger scales, which impacts the scalability and generalizability of current models. Although various multimodal fusion techniques have been proposed, ranging from early fusion to late fusion and hybrid approaches, these methods are frequently selected empirically without thorough evaluation, potentially limiting the robustness and effectiveness of the resulting models. These challenges highlight the need for the development of versatile foundation models that can adaptively handle multiple modalities, even in the presence of unpaired data. There is significant scope for systematically evaluating the effectiveness of different multimodal fusion strategies by examining how various fusion techniques influence classification performance on publicly available datasets.

To address these challenges, we propose a multimodal fusion-based approach for image-based disease detection that integrates information from multiple imaging modalities. Furthermore, we conduct a systematic analysis to assess the effectiveness of various fusion strategies, examining the impact of different multimodal fusion techniques on identification performance using publicly available datasets. The key contributions of this work in advancing the efforts of state-of-the-art multimodal fusion techniques can be summarized as follows:

- Design of multimodal fusion techniques for the detection of ocular disease using identical backbone networks and initial weights to identify the best fusion strategy.
- Detailed analysis of three fusion techniques, early, feature and late fusion, in terms of network parameters, performance, inference time, and interpretability.

The remainder of this article is organized as follows. Section 2 reviews the related literature, while Section 3 outlines the methodology and describes the data sets used. Sections 4 and 5, present the experimental results and their interpretation. Finally, Section 6 concludes the study and discusses the directions for future work.

2. Literature Review

Multimodal fusion enhances classification accuracy and robustness by integrating diverse data sources. It encompasses three primary strategies: early fusion, intermediate fusion, and late fusion, each offering a trade-off between integration depth and computational efficiency, while aiming to preserve fine-grained cross-modal relationships [50]. Early fusion of data from multiple modalities before feature extraction allows deep cross-modal integration, but can result in high-dimensional feature spaces and increased noise sensitivity [51]. Intermediate fusion, or representation-level fusion, is a method that separates and combines features from different modality processes, balancing multimodal information integration with model complexity [52]. Late fusion, a decision-level fusion approach, processes each modality independently and combines their outputs during decision-making, offering flexibility but potentially overlooking complex cross-modal interactions [53].

Recently, several studies have explored the use of multiple retinal imaging modalities for various decision-making tasks in ophthalmic diagnostics. Hervella et al. [54] explored the multimodal image encoding pre-training for the classification of DR. They introduced a novel self-supervised learning framework, termed multimodal image encoding, which learns feature representations from unlabeled pairs of retinal images. Compared to conventional supervised pre-training, their approach led to improved performance across multiple datasets. Wardhani et al. [55] developed a novel model for DR prediction using a multimodal early fusion approach that integrates fundus retinal images, OCT, and Electronic Health Records (EHR). The proposed method demonstrated superior performance

in DR detection compared to models based on unimodal data. Bidwai et al. [56] proposed a feature-level fusion approach that combines ResNet-extracted features from fundus and OCT images, using a DkSO-optimized Light GBM model for DR detection. Their method improved diagnostic accuracy while ensuring the interpretation of the model.

In the context of the diagnosis of glaucoma, Guangzhou et al. [57] proposed an ML-based approach using OCT and color fundus images. The method used a VGG19-based Convolutional Neural Network (CNN) to extract features from the multimodal input, which are later classified using a random forest algorithm. The thickness and deviation maps derived from OCT segmentation were integrated with CNN-based features, allowing the model to effectively differentiate between healthy and glaucomatous eyes. Mehta et al. [58] automated glaucoma detection using interpretable ML by integrating clinical data with multimodal retinal imaging, including color fundus photographs and OCT images. The authors used gradient-boosted decision trees to detect glaucoma based on extracted features and performed an interpretability analysis to identify the key features that contributed to the model's diagnostic decisions. Lim et al. [59] used fundus images, OCT scans, and demographic data for the detection of glaucoma. The features extracted using the Xception model were combined with machine learning classifiers, resulting in superior performance compared to single-modality baselines. A study by Sejong et al. [60] examined an ML model designed to elucidate the identification of glaucoma. To construct the model, they utilized fundus images, OCT scans, visual field data, and non-visual health information to illustrate the onset of glaucoma.

Multimodal fusion approaches have also been investigated for the detection of AMD and other ocular conditions [61,62]. Dong et al. [63] proposed a multimodal transformer system for the non-invasive diagnosis of Diabetic Nephropathy (DN) through retinal imaging. The study created a transformer-based DL system for the identification of DN and non-DN utilizing non-invasive multimodal data, comprising fundus images and clinical attributes. Jin et al. [64] developed a method to diagnose Choroidal Neovascularization (CNV) activity in AMD using the extracted features of the OCT and OCTA images. The authors found that the multimodal model outperformed the single-modality models in terms of accuracy and recall for detecting active CNV. Liu et al. [65] combined infrared (IR) and OCT images to predict visual acuity. By aggregating these modalities and applying a DL model, they achieved an accurate prediction of visual impairment, thus supporting early detection and timely medical intervention. Fusion techniques have also been explored for the detection of multiple ocular diseases. Hervella et al. [66] proposed an innovative self-supervised multimodal reconstruction-based pre-training framework aimed at improving diagnostic performance. Their work focuses on the detection of various retinal diseases, particularly two of the most prevalent vision-impairing ocular conditions: glaucoma and AMD.

Although multimodal data sets have gained significant attention for the detection of ocular diseases by enabling more accurate and robust diagnostic models, many existing studies rely on empirically selected fusion strategies without systematic evaluation. Furthermore, most of these approaches are validated using a single, often private dataset, which limits reproducibility and generalizability. Therefore, a comprehensive examination of different fusion techniques is essential to assess their impact on disease recognition. To address these limitations, this study investigates three fusion strategies, early, intermediate and late fusion, applied to a multimodal dataset, followed by an extensive evaluation of their effectiveness in improving classification performance. The findings provide valuable insights into the importance of multimodal fusion in the advancement of automated diagnosis of ocular diseases.

3. Methodology

In this study, we used three publicly available datasets. Section 3 provides detailed information about each dataset. For each of these datasets, we applied three multimodal fusion techniques, as discussed in Section 4. For each of these fusion techniques, we applied three standard classification models (VGG16, ResNet50, EfficientNetB0 and Swin Transformer (SwinT)), analyzing their influence on the performance of the model in the classification of ocular diseases. Figure 1 shows the entire methodology used to classify ocular diseases. These models are pre-trained on ImageNet as a feature extractor with the top classification layers removed. As input, it takes an image of shape (224, 224, 3) and outputs a feature map, which serves as input to the classification head. For all these models, All layers are made trainable and layers initialized using ImageNet weights. The algorithm used in this study is presented in Algorithm 1.

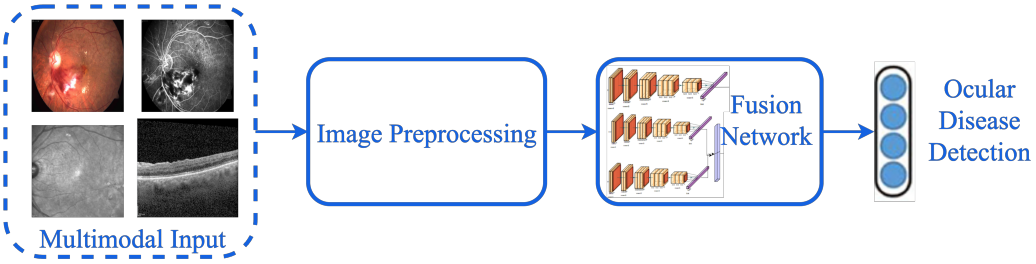


Figure 1. High-level framework for multimodal image-based ocular disease detection.

