IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 February 2026, accepted 20 February 2026, date of publication 25 February 2026, date of current version 2 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3668034

RESEARCH ARTICLE

Explainable Multimodal Foundational Models for Retinal Disease Stratification: A Robustness Study Across 15+ Heterogeneous Datasets

AINHOA OSA-SANCHEZ[1], AYMAN EL-BAZ[2], (Fellow, IEEE), IBON OLEAGORDIA-RUIZ[1], AND BEGONYA GARCIA-ZAPIRAIN[1]

[1]eVIDA Research Group, University of Deusto, 48007 Bilbao, Spain
[2]Department of Bioengineering, J. B. Speed School of Engineering, University of Louisville, Louisville, KY 40292, USA

Corresponding author: Ainhoa Osa-Sanchez (ainhoa.osa.sanchez@deusto.es)

ABSTRACT The automated stratification of retinal diseases remains a significant challenge due to data heterogeneity and the closed-box nature of deep learning models. Although foundational models have demonstrated remarkable success in general computer vision, their clinical reliability and interpretability in multimodal ophthalmology remain insufficiently explored. In this work, we introduce an Explainable Multimodal Foundational AI framework trained on a large-scale integrated corpus of 760,243 retinal images collected from over 15 heterogeneous repositories, encompassing both fundus photography and optical coherence tomography (OCT). We systematically evaluate self-supervised learning (SSL) paradigms DINO and iBOT across convolutional (ResNet) and Transformer-based (Vision Transformer, ViT) architectures. Our results show that ResNet-DINO achieves state-of-the-art performance, reaching 93.53% accuracy and a 0.935 F1-score in 6-class multimodal retinal disease classification, while exhibiting superior robustness under data-limited conditions, attributed to its inductive bias. Notably, we observe emergent clinical localization capabilities in Vision Transformer models (ViT-DINOv2 and ViT-iBOT). Using frozen pre-trained weights and without exposure to expert-labeled data or ground truth labels, these models autonomously highlight clinically relevant biomarkers, including subretinal fluid and drusen, demonstrating intrinsic pathological awareness. By bridging the semantic gap between unsupervised representation learning and targeted clinical diagnosis, this study establishes a benchmark for robust, explainable, and label-efficient AI in ophthalmology. Our findings indicate that large-scale foundational pre-training not only enhances diagnostic accuracy but also induces meaningful visual priors aligned with established clinical biomarkers, supporting the deployment of trustworthy AI systems in real-world clinical decision support.

INDEX TERMS Explainable AI (XAI), foundation models, self-supervised learning, multimodal fusion, retinal pathology, large-scale ophthalmic benchmark.

I. INTRODUCTION

The automated detection of eye diseases has become a central concern in clinical ophthalmology in recent years. This is due not only to the magnitude of the problem posed by diabetic retinopathy, age-related macular degeneration (AMD), and glaucoma, but also to the increasing difficulty of establishing reliable differential diagnoses between pathologies that share very similar visual signs [1]. It is worth remembering that, even with the accumulated experience of specialists, the overlap of clinical manifestations between maculopathy and retinal detachment remains a frequent obstacle in diagnostic practice [2], [3].

The emergence of foundational artificial intelligence (AI) models has opened up a particularly interesting technical possibility for these types of scenarios, where the

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

combination of imaging modalities such as optical coherence tomography (OCT) and fundus photography allows for a more comprehensive view of the retinal condition [3], [4]. Recent studies have shown that these models, when trained with multimodal learning strategies, achieve a noticeable improvement in diagnostic accuracy and generalizability to real-world clinical conditions (AMD, diabetic retinopathy, among others) [5], [6], [7]. It is worth noting that this type of integration is not limited to improving the accuracy of predictions; it also expands the structural and functional understanding of the eye in contexts where a single type of image would be insufficient.

Age-related macular degeneration (AMD) is a progressive disease affecting the macula of adults over 50, with cases projected to reach 288 million by 2040. Characterized by the accumulation of drusen, it exists in two forms: dry AMD, which progresses to geographic atrophy (GA), and wet AMD, which causes 90% of severe vision loss due to abnormal blood vessel leakage [54].

Recent literature emphasizes the automatic classification of AMD severity stages from normal retina to intermediate, GA, and wet stages to prevent irreversible vision loss [4], [7]. Architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have achieved accuracy rates between 93% and 99.5% [4], [8]. Additionally, explainable and self-supervised hierarchical approaches have successfully distinguished healthy retinas from those with varying pathologies with approximately 91% accuracy [9], [10].