3.1. Dataset Description

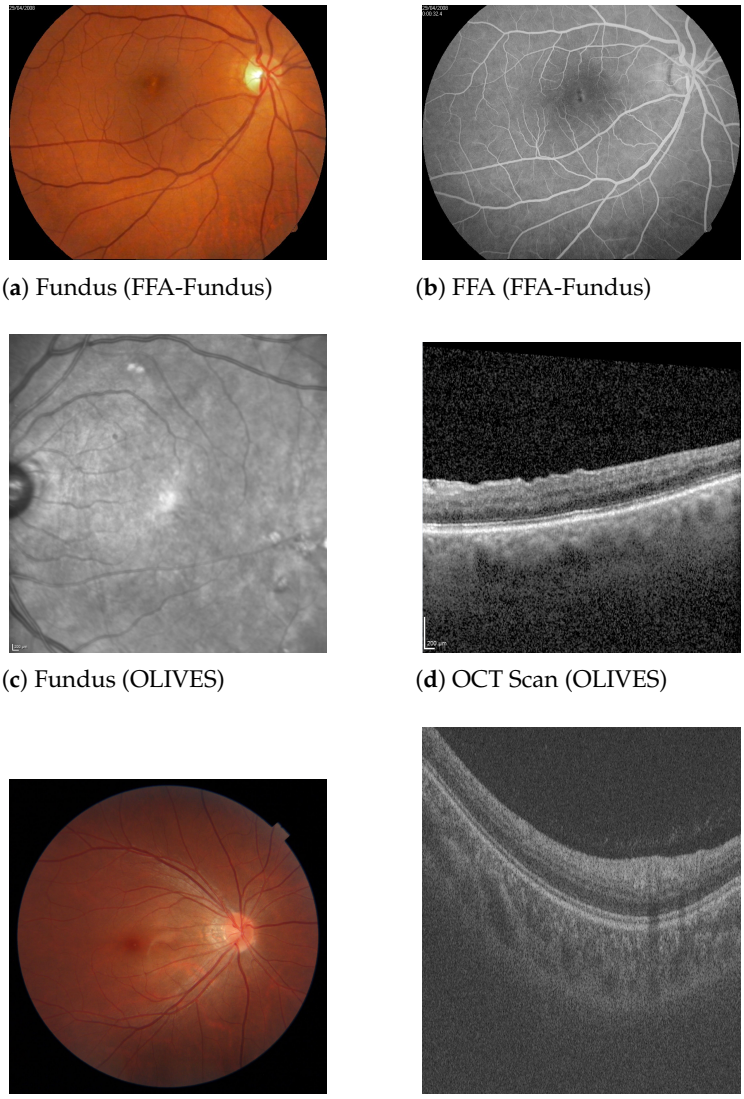
This section provides an overview of the datasets used in the development and evaluation of the proposed fusion techniques. The experiments utilized three publicly available datasets: FFA-Fundus [67], GAMMA [42] and OLIVES [68]. These datasets were chosen for their complementary characteristics, including the incorporation of multiple imaging modalities and diverse sample sizes representing various ocular diseases. This diversity facilitates experimentation across a wide range of ocular disease identification tasks, allowing for a comprehensive evaluation of the proposed models. Figure 2 shows the sample images from each of these datasets and details about the datasets utilized are presented in Table 1.

3.1.1. FFA-Fundus

The FFA-Fundus dataset [67] comprises 60 paired images, evenly divided into 30 pairs that depict normal and abnormal cases of fundus images in addition to their corresponding FFA images. Each pair of images is labeled with a binary designation that indicates whether the case is normal or abnormal.

3.1.2. GAMMA

The "Glaucoma Grading from Multimodality Images (GAMMA) Challenge" [42] provides 2D fundus and 3D OCT images of 300 patients tagged with severity of glaucoma. The three class labels used in this multimodal dataset are none, early and mid-advanced, indicating that there are no, early, mid-, or advanced glaucoma severity levels, respectively. Each of the fundus and OCT modalities includes 150, 78, and 72 images for the none, early, and mid-advanced classes, respectively, summing up to 300 fundus and 300 OCT images. In this study, both normal and abnormal cases of glaucoma are included to ensure a balanced class distribution and to facilitate a comparative analysis with other available data sets.



(**a**) Fundus (FFA-Fundus) (**b**) FFA (FFA-Fundus)

(**c**) Fundus (OLIVES) (**d**) OCT Scan (OLIVES)

(**e**) Fundus (GAMMA) (**f**) OCT Scan (GAMMA)

Figure 2. FFA-fundus [67], OLIVES [68] and GAMMA [42] dataset samples.

Table 1. Details about dataset used.

Data Source	Disease Classified	Modality	Image Count		
			Normal	Abnormal	Total
FFA-fundus	DR	FFA	30	30	60
		Fundus	30	30	60
GAMMA	Glaucoma	Fundus	150	150	300
		OCT	150	150	300
OLIVES	DR (NPDR/PDR)	Fundus	125	125	250
		OCT	125	125	250

3.1.3. OLIVES

The "Ophthalmic Labels to Investigate Visual Eye Semantics (OLIVES)" [68] provides fundus and OCT scans of 96 patients, along with binary DR labels that indicate Non-Proliferative DR (NPDR) or Proliferative DR (PDR). DR can be NPDR and PDR forms, depending on whether abnormal new blood vessels are absent or present in the retina [69]. Each patient has one fundus image and 49 OCT scans per visit. Among the 49 scans,

only one was included in the study, as most scans per patient were identical. For binary classification, our study uses 250 NPDR and 250 PDR fundus–OCT image pairs.

Algorithm 1 The proposed algorithm

Inputs:
(A) Unimodal and multimodal retinal images
▷ multimodal public datasets used in this study are FFA-Fundus, GAMMA, and OLIVES
(B) Classification label of each image
▷ available along with datasets
Output: An ocular disease-based trained classification models

Step 1: Image Preprocessing
- Resize the images to 224 × 224

Step 2: Training three models, namely VGG16, ResNet50, SwinT and EfficientNetB0, for ocular disease base classification task
(a) Unimodal using modality mod1
▷ modality mod1: FFA images from FFA-Fundus, Fundus images from GAMMA, OLIVES datasets
- image features are extracted
- classified using VGG16, ResNet50, SwinT and EfficientNetB0

(b) Unimodal using modality mod2
▷ modality mod2: Fundus images from FFA-Fundus, OCT images from GAMMA, OLIVES datasets
- image features are extracted
- classified using VGG16, ResNet50, SwinT and EfficientNetB0

(c) Multimodal using modality mod1 and mod2
(i) early fusion
▷ fusion at the input level before feature extraction
- integration level: input level fusion
- fusion method: concatenation or summation of feature vectors
- advantages: Leverages complementary modality information, richer representation and improved prediction accuracy

(ii) feature fusion
▷ fusion occurs during the feature extraction stage
- integration level: feature-level fusion
- fusion method: concatenation of feature vectors
- advantages: makes use of richer representation, complementary modality information, improved prediction accuracy

(iii) late fusion
▷ fusion occurs at decision level
- integration level: decision-level fusion
- fusion method: averaging predicted class probabilities
- advantages: equal modality contribution, enhanced robustness and precision.

Step 3: The trained classification models are saved based on an early stopping criterion.
Step 4: The best-performing saved models are deployed for inference on unseen test images.

3.2. Multimodal Fusion

This study focuses mainly on investigating three fusion techniques, as illustrated in Figures 3–5, and evaluates their impact on model performance in the classification of ocular diseases. Early fusion, particularly pixel-level fusion, is implemented at the input stage before the images are fed into the model. In this method, two modality RGB images are concatenated along the channel axis, resulting in a six-channel input. The first three channels represent one modality image, while the last three correspond to the second modality image. To ensure compatibility with classification models that require three-channel input, the six-channel representation is reduced to three channels using a 1×1 CNN layer with ReLU activation. This method facilitates efficient channel-wise fusion while reducing the dimensionality. Subsequently, backbone networks are used for the detection of ocular disease. Figure 3 illustrates the model architecture [70] proposed for the early fusion technique, using the VGG pre-trained backbone network. In this figure,

image_mod1 and image_mod2 represent input-modality images from their respective datasets. The custom classification head is trainable and consists of a flat layer following the feature extractor layers. This converts the multidimensional feature maps from the convolution backbone into a one-dimensional vector, making it suitable for processing by fully connected layers. Next, dense layers with ReLU activation are used to learn complex patterns from extracted features, enabling non-linearity and capturing higher-level feature representations. Dropout layers are used to reduce overfitting by randomly deactivating neurons during training and to enhance generalization. Finally, the output layer is a dense layer with sigmoid activation. The loss function is the binary cross-entropy (binary_cross-entropy). The model is trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and $\beta = 0.5$.