It is crucial to remember that a significant part of the challenge lies in the robustness and clinical reliability of these models. Differences in image quality, the scarcity of samples of rare diseases, and the difficulties in ensuring multimodal consistency are factors that still demand robust technical solutions [11], [12]. This situation suggests that, while AI can overcome certain diagnostic limitations, its value largely depends on the careful integration of its results with medical interpretation [13], [14].

Against this backdrop, the present study establishes a new benchmark by synthesizing a massive, heterogeneous repository from over 15 public datasets, encompassing 636,000 images. This integration is non-trivial, as it merges data from diverse clinical environments, varying imaging protocols, and disparate sensor qualities. The research systematically examines how foundational AI models can be expanded to achieve a more precise and explainable stratification of ocular diseases through the following research questions:

1) *Generalizability across Diverse Pathologies:* To what extent can foundational AI models, pre-trained via self-supervised learning on massive-scale data, differentiate between retinal diseases with overlapping clinical signatures across heterogeneous acquisition protocols?
2) *Multimodal Synergy vs. Unimodal Performance:* Does the strategic integration of OCT and fundus photography significantly enhance diagnostic accuracy and robustness compared to unimodal architectures, particularly in data-constrained scenarios?
3) *Robustness to Image Heterogeneity:* How does the unification of multiple datasets characterized by varying image qualities, noise levels, and hardware-specific artifacts influence the stability of the learned visual representations compared to single-source training?
4) *Clinical Interpretability and Trust:* How do different foundational architectures (CNN vs. Vision Transformers) align with clinical biomarkers through explainable AI (XAI) techniques, such as integrated gradients and activation maximization?

To address these questions, we leverage a massive corpus of 760,243 images, utilizing the computational power of state-of-the-art hardware (NVIDIA RTX 5090) to ensure exhaustive hyperparameter optimization. The ultimate goal of this work is to demonstrate that by leveraging large-scale, multimodal data diversity, we can overcome the limitations of traditional models that are often overfitted to specific clinical settings. We aim to bridge the gap between technical reliability and the practical transparency needed for real-world ophthalmic decision support.

II. RELATED STUDIES

In the field of automated diagnosis of eye diseases, recent years have revealed a particularly dynamic and heterogeneous landscape. ViT models offer interesting possibilities due to their ability to generalize across modalities such as fundus and OCT imaging through self-supervision, achieving up to 98% accuracy in glaucoma and 99.6% in 2D OCT. Models such as OCT-Trans (96.68%) or hybrid variants like ViT-Large+GRU (92.92%) reveal that integrating context and temporality can be more productive than simply increasing network depth. This trend suggests an ongoing conceptual shift in the nature of retinal analysis.

Meanwhile, CNNs remain fundamental due to their efficiency and accuracy. VGG16 reaches up to 98.7%, adjusted ResNet-50 achieves 99.9%, and DenseNet201 achieves 99.69%. Lightweight models such as OctNET attain equivalent performance with only 6.9% of the parameters required by ResNet-50, highlighting the potential of optimization rather than scaling as a principle for network design.

Finally, ensemble strategies confirm the value of complementarity: voting and stacking schemes have improved performance up to 92% in diabetic retinopathy OCTA, while fusion of capillary plexus layers yields improvements of approximately 5.4%. In heterogeneous clinical contexts, federated learning as shown by Julian Lo et al. achieves an area under the receiver operating characteristic (AUROC) ranging from 0.954 to 0.960, maintaining accuracy without compromising privacy. This may represent a viable pathway for redefining the future of diagnostic intelligence. Kumar et al. [15], for example, comprehensively addressed the comparison between visual modalities, including 2D and 3D fundus and OCT imaging, to identify pathologies such as glaucoma and diabetic retinopathy. Their work not only

TABLE 1. Summarization of the discussed related studies.