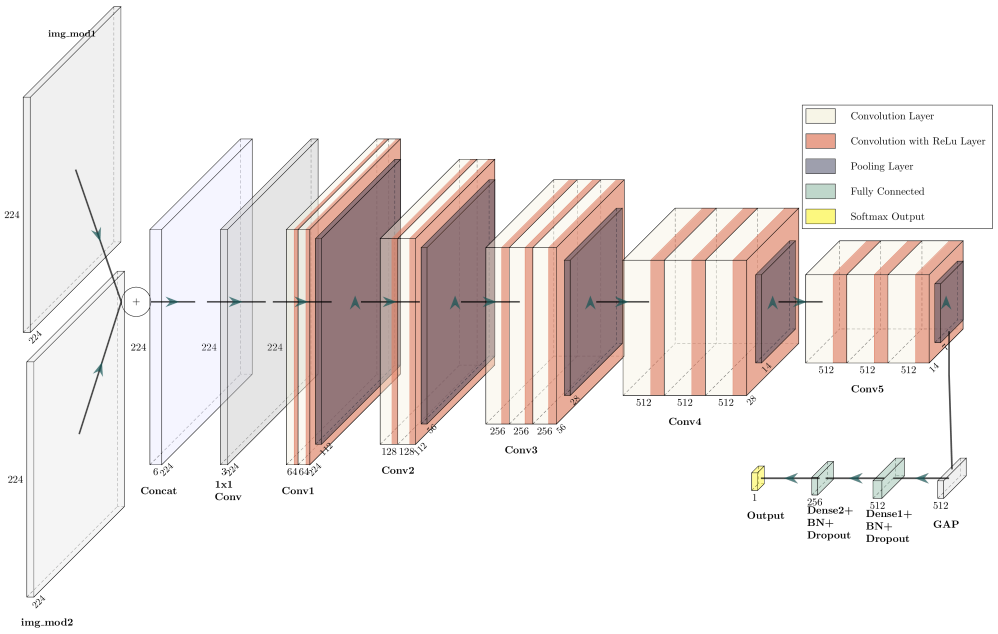


Figure 3. Proposed early fusion technique combining input features and passing them through the CNN network.

Feature fusion, also known as intermediate fusion, takes place in the feature extraction stage. In this study, images from two different input modalities are processed through a shared backbone network, such as VGG16, ResNet50, or EfficientNetB0, SwinT to extract high-level feature representations. Each image is passed through the backbone, resulting in 2D feature maps which are then flattened using a global average pooling layer. Figure 4 illustrates the model architecture proposed for the feature fusion technique, using the VGG pre-trained backbone network.

In addition, multimodal performance is examined through late fusion, where integration occurs at the decision level. In this approach, the predictions of two independent model branches, each handling a different modality, are combined to generate the final classification output. Two separate instances of the backbone network process their respective inputs independently, extracting features and generating distinct predictions. The outputs from both branches are then fused at the decision stage by averaging the predicted class probabilities. This method ensures equal contributions from both modalities, enhancing the robustness of the model and the precision of the classification. Figure 5 illustrates the model architecture proposed for the late fusion technique, using the VGG pre-trained backbone network.

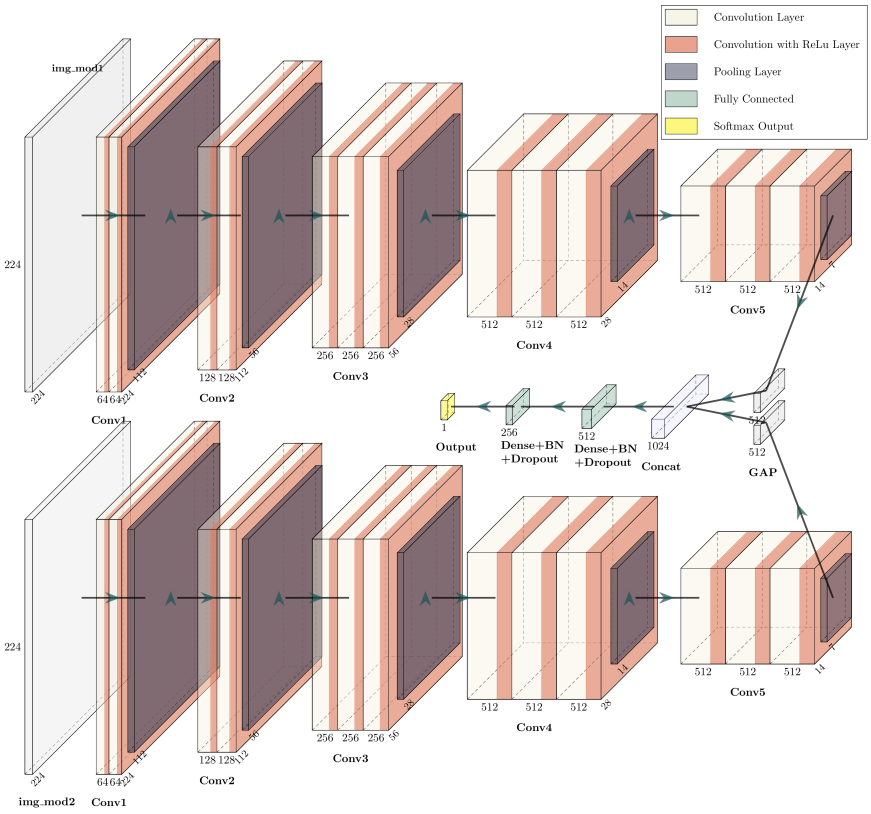


Figure 4. Proposed fusion method integrating modality features before decision-level classification.

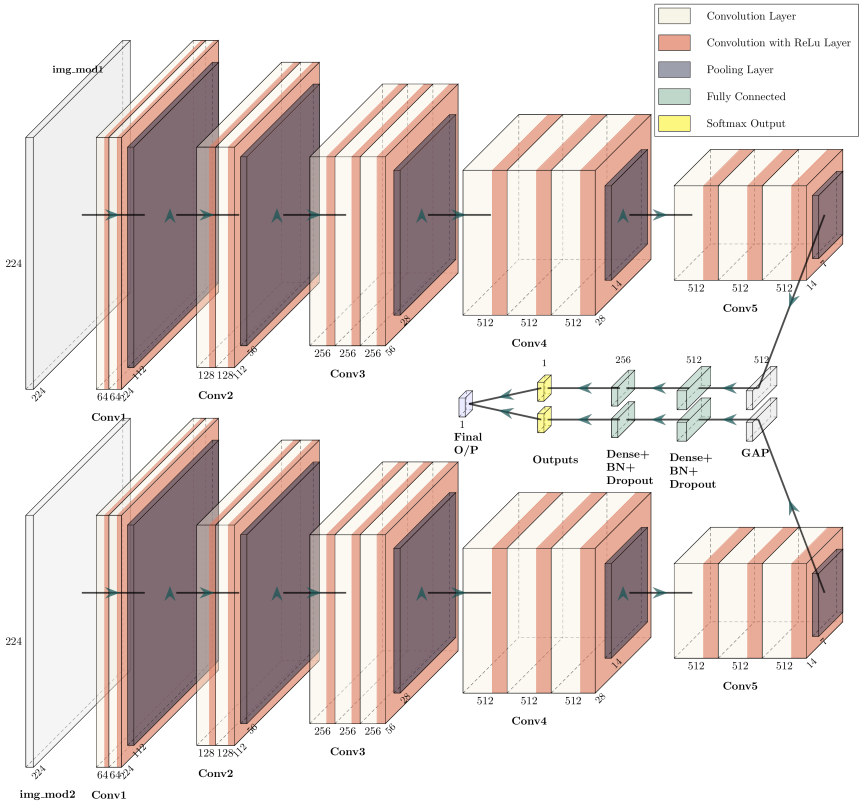


Figure 5. Proposed late fusion approach combining modality-specific networks at the decision level.

The findings of this study provide valuable information on optimal fusion strategies to effectively utilize complementary information from multiple modalities, thus enhancing

diagnostic precision and robustness. In the case of the FFA-Fundus dataset, fundus images were fused with their corresponding FA images to identify normal and abnormal cases. For the OLIVES dataset, fundus images were combined with OCT images to detect NPDR and PDR. Similarly, fundus images and their corresponding OCT counterparts from the GAMMA dataset were subjected to the same fusion techniques and classified as normal or indicative of glaucoma.

4. Results

The experimental setup utilized a high-performance computing (HPC) system configured with CUDA version 12.2, 6912 CUDA cores, driver version 535.183.01, 432 TMUs and 80 GB memory. To ensure a comprehensive evaluation of the proposed models, we used various quantitative metrics in different tasks. This evaluation encompasses image generation and unimodal and multimodal disease classification across three datasets: FFA-Fundus [67], GAMMA [42], and OLIVES [68]. Accuracy represents the proportion of correct predictions to the total predictions generated by the model. The F1 score assesses the performance of the model by calculating the harmonic mean of precision and recall. The Area Under Curve (AUC) score indicates the effectiveness of the model in distinguishing between different classes. Lastly, the Kappa score measures the degree of agreement between the predicted and actual labels, providing insight into their consistency.