Reference	Year	Task	Modality	Dataset	Model Type	Cons	Performance
Kumar et al. [15]	2025	Multi.	Fundus/OCT	OCT2017	EffNet, ViT	High compute; data diversity	99.6% (EffNet)
Hauri-Rosales [16]	2024	Multi.	2D OCT	84,484 img.	ResNet-50	ImageNet weight dependency	99.9% Acc.
Elsharkawy [17]	2025	Multi.	OCT	541 cases	OCT-Trans	Small sample; adaptation req.	96.6% Acc.
Ebrahimi [18]	2023	Multi.	OCTA	136 subj.	VGG16 Fusion	Multiscale alignment req.	92.6% Acc.
Lo et al. [19]	2021	Bin.	OCTA	700 eyes	Fed. ResU-Net	Federated complexity	AUROC 0.95-0.96
Sunija et al. [20]	2020	Multi.	OCT	83,484 img.	OctNET	Lacks multimodal depth	99.6% Acc.

confirms the superior accuracy of CNNs with EfficientNet reaching 99.6% in 2D OCT but also suggests something more profound: ViTs, even though they achieve slightly lower performance in specific tasks (e.g., 98% in fundus imaging), represent a promising avenue for cross-modal generalization. This finding invites a rethinking of the very notion of ''optimal performance,'' shifting it toward a more comprehensive understanding of visual learning.

It is noteworthy that when dealing with volumetric 3D OCT analysis, the same authors opted for a hybrid solution (ViT-Large+GRU) that achieved 92.92% accuracy by leveraging sequential dependencies between slices. This integration of spatial and temporal structures redefines the model's ability to capture subtle inter-slice variations, which purely spatial approaches often overlook.

Hauri-Rosales et al. [16] continued this research line from another angle by optimizing CNN architectures particularly ResNet-50 via fine-tuning strategies and pre-trained ImageNet weights. Their systematic experimental review demonstrated that proper calibration can achieve accuracies as high as 99.9%, even when training is limited to 10% of the full dataset (approximately 8,000 images). This nontrivial observation shows that transfer learning not only compensates for data scarcity but also redefines our understanding of efficiency in biomedical training paradigms.

Likewise, Elsharkawy et al. [17] introduced OCT-Trans, a Transformer specifically adapted to OCT morphology and texture. Achieving 96.68% accuracy in distinguishing AMD and diabetic retinopathy, the model also demonstrated significant improvements over ViT-B and ConvNeXt-T according to Wilcoxon tests ($p < 0.05$). This architectural customization suggests a critical evolution: Transformers are shifting from general-purpose visual models to domain-specific, multimodal structures suited to the medical imaging field.

In a related context, Ebrahimi et al. [18] investigated the impact of layer fusion strategies in OCTA images for diabetic retinopathy detection. Their work revealed that intermediate fusion architectures integrating information from multiple vascular plexuses both superficial and deep improve accuracy by approximately 5.4% compared to single-layer models, reaching 92.65%. This indicates that intelligent, multiscale integration can outperform brute-force network deepening, underscoring the importance of anatomical and structural context.

Meanwhile, Lo et al. [19] explored federated learning (FL) as a complementary paradigm. Their multi-institutional OCTA study achieved AUROC values between 0.954 and 0.960. Beyond numerical improvements, the ethical and practical implications are notable: decentralized training not only preserves patient privacy but also generates models with improved robustness and generalizability. Thus, FL emerges as a promising solution for medical AI development under equity and data protection constraints.

Finally, Sunija et al. [20] proposed OctNET, a lightweight six-block CNN optimized for computational efficiency. Despite its compact architecture, OctNET achieved 99.69% accuracy, matching ResNet-50 while using only 6.9% of its parameters. This finding underscores that purely quantitative metrics such as raw accuracy or model complexity may fail to capture the true value of lightweight solutions in constrained environments, where energy, hardware, and accessibility limitations shape their real-world impact.

III. MATERIALS

To train the foundational model, multiple publicly available datasets were compiled and integrated. All datasets underpinning this study were sourced exclusively from open repositories, ensuring transparent and reproducible experimentation within ophthalmic imaging. Large-scale data were indispensable not only for model initialization but also for capturing a broad spectrum of anatomical and acquisition variations. Two main repositories *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification* [21] (253,452 images) and the *Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images* [22] (108,309 images) formed the core of the pretraining stage, providing a diverse foundation from which

TABLE 2. Integrated ophthalmic datasets: partitioning for self-supervised pre-training and supervised specific training.

Dataset Name	Modality	Count	Role in Proposed Framework
[HTML]F2F2F2	*Phase 1: Self-Supervised Pre-training (No Labels used)*		
Labeled OCT & Chest X-Ray (Subset)	OCT	253,452	Large-scale Feature Initialization
Large Dataset Labeled OCT (Subset)	OCT	108,309	Visual Representation Learning
OCT-Retina-Disease	OCT	109,300	Intra-modality Diversity
AIROGS	Fundus	101,442	Global Anatomical Features
EYEPACS Diabetic	Fundus	88,712	General Pathological Patterns
RetinalOCT_Dataset-C8	OCT	24,000	Spatial Feature Expansion
DDR Dataset	Fundus	17,458	Structural Diversity
Labeled Retinal OCT	OCT	16,803	Anatomical Representation
MM-Retinal	Multimodal	4,391	Multimodal Weight Alignment
OIMHS	OCT	3,859	Feature Refinement
Composite Retinal (Fundus + OCT)	Multimodal	2,560	Cross-modal Pre-training
Messidor	Fundus	900	Pattern Diversity
[HTML]F2F2F2	*Phase 2: Specific Training and Evaluation (With Labels)*		
Eye Disease Image Dataset	Mixed	21,577	Multi-disease Classification
REFUGE_2	Fundus	2,400	Glaucoma Stratification
OCTDL	OCT	2,064	Disease Fine-tuning
Glaucoma Detection	Fundus	2,007	Glaucoma Validation
Glaucoma Fundus	Fundus	1,544	Targeted Pathological Analysis
Fundus CATT	Fundus	864	AMD Severity Grading
sjchoi86-HRF	Fundus	601	High-resolution Evaluation
Total Integrated Images		**760,243**	**Combined Ophthalmic Scale**

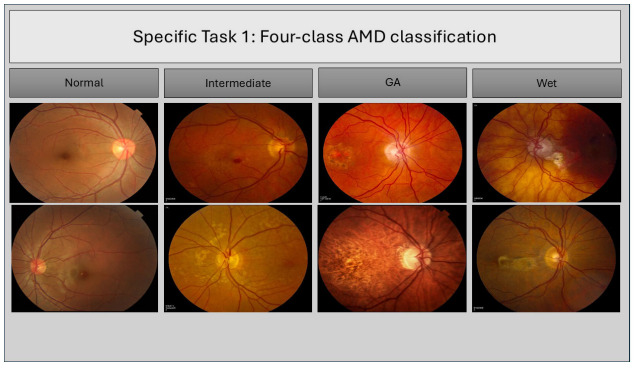

FIGURE 1. Samples from CATT dataset for AMD detection.

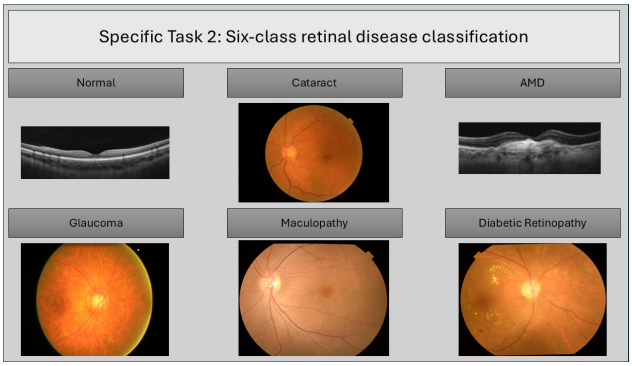

FIGURE 2. Samples from the different diseases from the datasets.

the proposed foundational models could extract transferable visual features.

During the initial learning phases, additional datasets reinforced this process by introducing alternative scanning conditions and disease categories. Sets such as *OCT-Retina-Disease* (109,300 images) [23] and *Labeled Retinal Optical Coherence Tomography* (16,803 images) [24] contributed to refining intra-modality representations. Complementary collections including *AIROGS* (101,442 images) [25], *EYEPACS Diabetic* (88,712 images) [26], *DDR Dataset* (17,458 images), and *MM-Retinal* (4,391 images) further diversified the data spectrum. Smaller archives such as *Messidor* (900 images) [27], *RetinalOCT_Dataset-C8* (24,000 images) [28], *OIMHS* (3,859 images) [29], and *Composite Retinal Fundus and OCT Dataset* (2,560 images) [30] expanded the overall visual representation coverage.

Collectively, these resources supported a large-scale pre-training phase aimed at deriving representations resilient to variations in both imaging modality and acquisition protocol, establishing a robust foundation for subsequent fine-tuning and multimodal adaptation stages. When the model entered the validation and adaptation phase, the focus shifted toward assessing its capacity to generalize across specific diagnostic targets. For glaucoma-related examinations and optic disc analysis, we employed several specialized datasets, including *Glaucoma Detection* [31] (2,007 images), *OCTDL* [32] (2,064 images), *Glaucoma Fundus* [33] (1,544 images), *sjchoi86-HRF* [34] (601 images), and *REFUGE_2* [35] (2,400 images), see Fig. 2.