Three fusion methods, Early, Late, and Feature Fusion, were studied using multimodal data comprising fundus, OCT, and FA images. The results are summarized in Tables 2–4. To ensure a fair comparison among the experimental techniques, the same random seed is used during training and testing of the proposed models. Multimodal approaches demonstrated improvements over unimodal models when evaluated in different architectures (VGG16, ResNet, and EfficientNet). Figure 6 illustrates the plots of loss and accuracy of training and validation data. The accuracy and loss plots demonstrate that training and validation performance are similar without any significant gaps, which ensures no overfitting and stable learning.

Table 2. Results obtained for FFA-Fundus Dataset [67].

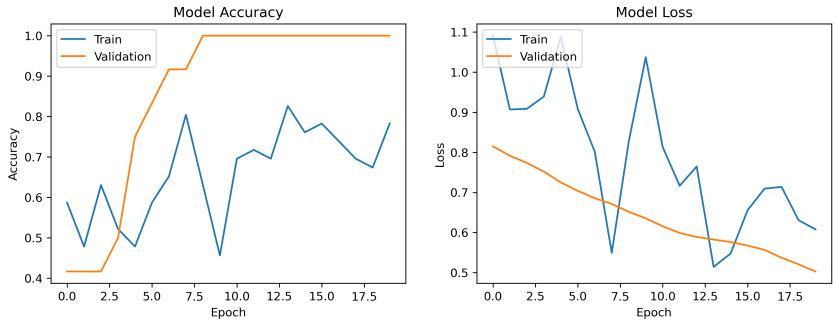
Fusion Type	Model	Test Accuracy	F1 Score	AUC	Kappa Score
Unimodal (FFA)	VGG16	66.67	77.14	33.33	22.58
	ResNet	58.33	71.43	54.55	16.67
	EfficientNet	91.67	100	90.91	83.33
	SwinT	100	100	100	100
Unimodal (Fundus)	VGG16	75	77.15	72.73	50
	ResNet	100	100	100	100
	EfficientNet	66.67	100	71.43	38.46
	SwinT	81.81	75.00	80.00	62.07
Early Fusion	VGG16	66.67	45.71	33.33	22.58
	ResNet	66.67	57.14	33.33	22.58
	EfficientNet	66.67	77.14	66.67	35.14
	SwinT	45.45	62.50	70.00	65.43
Feature Fusion	VGG16	66.67	94.29	71.43	38.46
	ResNet	91.67	94.29	90.91	83.33
	EfficientNet	100	100	100	100
	SwinT	80.82	75.00	70.00	62.07
Late Fusion	VGG16	75.00	91.43	66.67	47.06
	ResNet	75.00	94.29	76.92	52.63
	EfficientNet	91.67	100	88.89	82.35
	SwinT	91.66	88.89	86.66	81.36

Table 3. Results obtained for GAMMA Dataset [42].

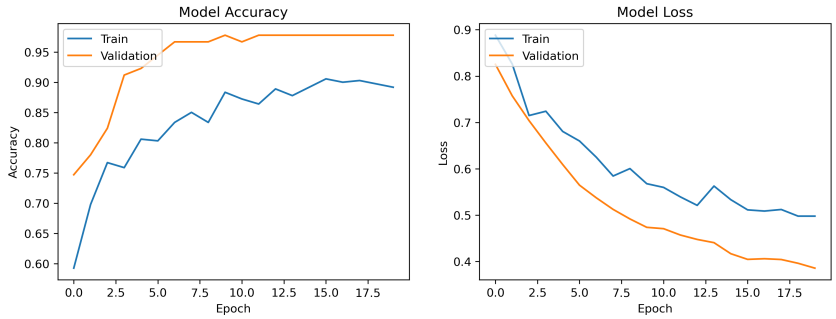
Fusion Type	Model	Test Accuracy	F1 Score	AUC	Kappa Score
Unimodal (Fundus)	VGG16	83.33	91.52	82.76	66.96
	ResNet	86.66	94.19	87.09	73.33
	EfficientNet	85.00	94.87	86.15	69.79
	SwinT	100	100	100	100
Unimodal (OCT)	VGG16	81.67	91.29	83.58	62.92
	ResNet	83.33	91.41	83.87	66.67
	EfficientNet	90	96.88	91.18	79.73
	SwinT	62.71	73.17	62.07	64.45
Early Fusion	VGG16	83.33	91.52	84.85	66.37
	ResNet	75.00	88.62	74.58	50.33
	EfficientNet	83.33	97.32	86.11	65.91
	SwinT	58.33	56.57	59.56	61.67
Feature Fusion	VGG16	93.33	97.32	93.94	86.55
	ResNet	91.66	98.10	92.06	83.29
	EfficientNet	93.33	98.55	93.75	86.61
	SwinT	98.33	98.36	99.78	96.67
Late Fusion	VGG16	93.33	96.32	93.75	86.61
	ResNet	91.66	98.21	91.80	83.37
	EfficientNet	93.33	98.88	94.12	86.49
	SwinT	90.00	89.66	98.89	80.00

Table 4. Results obtained for OLIVES Dataset [68].

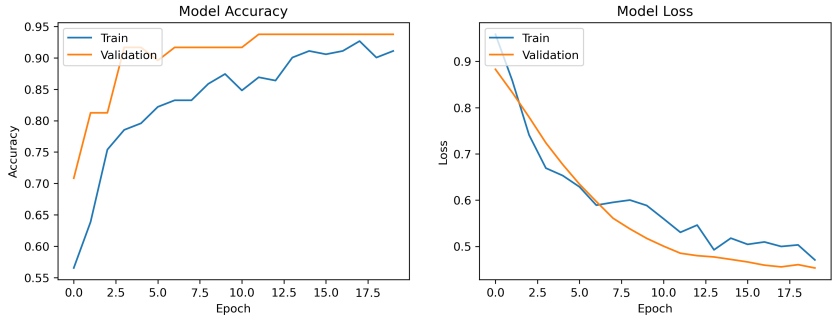
Approach	Model	Test Accuracy	F1 Score	AUC	Kappa Score
Unimodal (Fundus)	VGG16	94.05	95.22	94.44	88.07
	ResNet50	94.05	97.79	94.44	88.07
	EfficientNet	93.06	96.48	93.33	86.15
	SwinT	96.00	93.33	96.00	94.00
Unimodal (OCT)	VGG16	71.29	83.52	60.27	41.66
	ResNet	82.18	92.15	83.02	64.49
	EfficientNet	87.13	94.11	85.39	74.09
	SwinT	52.00	52.00	61.29	59.00
Early Fusion	VGG16	79.20	93.72	81.74	58.79
	ResNet50	89.10	94.66	88.89	78.21
	EfficientNet	94.05	95.06	93.62	88.08
	SwinT	66.00	73.02	74.08	61.99
Feature Fusion	VGG16	94.05	94.58	93.75	88.09
	ResNet50	94.05	95.09	93.62	88.08
	EfficientNet	95.04	95.41	94.74	90.07
	SwinT	54.00	64.62	71.36	71.99
Late Fusion	VGG16	95.04	95.80	94.62	90.06
	ResNet50	93.06	95.49	92.63	86.10
	EfficientNet	96.03	96.59	95.74	92.05
	SwinT	50.00	60.32	79.84	72.71



(**a**) FFA-Fundus feature fusion



(**b**) OLIVES late fusion



(**c**) GAMMA late fusion

Figure 6. Plots of accuracy and loss of training data.

5. Discussion

Three fusion techniques are evaluated in multiple ocular diseases. Among the three fusion strategies, feature fusion consistently outperformed the other two methods in all metrics (see Tables 2–4). These results reflect the benefits of combining complementary modalities to improve diagnostic precision. For comparative analysis among F1 score, accuracy, AUC, and kappa score, the kappa score was selected as the primary metric. Unlike accuracy and the F1 score, which can be affected by class imbalance, the kappa score is chosen because the score accounts for the agreement occurring by chance, providing a more robust measure of model performance. Among unimodal DL models, the fundus-only SwinT approach showed superior performance, capturing key features such as microaneurysms and exudates. Among the evaluated pre-trained models, the SwinT achieved the best performance in the unimodal (Fundus) classification task. This superiority can be attributed to its ability to model long-range dependencies through self-attention while preserving hierarchical feature representations. In contrast to CNN-based architectures, which depend on fixed local receptive fields, the SwinT captures local and global contextual information via shifted window attention, enabling a more expressive representation of complex spatial patterns. However, its performance declines in multimodal settings. This

limitation is mainly due to the fact that SwinT is optimized primarily for the hierarchical visual structure of a single modality; introducing a weaker or misaligned second modality can disrupt its inductive bias and hinder learning. For most of the other models, multimodal fusion approaches achieved superior performance compared to unimodal models, underscoring the advantage of leveraging complementary information across modalities.

Multimodal fusion techniques outperformed unimodal models, highlighting the advantage of combining complementary information between two modalities. In the FFA dataset, the performance matched the fundus-only model, likely due to limited test images. Both late fusion with EfficientNet and feature-level fusion achieved comparable results, indicating their ability to capture modality-specific features effectively. In contrast to VGG's deep yet uniform architecture and ResNet's dependence on residual connections, EfficientNet employs a compound scaling method that harmonizes the depth, width, and resolution of the network. This approach enhances the model's ability to learn effectively from diverse and complementary features present in multimodal data. Furthermore, its lightweight design mitigates the risks of overfitting, particularly when integrating modalities with varying feature distributions, making it particularly well suited for complex fusion tasks. However, early fusion did not yield significant performance gains, possibly due to suboptimal integration of heterogeneous feature spaces at the input level.

Table 5 presents the performance benchmarking results in comparison with existing methods. The proposed fusion technique achieves performance on par with baseline approaches while maintaining a simpler architectural design. Although many existing methods involve multi-stage pipelines for lesion or biomarker identification, the proposed framework employs a single training phase using fused multimodal features. This streamlined design improves computational efficiency and enhances practical applicability. Furthermore, prior approaches often rely on complex multi-branch networks, ensemble strategies, or expert-driven biomarker annotations, which increase model complexity and limit scalability. In contrast, the proposed model is lightweight, efficient, and more generalizable, making it better suited for real-world clinical deployment without requiring specialized infrastructure or extensive manual intervention.

Table 5. Performance benchmarking with existing work.

Dataset	Existing Work	Performance (Metric)	Remarks
FFA-Fundus	Hajeb et al. [67]	100 (Accuracy)	Employs five lesion segmentation to determine the final classification;
GAMMA	Wu et al. [42]	82.80 $\pm$ 0.03 (Kappa Score)	Achieves an accuracy of 79.55% by employing a 3D network for the OCT branch and a DENet architecture for the fundus branch for three-class classification. The method further incorporates ensemble strategies, which further increases the number of training parameters.
OLIVES	Prabhushankar et al. [68]	82.33% $\pm$ 3.59 (Balanced/Weighted accuracy)	Employs a data split of 70% for training, 20% for validation, and 10% for testing in the detection of DR/DME using OCT and biomarker inputs. However, accurately identifying these biomarkers is time-consuming and requires expert knowledge.

To further analyze the number of trainable parameters and the interpretation of the model, we used the top-performing model per dataset. Table 6 lists the number of parameters required for these top-performing models. It can be observed that multimodal networks require a larger number of parameters, which results in longer training and

inference times. However, instead of training two separate networks for each modality, both modalities can be integrated within a fusion network.The last column reports the inference time for a single image using the unimodal EfficientNet and for a pair of images using the multimodal EfficientNet. These experiments were conducted on a system configured with CUDA version 12.2, using an NVIDIA A100 80 GB PCIe GPU. CUDA cores, 432 TMUs, and with 80 GB of memory.

From Table 6, it can be observed that late and feature fusion strategies incur higher inference time compared to unimodal and early fusion approaches, due to the additional computational stages required for separate feature extraction and subsequent aggregation. Notably, GPU inference time does not scale linearly with the number of trainable parameters, particularly when comparing CNN- and SwinT-based architectures. This behavior can be attributed to architectural design and the degree of computational parallelism, as transformer-based operations are highly optimized for large-scale matrix multiplications that efficiently utilize GPU cores. However, deployment of SwinT models in general clinical settings remains challenging, as CPU-based inference results in substantially higher latency due to the computational intensity of attention mechanisms. In contrast, EfficientNet-based models demonstrate more stable and practical performance on CPU systems. Therefore, considering both deployment feasibility and runtime performance, feature- and late-fusion EfficientNet architectures provide a more suitable balance for real-world clinical applications.

Table 6. Trainable parameters and inference time of unimodal and multimodal models.

Dataset	Fusion Type	Trainable Parameters	Multiplicative Factor (×)	Inference Time (ms)	Inference Time (ms) Under Clinical Setup
FFA Fundus	Unimodal CNN	788,993	–	105.75 ± 10.34	423.00 ± 621.00
	Early Fusion CNN	789,014	1.00	201.83 ± 7.31	541.13 ± 110.55
	Feature Fusion CNN	1,444,353	1.83	391.91 ± 9.68	1058.40 ± 178.56
	Late Fusion CNN	1,577,986	2.00	397.39 ± 10.60	1054.21 ± 167.73
	Unimodal SwinT	87,005,881	110.27	22.66 ± 1.67	1122.35 ± 555.64
	Early Fusion SwinT	87,005,902	110.27	32.69 ± 10.53	6443.00 ± 2776.47
	Feature Fusion SwinT	174,536,049	221.24	51.69 ± 9.67	$14,360.21 \pm 2210.62$
	Late Fusion SwinT	173,488,498	219.92	52.39 ± 15.23	$29,211.38 \pm 3147.46$

To understand the effect of fusion on model interpretation, we employed Grad-CAM to visualize the key regions learned by the model. The visualization of the output heatmap for a sample input is shown in Figure 7. From the visualizations, it can be seen that both unimodal and multimodal networks learn similar regions. For fundus input images, the lesions are generally identified correctly, while this is not consistently the case for OCT images, particularly in the GAMMA dataset. This discrepancy could be attributed to the large noise regions present in OCT images. In contrast, for the OLIVES dataset, the OCT layers are better captured due to layer-focused ROI inputs.

For the FFA-Fundus images (refer Figure 7A,B), multimodal fusion helps to restrict the learned regions more precisely within the fundus region compared to using single-modality images. Both single and multimodal networks identify similar lesions (e.g., exudates, red lesions, microaneurysms). In the case of GAMMA, the optic disc regions are learned correctly from fundus images by both networks. The late fusion of both modalities does not significantly alter the regions learned in this case. In the OLIVES dataset, the OCT layer regions are correctly learned by the network. Again, late fusion of both modalities does not lead to substantial differences compared to unimodal training. Overall, the results suggest that the regions learned by multimodal fusion networks are largely consistent with those learned by individual unimodal networks, with only dataset-specific variations.

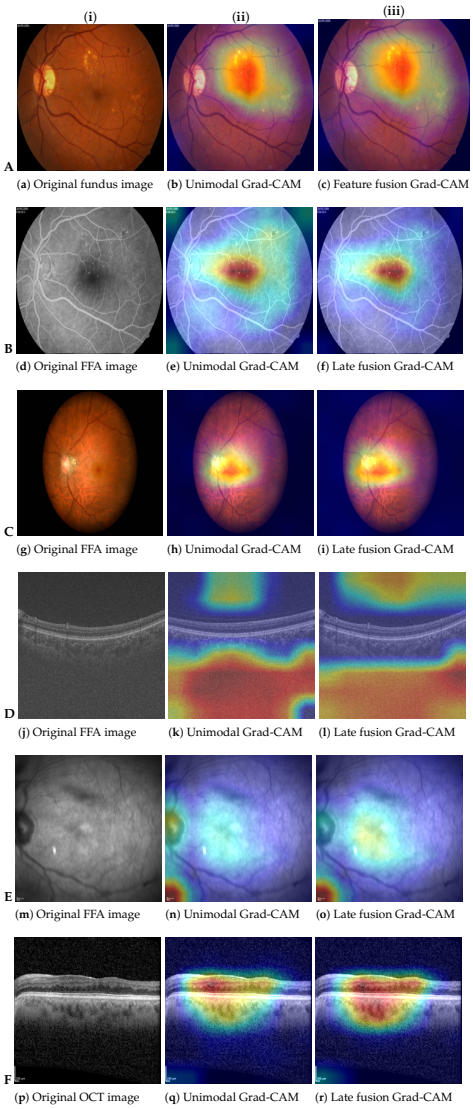


Figure 7. Visualization of heatmaps overlaid on input images (red to blue, denoting high to low importance). Rows **A** and **B**, **C** and **D**, and **E** and **F** present sample images from the FFA-Fundus, GAMMA, OLIVES datasets respectively. Column (**i**) presents the original images, column (**ii**) shows the Grad-CAM representations generated from the unimodal network, while column (**iii**) displays the Grad-CAM representations obtained from the multimodal network that leverages both input modalities.