Meanwhile, the *Eye Disease Image Dataset* [36] (21,577 images) provided a comprehensive testbed for multi-disease screening scenarios, enabling verification of both classification and severity grading of retinal conditions under heterogeneous data environments. For AMD classification and maculopathy assessment, the *Fundus CATT* dataset [37] (864 images) was utilized, see Fig. 1.

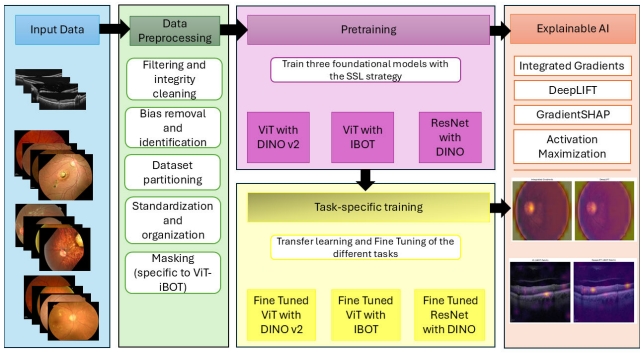

FIGURE 3. An outline of the explainability and tuning of the suggested framework.

It is important to note that all datasets used in this study comply with the usage conditions prescribed by their respective repositories and were obtained exclusively from official distribution platforms. No patient-identifiable clinical materials were introduced at any stage unless explicitly addressed elsewhere in this manuscript, ensuring full compliance with data governance and ethical research standards.

IV. METHODOLOGY

The proposed framework, whose general scheme is presented in Fig. 3, begins with exhaustive preprocessing of fundus and OCT images to highlight regions of interest and ensure data integrity through filtering techniques and multi-crop magnification. Foundational models based on ViT and ResNet-50 are then employed.

A. DATA PREPROCESSING

During the initial preparation stage, particular attention was given to the integrity and informational value of the images. Files that displayed corruption, duplication, or that contained auxiliary layers such as binary masks or label overlays intended for annotation rather than visual inspection were excluded. This filtering ensured that the material retained was visually meaningful and suitable for representation learning under self-supervised conditions. It is worth noting that any form of metadata or textual trace linked to class identity, whether in file names or image headers, was deliberately removed before initiating pretraining, thus eliminating any inadvertent guidance that could bias feature extraction.

Once this process had been completed, the image bank was reduced to 636,852 elements, all of them available for self-supervised representation learning without reference to downstream labels. That decision responded to a methodological choice: to privilege the observation of image structure itself rather than predefined diagnostic categories. In contrast, for subsequent supervised experiments, only instances with verified annotations were considered. From this narrower pool of 16,885 images in total, three disjoint partitions were produced following a 70/15/15 proportion to separate training, validation, and testing material, respectively.

To facilitate reproducibility and coherent pipeline integration, each subset was stored in its own directory and linked to a CSV annotation file, but only at the supervised stage. During pretraining, these associations were intentionally withheld. Such an arrangement, while seemingly simple, proved effective in maintaining strict independence between unsupervised representation learning and supervised evaluation—a distinction that underpins the experimental design as a whole.

B. PRETRAINING

This study compares three pre-trained visual models that share the same reference backbone but differ in the type of self-supervised learning strategy used. A ResNet-50 model trained using DINO and two configurations of the Vision Transformer ViT-B/16 are employed: one using the iBOT strategy and the other using DINOv2. This choice is not purely technical: it aims to examine, in a controlled manner, how the nature of the self-supervised learning objective influences the robustness of the visual representations when the transformer architecture remains fixed. Ultimately, the comparison between the convolutional encoder and the self-supervised learning models reveals differences not only in structure but also in how both mechanisms capture the variability of the biomedical image.

1) SSL ARCHITECTURE

The self-learning framework used falls within the student-teacher method, which relies on distillation through joint embeddings. The input samples undergo a multi-crop augmentation process where, from each image, two global views and a series of local cropped views are generated. These transformations include horizontal rotations, chromatic alterations, normalization, and occasional grayscale conversion. While the student encoder processes all the generated views, the primary updated by momentum as an exponential average of the student's parameters receives only the global views. This dynamic, by stabilizing the master's projections, induces scale-invariant representations, which is particularly useful in medical domains with morphological heterogeneity.