To quantitatively assess clinical localization consistency, 81 diabetic retinopathy (DR) fundus images from the "Indian Diabetic Retinopathy Image Dataset" (IDRiD) dataset [71] with expert-verified lesion annotations were utilized. These images included clinically relevant features such as soft exudates, hard exudates, microaneurysms, and hemorrhages. Pixel-level lesion masks were used as ground truth, and morphological processing was applied to aggregate scattered annotations into meaningful regions of interest. Correspondingly, Grad-CAM heatmaps were thresholded to generate binary localization maps for quantitative comparison with expert annotations. Grad-CAM–derived localization maps showed substantial agreement with expert-annotated lesion masks on correctly detected DR images, achieving an average Dice coefficient of 0.6327 and IoU of 0.4723 on unseen IDRiD images. Representative examples of original fundus images, expert annotations, and Grad-CAM–identified regions are shown in Figure 8, illustrating consistent localization across diverse imaging conditions and lesion types. While prominent lesions were accurately localized, reduced sensitivity was observed in mild DR cases (refer Figure 8E) with subtle microaneurysms and hemorrhages, likely due to their limited representation in the

training data. These findings highlight the model's ability to accurately identify clinically relevant pathological regions and indicate that further fine-tuning on larger and more diverse multimodal datasets could enhance sensitivity to early DR lesions and improve overall localization performance.

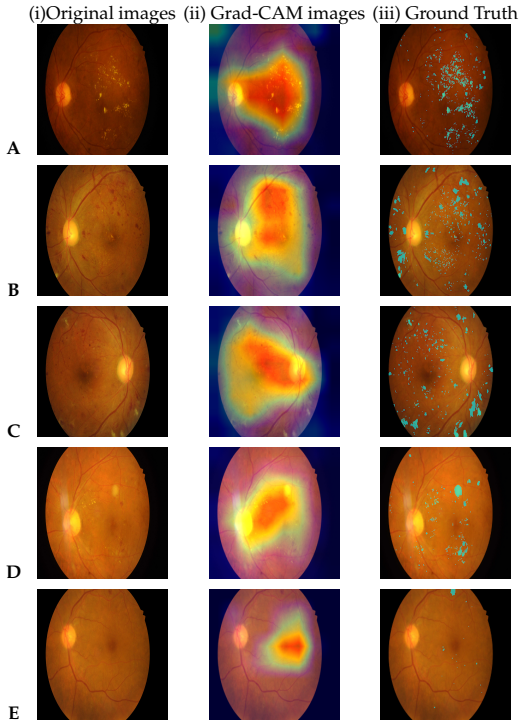


Figure 8. Visualization of Grad-CAM heatmaps (red to blue, denoting high to low importance) and expert-annotated lesion masks on IDRiD fundus images from the segmentation folder. Images (**A**–**D**) represent DR fundus images with more visible lesions (microaneurysms, haemorrhages, and soft and hard exudates), while image (**E**) represents a DR fundus image with very few visible lesions.

In addition, to understand the generability of the best performing multimodal fusion networks, EfficientNet is tested on another publicly available dataset from the GitHub repository Traslational-Visual-Health-Laboratory (TVHL) [72]. This dataset provides fundus and OCT image pairs, DR disease labels are supplied in a CSV file. Although fundus and OCT images along with DR labels are provided in a large number, as image pairs fewer images are available. Thus, the number of training images used is 500 (250 normal and 250 DR images). For validation, 50 normal and 50 DR images are used. The results of the EfficientNet model using the TVHL dataset are shown in Table 7. Similar results were obtained with unimodal fundus imaging contributing to the performance of DR detection, while among multimodal fusion techniques, late fusion showed superior performance. From the overall experimentation, the following conclusions can be drawn.

- Unimodal models using a SwinT with fundus images alone achieved better performance than their multimodal counterparts.
- CNN-based models performed significantly better with late fusion and feature-level fusion compared to early fusion techniques.
- The interpretability of the model was found to be modality- and architecture-specific. The fusion of modalities within the same network does not necessarily alter the Grad-CAM visualizations.

- The average percentage increase in the inference time for multimodal fusion was up to 50%, while the average percentage increase in the number of parameters was approximately 75% due to fusion.
- Multimodal fusion was more efficient than training separate models for each modality in terms of both overall training and inference time.

Table 7. Efficient model performance using TVHL dataset.

Model	Accuracy	AUC	F1 Score	Kappa Score
Unimodal (Fundus)	88.89	95.51	88.66	77.79
Unimodal (OCT)	69.70	64.98	62.50	69.63
Early fusion	43.43	42.45	48.15	43.33
Intermediate fusion	74.75	71.14	66.67	69.75
Late fusion	78.79	85.10	76.4	76.66

Although this study provides valuable information on the comprehensive evaluation of fusion techniques to recognize ocular diseases, several inherent limitations must be recognized. Although a comprehensive analysis has been performed to evaluate the impact of multimodal fusion, the limited availability of publicly accessible data has confined this analysis to detection tasks at present. However, with sufficient data availability, these techniques could easily be adapted for severity classification tasks with only minor modifications. While our findings indicate that multimodal fusion techniques enhance disease recognition, the improvement is marginal, and further validation is required to establish their advantage over unimodal approaches across different datasets and clinical settings. Additionally, the fusion mechanisms explored in this study are widely used conventional fusion techniques. More advanced cross-modal fusion strategies, such as attention-based and transformer-based methods, could be further investigated to enhance cross-modal interaction and improve feature representation. It is also important to note that the proposed framework is primarily designed for scenarios where multimodal imaging data are accessible, and its applicability may be limited in situations where such data are not routinely collected. Future research can explore methods for generalizing this approach to broader clinical applications and optimizing fusion strategies to enhance diagnostic accuracy.

6. Conclusions

This study illustrates the effectiveness of multimodal fusion techniques in improving automated recognition of ocular diseases. Experimental evaluations conducted on publicly available datasets, including FFA-Fundus, GAMMA, and OLIVES, demonstrate that integrating various imaging modalities through intermediate and late fusion strategies improves ocular detection performance. These findings underscore the potential of multimodal ophthalmic imaging in improving diagnostic precision. Future research may focus on generating the necessary modality images through image synthesis, refining fusion strategies, and expanding this approach to broader clinical applications for improved real-world implementation.

Author Contributions: Investigation: V.K.M. and V.M.; Conceptualization: All authors; Methodology: V.K.M., P.G., R.A. and V.M.; Supervision: R.N.R., S.V.B. and V.M.; Writing—original draft: V.K.M., P.G., R.A. and V.M.; Writing—review and editing: All authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in [42,67,68,71,72].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. August 2023. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 23 February 2025).
2. Ringel, M.J.; Tang, E.M.; Tao, Y.K. Advances in multimodal imaging in ophthalmology. *Ther. Adv. Ophthalmol.* **2021**, *13*, 1–12. [CrossRef]
3. Xu, X.; Li, J.; Zhu, Z.; Zhao, L.; Wang, H.; Song, C.; Chen, Y.; Zhao, Q.; Yang, J.; Pei, Y. A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering* **2024**, *11*, 219. [CrossRef] [PubMed]
4. Veena, K.; Tummala, V.; Sangaraju, Y.S.V.; Reddy, M.S.V.; Kumar, P.; Mayya, V.; Kulkarni, U.; Bhandary, S.; Shailaja, S. FFA-Lens: Lesion detection tool for chronic ocular diseases in Fluorescein angiography images. *SoftwareX* **2024**, *26*, 101646. [CrossRef]
5. Mayya, V.; Kandala, R.N.; Gurupur, V.; King, C.; Vu, G.T.; Wan, T.T. Need for an Artificial Intelligence-based Diabetes Care Management System in India and the United States. *Health Serv. Res. Manag. Epidemiol.* **2024**, *11*, 1–12. [CrossRef] [PubMed]
6. Martinc, M.; Haider, F.; Pollak, S.; Luz, S. Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. *Front. Aging Neurosci.* **2021**, *13*, 642647. [CrossRef]
7. Ning, Z.; Xiao, Q.; Feng, Q.; Chen, W.; Zhang, Y. Relation-induced multi-modal shared representation learning for Alzheimer's disease diagnosis. *IEEE Trans. Med. Imaging* **2021**, *40*, 1632–1645. [CrossRef]
8. Shi, J.; Zheng, X.; Li, Y.; Zhang, Q.; Ying, S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 173–183. [CrossRef]
9. Zhou, T.; Liu, M.; Thung, K.H.; Shen, D. Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. *IEEE Trans. Med. Imaging* **2019**, *38*, 2411–2422. [CrossRef]
10. Cai, H.; Huang, X.; Liu, Z.; Liao, W.; Dai, H.; Wu, Z.; Zhu, D.; Ren, H.; Li, Q.; Liu, T.; et al. Exploring Multimodal Approaches for Alzheimer's Disease Detection Using Patient Speech Transcript and Audio Data. *arXiv* **2023**, arXiv:2307.02514.
11. Park, C.; Ha, J.; Park, S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Syst. Appl.* **2020**, *140*, 112873. [CrossRef]
12. Golovanevsky, M.; Eickhoff, C.; Singh, R. Multimodal attention-based deep learning for Alzheimer's disease diagnosis. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 2014–2022. [CrossRef] [PubMed]
13. Huang, Y.; Zeng, P.; Zhong, C. Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC Bioinform.* **2024**, *25*, 132. [CrossRef] [PubMed]
14. Choi, J.M.; Chae, H. moBRCA-net: A breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC Bioinform.* **2023**, *24*, 169. [CrossRef]
15. Parshionikar, S.; Bhattacharyya, D. An enhanced multi-scale deep convolutional orchard capsule neural network for multi-modal breast cancer detection. *Healthc. Anal.* **2024**, *5*, 100298. [CrossRef]
16. Guo, D.; Lu, C.; Chen, D.; Yuan, J.; Duan, Q.; Xue, Z.; Liu, S.; Huang, Y. A multimodal breast cancer diagnosis method based on Knowledge-Augmented Deep Learning. *Biomed. Signal Process. Control* **2024**, *90*, 105843. [CrossRef]
17. Yan, R.; Zhang, F.; Rao, X.; Lv, Z.; Li, J.; Zhang, L.; Liang, S.; Li, Y.; Ren, F.; Zheng, C.; et al. Richer fusion network for breast cancer classification based on multimodal data. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 134. [CrossRef]
18. Arya, N.; Saha, S. Multi-modal advanced deep learning architectures for breast cancer survival prediction. *Knowl.-Based Syst.* **2021**, *221*, 106965. [CrossRef]
19. Sun, D.; Wang, M.; Li, A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *16*, 841–850. [CrossRef]
20. Tong, L.; Mitchel, J.; Chatlin, K.; Wang, M.D. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 225. [CrossRef] [PubMed]
21. Liu, T.; Huang, J.; Liao, T.; Pu, R.; Liu, S.; Peng, Y. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM* **2022**, *43*, 62–74. [CrossRef]
22. Liu, H.; Shi, Y.; Li, A.; Wang, M. Multi-modal fusion network with intra-and inter-modality attention for prognosis prediction in breast cancer. *Comput. Biol. Med.* **2024**, *168*, 107796. [CrossRef]
23. Dong, Y.; Yang, X. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing* **2021**, *441*, 279–290. [CrossRef]
24. Ehghaghi, M.; Rudzicz, F.; Novikova, J. Data-driven approach to differentiating between depression and dementia from noisy speech and language data. *arXiv* **2022**, arXiv:2210.03303. [CrossRef]

25. Diep, B.; Stanojevic, M.; Novikova, J. Multi-modal deep learning system for depression and anxiety detection. *arXiv* **2022**, arXiv:2212.14490. [CrossRef]
26. Mao, K.; Zhang, W.; Wang, D.B.; Li, A.; Jiao, R.; Zhu, Y.; Wu, B.; Zheng, T.; Qian, L.; Lyu, W.; et al. Prediction of depression severity based on the prosodic and semantic features with bidirectional LSTM and time distributed CNN. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2251–2265. [CrossRef]
27. Jan, A.; Meng, H.; Gaus, Y.F.B.A.; Zhang, F. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Trans. Cogn. Dev. Syst.* **2017**, *10*, 668–680. [CrossRef]
28. Uddin, M.A.; Joolee, J.B.; Sohn, K.A. Deep multi-modal network based automated depression severity estimation. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2153–2167. [CrossRef]
29. Yang, L.; Jiang, D.; Xia, X.; Pei, E.; Oveneke, M.C.; Sahli, H. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 53–59.
30. Kunekar, P.R.; Gupta, M.; Agarwal, B. Deep learning with multi modal ensemble fusion for epilepsy diagnosis. In *Proceedings of the 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*; IEEE: Jaipur, India, 2020; pp. 80–84.
31. Ahmedt-Aristizabal, D.; Fookes, C.; Denman, S.; Nguyen, K.; Fernando, T.; Sridharan, S.; Dionisio, S. A hierarchical multimodal system for motion analysis in patients with epilepsy. *Epilepsy Behav.* **2018**, *87*, 46–58. [CrossRef] [PubMed]
32. Pominova, M.; Artemov, A.; Sharaev, M.; Kondrateva, E.; Bernstein, A.; Burnaev, E. Voxelwise 3d convolutional and recurrent neural networks for epilepsy and depression diagnostics from structural and functional mri data. In *Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*; IEEE: New York, NY, USA, 2018; pp. 299–307.
33. Torres-Velázquez, M.; Hwang, G.; Cook, C.J.; Hermann, B.; Prabhakaran, V.; Meyerand, M.E.; McMillan, A.B. Multi-channel deep neural network for temporal lobe epilepsy classification using multimodal mri data. In *Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)*; IEEE: New York, NY, USA, 2020; pp. 1–4.
34. Dong, D.; Fu, G.; Li, J.; Pei, Y.; Chen, Y. An unsupervised domain adaptation brain CT segmentation method across image modalities and diseases. *Expert Syst. Appl.* **2022**, *207*, 118016. [CrossRef]
35. Sun, K.; Zhang, Y.; Liu, J.; Yu, L.; Zhou, Y.; Xie, F.; Guo, Q.; Zhang, H.; Wang, Q.; Shen, D. Achieving multi-modal brain disease diagnosis performance using only single-modal images through generative AI. *Commun. Eng.* **2024**, *3*, 96. [CrossRef]
36. Venugopalan, J.; Tong, L.; Hassanzadeh, H.R.; Wang, M.D. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci. Rep.* **2021**, *11*, 3254. [CrossRef]
37. Qiu, S.; Miller, M.I.; Joshi, P.S.; Lee, J.C.; Xue, C.; Ni, Y.; Wang, Y.; De Anda-Duran, I.; Hwang, P.H.; Cramer, J.A.; et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat. Commun.* **2022**, *13*, 3404. [CrossRef]
38. Yoon, J.M.; Yoo, H.; Jung, S.H.; Lee, J.H.; Ham, D.I. Multimodal deep learning-based prediction of response to anti-vascular endothelial growth factor treatment in diabetic macular edema. *Investig. Ophthalmol. Vis. Sci.* **2024**, *65*, 2360.
39. Parravano, M.; Cennamo, G.; Di Antonio, L.; Grassi, M.O.; Lupidi, M.; Rispoli, M.; Savastano, M.C.; Veritti, D.; Vujosevic, S. Multimodal imaging in diabetic retinopathy and macular edema: An update about biomarkers. *Surv. Ophthalmol.* **2024**, *69*, 893–904. [CrossRef] [PubMed]
40. Li, Y.; El Habib Daho, M.; Conze, P.H.; Al Hajj, H.; Bonnin, S.; Ren, H.; Manivannan, N.; Magazzeni, S.; Tadayoni, R.; Cochener, B.; et al. Multimodal information fusion for glaucoma and diabetic retinopathy classification. In *Proceedings of the International Workshop on Ophthalmic Medical Image Analysis*; Springer: Singapore, 2022; pp. 53–62.
41. Huang, X.; Sun, J.; Gupta, K.; Montesano, G.; Crabb, D.P.; Garway-Heath, D.F.; Brusini, P.; Lanzetta, P.; Oddone, F.; Turpin, A.; et al. Detecting glaucoma from multi-modal data using probabilistic deep learning. *Front. Med.* **2022**, *9*, 923096. [CrossRef] [PubMed]
42. Wu, J.; Fang, H.; Li, F.; Fu, H.; Lin, F.; Li, J.; Huang, Y.; Yu, Q.; Song, S.; Xu, X.; et al. Gamma challenge: Glaucoma grading from multi-modality images. *Med. Image Anal.* **2023**, *90*, 102938. [CrossRef]
43. Li, Y.; Han, Y.; Li, Z.; Zhong, Y.; Guo, Z. A transfer learning-based multimodal neural network combining metadata and multiple medical images for glaucoma type diagnosis. *Sci. Rep.* **2023**, *13*, 12076. [CrossRef] [PubMed]
44. El Habib Daho, M.; Li, Y.; Zeghlache, R.; Atse, Y.C.; Le Boité, H.; Bonnin, S.; Cosette, D.; Deman, P.; Borderie, L.; Lepicard, C.; et al. Improved automatic diabetic retinopathy severity classification using deep multimodal fusion of UWF-CFP and OCTA images. In *Proceedings of the International Workshop on Ophthalmic Medical Image Analysis*; Springer: Cham, Switzerland, 2023; pp. 11–20.
45. Li, Y.; El Habib Daho, M.; Conze, P.H.; Zeghlache, R.; Le Boité, H.; Bonnin, S.; Cosette, D.; Magazzeni, S.; Lay, B.; Le Guilcher, A.; et al. Hybrid fusion of high-resolution and ultra-widefield OCTA acquisitions for the automatic diagnosis of diabetic retinopathy. *Diagnostics* **2023**, *13*, 2770. [CrossRef]
46. Xiong, J.; Li, F.; Song, D.; Tang, G.; He, J.; Gao, K.; Zhang, H.; Cheng, W.; Song, Y.; Lin, F.; et al. Multimodal machine learning using visual fields and peripapillary circular OCT scans in detection of glaucomatous optic neuropathy. *Ophthalmology* **2022**, *129*, 171–180. [CrossRef]