In the case of the DINO [38], [39] model, for both ResNet-50 and ViT-B/16, training is based on the student reproducing the probability distribution estimated by the primary for different views of the same image. Normalization through centering and temperature adjustment manages the cross-entropy used as the main loss. The advantage of this scheme lies in its elimination of explicit negative pairs, favoring a type of representation that tends to be richer and more transferable to subsequent diagnostic or classification tasks in clinical images.

The iBOT [40], [41] method, applied exclusively to the ViT-B/16 backbone, introduces a combination of distillation and masked image modeling. A subset of patches is hidden

in the student's input, who must reconstruct their representations based on the teacher's predictions. This incorporation of masked image modeling gives the model a dual capability: capturing large-scale relationships while simultaneously integrating the local texture that characterizes histological or cellular tissues. It is worth noting that in histopathology images, this dual sensitivity global and local is critical for reliable automated interpretations.

The DINOv2 [38], [42] procedure maintains the student-teacher philosophy but with readjusted training. Both the selection and specificity of the views and the normalization of the teacher's outputs are refined. This refinement is accompanied by more careful control of hyperparameters and extended training times, with the goal of generating stable, uniformly distributed, and easily reusable embeddings in transfer scenarios, even when the amount of labeled data is limited. It is worth noting that this type of adaptability is especially valuable in contexts where the cost of annotation is high or access to specialists is restricted.

2) MODELS

ResNet-50 [43], [44] acts as the reference convolutional backbone, composed of four residual stages with bottleneck blocks and a global pooling layer that produces a 2048-dimensional embedding. A nonlinear projection head is added to this encoder, used only during pretraining with DINO. After pretraining, the backbone is reused as a feature extractor for downstream tasks, where task-specific heads are added (classification, segmentation, etc.).

The first Vision Transformer considered is ViT-B/16 [45], [46], which divides the image into 16×16 pixel patches and processes them using a self-addressing Transformer layer stack. In the ViT-iBOT variant, this backbone is trained with the combination of image-level distillation and masked image modeling described above. The patchy nature of the model, along with its objective of reconstructing masked patches, makes it particularly powerful for exploiting fine morphological patterns in biomedical images.

The third model is also based on ViT-B/16 [47], [48], but it is trained using the DINOv2 recipe. By keeping the architecture fixed and varying only the SSL method, the effect of the loss formulation and the training recipe on the quality of the representations can be isolated. This model is geared towards producing more stable and uniform features, with good out-of-domain performance and in tasks with little labeled data.

C. TASK-SPECIFIC TRAINING

Our foundation models were adapted to downstream retinal disease classification tasks using feature-based transfer learning and end-to-end fine-tuning strategies. Three self-supervised foundation models pretrained on large-scale retinal imaging data were evaluated: ResNet-50 trained with DINO, ViT-B/16 trained with iBOT, and ViT-B/16 trained with DINOv2. These models were transferred to two clinically relevant classification tasks to assess their ability to generalize to labeled medical data.

Two task-specific classification problems were considered:

- **Four-class AMD classification:** This task aimed to predict the severity of AMD from color fundus photographs from the CATT dataset, with categories: Normal, Intermediate AMD, Geographic Atrophy, and Neovascular (Wet) AMD. Given the limited number of labeled samples, this task was particularly suitable for transfer learning, where pretrained features were used as input to a task-specific classification head without extensive backbone fine-tuning. For the ViT models, the [CLS] token embedding served as the global image representation, while for ResNet-50, the pooled convolutional features were used.
- **Six-class retinal disease classification:** The second task involved a larger and more heterogeneous dataset combining fundus and OCT images. The target classes were AMD, Diabetic Retinopathy, Glaucoma, Maculopathy, Normal, and Retinal Detachment. Due to the increased dataset size and task complexity, this setting enabled end-to-end fine-tuning of the pretrained backbones together with task-specific heads, allowing the models to adapt fully to the new domain while leveraging the pretrained representations.

D. EXPLAINABILITY AND INTERPRETABILITY

For AI models to generate accurate and intelligible predictions in the context of AMD and other retinal diseases, it is imperative that their decisions are both understandable and clinically interpretable [49]. In this work, explainability is achieved through attribution and representation-based techniques grounded on gradients and internal activations, such as Integrated Gradients, DeepLIFT, GradientSHAP, and Activation Maximization, which have been extensively validated in biomedical and multimodal deep learning settings.