47. Lin, Y.T.; Zhou, Q.; Tan, J.; Tao, Y. Multimodal and multi-omics-based deep learning model for screening of optic neuropathy. *Heliyon* **2023**, *9*, e22244. [CrossRef]
48. Cao, Z.; Sun, C.; Wang, W.; Zheng, X.; Wu, J.; Gao, H. Multi-modality fusion learning for the automatic diagnosis of optic neuropathy. *Pattern Recognit. Lett.* **2021**, *142*, 58–64. [CrossRef]
49. El-Ateif, S.; Idri, A. Multimodality Fusion Strategies in Eye Disease Diagnosis. *J. Imaging Inform. Med.* **2024**, *37*, 2524–2558. [CrossRef]
50. Boulahia, S.Y.; Amamra, A.; Madi, M.R.; Daikh, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach. Vis. Appl.* **2021**, *32*, 121. [CrossRef]
51. Waqas, A.; Tripathi, A.; Ramachandran, R.P.; Stewart, P.A.; Rasool, G. Multimodal data integration for oncology in the era of deep neural networks: A review. *Front. Artif. Intell.* **2024**, *7*, 1408843. [CrossRef]
52. Guarrasi, V.; Aksu, F.; Caruso, C.M.; Di Feola, F.; Rofena, A.; Ruffini, F.; Soda, P. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *arXiv* **2024**, arXiv:2408.02686. [CrossRef]
53. Pereira, L.M.; Salazar, A.; Vergara, L. On comparing early and late fusion methods. In *Proceedings of the International Work-Conference on Artificial Neural Networks*; Springer: Cham, Switzerland, 2023; pp. 365–378.
54. Hervella, Á.S.; Rouco, J.; Novo, J.; Ortega, M. Multimodal image encoding pre-training for diabetic retinopathy grading. *Comput. Biol. Med.* **2022**, *143*, 105302. [CrossRef] [PubMed]
55. Wardhani, K.D.K.; Kasim, S.; Erianda, A.; Hassan, R. Deep Learning-based Method in Multimodal Data for Diabetic Retinopathy Detection. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2024**, *14*, 1602–1608. [CrossRef]
56. Bidwai, P.; Gite, S.; Pahuja, N.; Pahuja, K.; Kotecha, K.; Jain, N.; Ramanna, S. Multimodal image fusion for the detection of diabetic retinopathy using optimized explainable AI-based Light GBM classifier. *Inf. Fusion* **2024**, *111*, 102526. [CrossRef]
57. An, G.; Omodaka, K.; Hashimoto, K.; Tsuda, S.; Shiga, Y.; Takada, N.; Kikawa, T.; Yokota, H.; Akiba, M.; Nakazawa, T. Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images. *J. Healthc. Eng.* **2019**, *2019*, 4061313. [CrossRef]
58. Mehta, P.; Petersen, C.A.; Wen, J.C.; Banitt, M.R.; Chen, P.P.; Bojikian, K.D.; Egan, C.; Lee, S.I.; Balazinska, M.; Lee, A.Y.; et al. Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. *Am. J. Ophthalmol.* **2021**, *231*, 154–169. [CrossRef]
59. Lim, W.S.; Ho, H.Y.; Ho, H.C.; Chen, Y.W.; Lee, C.K.; Chen, P.J.; Lai, F.; Jang, J.S.R.; Ko, M.L. Use of multimodal dataset in AI for detecting glaucoma based on fundus photographs assessed with OCT: Focus group study on high prevalence of myopia. *BMC Med. Imaging* **2022**, *22*, 206. [CrossRef]
60. Oh, S.; Park, Y.; Cho, K.J.; Kim, S.J. Explainable Machine Learning Model for Glaucoma Diagnosis and Its Interpretation. *Diagnostics* **2021**, *11*, 510. [CrossRef]
61. Liu, T.; Liu, Y.; Gastonguay, M.S.; Midgett, D.; Kuo, N.; Zhao, Y.; Ullah, K.; Alexander, G.; Hartman, T.; Koseoglu, N.D.; et al. Predicting Imminent Conversion to Exudative Age-Related Macular Degeneration Using Multimodal Data and Ensemble Machine Learning. *Ophthalmol. Sci.* **2025**, *5*, 100785. [CrossRef]
62. Yang, C.N.; Hsieh, Y.T.; Yeh, H.H.; Chu, H.S.; Wu, J.H.; Chen, W.L. Prediction of Visual Acuity After Cataract Surgery by Deep Learning Methods Using Clinical Information and Color Fundus Photography. *Curr. Eye Res.* **2025**, *50*, 276–281. [CrossRef]
63. Dong, Z.; Wang, X.; Pan, S.; Weng, T.; Chen, X.; Jiang, S.; Li, Y.; Wang, Z.; Cao, X.; Wang, Q.; et al. A multimodal transformer system for noninvasive diabetic nephropathy diagnosis via retinal imaging. *npj Digit. Med.* **2025**, *8*, 50. [CrossRef]
64. Jin, K.; Yan, Y.; Chen, M.; Wang, J.; Pan, X.; Liu, X.; Liu, M.; Lou, L.; Wang, Y.; Ye, J. Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration. *Acta Ophthalmol.* **2021**, *100*, e512–e520. [CrossRef] [PubMed]
65. Liu, T.Y.A.; Ling, C.; Hahn, L.; Jones, C.K.; Boon, C.J.; Singh, M.S. Prediction of visual impairment in retinitis pigmentosa using deep learning and multimodal fundus images. *Br. J. Ophthalmol.* **2022**, *107*, 1484–1489. [CrossRef] [PubMed]
66. Hervella, Á.S.; Rouco, J.; Novo, J.; Ortega, M. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Syst. Appl.* **2021**, *185*, 115598. [CrossRef]
67. Hajeb Mohammad Alipour, S.; Rabbani, H.; Akhlaghi, M.R. Diabetic Retinopathy Grading by Digital Curvelet Transform. *Comput. Math. Methods Med.* **2012**, *2012*, 761901. [CrossRef]
68. Prabhushankar, M.; Kokilepersaud, K.; Logan, Y.-y.; Corona, S.T.; AlRegib, G.; Wykoff, C. OLIVES Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics. Dataset. 2022. Available online: https://zenodo.org/records/7105232 (accessed on 24 April 2025).
69. Chong, D.D.; Das, N.; Singh, R.P. Diabetic retinopathy: Screening, prevention, and treatment. *Clevel. Clin. J. Med.* **2024**, *91*, 503–510. [CrossRef]
70. Iqbal, H. HarisIqbal88/PlotNeuralNet v1.0.0, 2018. Dataset, Version v1.0.0. Available online: https://zenodo.org/records/2526396 (accessed on 24 June 2025).

71. Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Meriaudeau, F. Indian Diabetic Retinopathy Image Dataset (IDRiD). 2018. Available online: https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid (accessed on 24 July 2025).
72. Hughes Cano, J.A.; Olivares Pinto, U.; Thébault, S.C. Dataset of Eye Fundus and OCT Images for the Study of Diabetic Macular Edema and Diabetic Retinopathy. GitHub Repository: Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET. 2022. Available online: https://github.com/Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET (accessed on 14 June 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.