Integrated Gradients, DeepLIFT, and GradientSHAP provide pixel-level attribution maps by linking the model's prediction to specific regions of the input image. Integrated Gradients attributes importance by integrating the gradients of the model output along a path between a reference input and the real image, satisfying desirable properties such as completeness [50]. DeepLIFT assigns contributions by propagating activation differences with respect to a reference input, offering improved behavior in regions with vanishing gradients [51]. GradientSHAP combines ideas from Integrated Gradients and Shapley values, estimating attribution scores by averaging gradients over multiple noisy reference inputs, resulting in more stable and robust explanations. These methods generate heatmaps over OCT and fundus images that highlight the retinal structures supporting each prediction [52].

Activation Maximization [53] complements these attribution techniques by providing insight into the internal representations learned by the model. Rather than explaining

individual predictions, this method optimizes a synthetic input to maximize the activation of specific neurons, channels, or class logits, revealing prototypical patterns encoded by the network. In the context of AMD, Activation Maximization can expose characteristic configurations of drusen, retinal thickening, or intraretinal fluid that the model associates with particular disease stages. In multimodal architectures, activations can be maximized independently for the OCT and fundus branches, facilitating modality-specific interpretability.

In diagnosing and monitoring AMD, these explainability methods can significantly aid clinicians by providing visual and mechanistic insight into how deep learning models base their decisions on retinal images. For instance:

Integrated Gradients, DeepLIFT, and GradientSHAP can reveal which retinal regions are most influential for a given prediction, allowing clinicians to verify that the model focuses on clinically meaningful biomarkers such as drusen deposits, edema, or areas of atrophy. These attribution maps support both local explanations at the patient level and the identification of consistent spatial patterns across the population.

Activation Maximization enables the analysis of global model behavior by visualizing prototypical features learned by internal filters. This can help clinicians and researchers understand what constitutes a "typical" representation of AMD-related pathology from the model's perspective, complementing case-specific heatmaps with higher-level interpretability.

Combining gradient-based attribution methods with activation-based analysis offers a complementary and comprehensive interpretability framework. Attribution techniques provide localized, patient-specific explanations of individual predictions, while Activation Maximization reveals global and modality-specific patterns learned by the model. Together, these methods enhance transparency, foster clinical trust, and improve the reliability of AI systems as diagnostic support tools for AMD and other retinal diseases.

E. PERFORMANCE ASSESSMENT

In evaluating classification models, it is common to use a set of metrics that allow for an objective assessment of their performance and, at the same time, an understanding of their limitations. It is crucial to remember that a model's value is not defined solely by its accuracy but also by the nature of its errors. In this sense, measures such as precision, sensitivity, and specificity complement each other to offer a more nuanced view of its overall performance [11].

It is worth recalling that accuracy or the overall success rate is obtained by dividing the number of correct predictions by the total number of samples analyzed:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Although high accuracy is generally considered positive, it does not always accurately describe the system's behavior, especially when the data is unbalanced or when the cost of errors varies between classes. This necessitates considering additional indicators that more faithfully reflect the model's ability to distinguish between positive and negative cases.

From a more detailed perspective, sensitivity (also called recall or true positive rate) evaluates the model's ability to correctly identify positive cases:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

High sensitivity is valuable when the goal is to minimize false negatives, as in medical or early diagnostic applications, where omitting a real case can have significant consequences.

In contrast, specificity analyzes the model's behavior with respect to the negative class and is defined as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

Its value becomes important when seeking to reduce false positives, a common situation in clinical or quality control systems.

Precision, on the other hand, is expressed as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

This metric represents the proportion of positive predictions that are true, an aspect closely linked to the reliability of automated decisions.

To balance the relationship between sensitivity and accuracy, the F1-score is used, which represents their harmonic mean:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{5}$$

It should be noted that a high F1 score not only indicates good detection capability but also consistency in the classification of positive samples, especially in contexts with unbalanced classes.

V. EXPERIMENTS AND DISCUSSION

The classification models were trained using a dedicated local workstation with an NVIDIA GeForce RTX 5090 graphics card. It offers sufficient processing power for demanding deep learning applications. The purpose of this hardware configuration was to specifically utilize the parallel processing power of the GPU, which is essential for accelerating model training and enabling efficient manipulation of system files.

Python 3.10 and PyTorch, a deep learning framework well known for its adaptability and effectiveness in research settings, were used to run the model building, training, validation, and testing processes.

The target of the experiments is to answer the current study research questions. This is accomplished by applying machine learning algorithms such as CNN, MLP, and Transformer Classifier to examine the numerical features collected from retinal images. To ensure that each model is optimized, these models are trained and evaluated using a set of experiments including early stopping, loss, and precision

TABLE 3. Summarization of the obtained results for task 1.

Model	Accuracy	F1	Recall	Precision
ResNet-DINO	0.9205	0.920	0.925	0.923
ViT-iBOT	0.5284	0.512	0.5283	0.552
ViT-DINO V2	0.5000	0.4967	0.500	0.500

TABLE 4. Summarization of the obtained results for task 2.

Model	Acc	F1 M	F1 W	Rec W	Prec W
ResNet-DINO	0.9353	0.9464	0.9354	0.94	0.94
ViT-iBOT	0.9096	0.9236	0.9093	0.91	0.91
ViT-DINO V2	0.7648	0.7740	0.7600	0.78	0.78

Acc: accuracy; *F1 M*: macro F1-score; *F1 W*: weighted F1-score; *Rec W*: weighted recall; *Prec W*: weighted precision.

tracking. Performance metrics of precision, accuracy, recall, and sensitivity are calculated through test runs. Using the Optuna library, hyperparameters of each model are tuned and the results of each tuning are saved to evaluate how well the models solve the key research problems.

A. HYPERPARAMETER TUNING OF PRETRAINING MODELS

All models were trained using the categorical cross-entropy loss, which is well suited for multi-class classification tasks and is commonly used in medical image analysis due to its stable gradients and effectiveness in probabilistic classification. To ensure a fair and robust comparison among the different pretrained architectures evaluated in this study, hyperparameter optimization was conducted using Optuna, with the validation loss as the optimization objective. Specifically, Optuna aimed to minimize the categorical cross-entropy loss computed on the validation set, thereby directly encouraging improved generalization performance during fine-tuning.

The hyperparameter search space was defined consistently across architectures. The learning rate was sampled on a logarithmic scale within the range $[1 \times 10^{-5}, 5 \times 10^{-4}]$, and the batch size was selected from the discrete set $\{16, 32, 64\}$. This shared search space ensured methodological consistency between CNN-based models (ResNet-DINO) and Vision Transformer-based models (ViT-DINO and ViT-iBOT).

For the ResNet-DINO model, Optuna identified an optimal configuration with a learning rate of 2.85×10^{-5} and a batch size of 16. For the ViT-DINO model, the best-performing configuration consisted of a learning rate of 1.11×10^{-4} and a batch size of 16.

In the case of ViT-iBOT, additional architecture-specific hyperparameters were included in the optimization process. Alongside the learning rate and batch size, the mask ratio, which controls the proportion of image patches masked during training, was optimized within the range [0.1, 0.7]. The optimal configuration selected by Optuna corresponded to a learning rate of 3.66×10^{-4}, a batch size of 32, and a mask ratio of 0.64.

All selected hyperparameter configurations were fixed after optimization and subsequently used for training and evaluation on the downstream classification tasks reported in this study.

B. REPORTED RESULTS

Model performance was evaluated using multiple complementary metrics, including Accuracy, F1-score (macro and weighted), Precision, and Recall. These metrics were selected to provide a comprehensive assessment of classification performance, particularly in the presence of class imbalance, which is common in ophthalmological datasets.

All reported results correspond to evaluations on an independent test set that was not used during training or validation. This ensures that the reported performance reflects the true generalization capability of each model.

1) SPECIFIC TASK 1

In the first use case, table 3, models were evaluated on a four-class classification task focused on AMD using fundus images. The dataset was divided into training (516 images), validation (172 images), and test (176 images) sets. The four classes included Normal, Intermediate AMD, GA, and Neovascular (Wet) AMD. Note: The near-random performance of ViT architectures in Task 1 highlights the 'Semantic Gap' in low-data regimes, a phenomenon further explored in the Explainability section to demonstrate their emergent pathological awareness despite low classification accuracy.

The ResNet-DINO model achieved the best overall performance, reaching an accuracy of 0.9205 and a weighted F1-score of 0.920. Precision and recall values were also consistently high (0.923 and 0.925, respectively), indicating a balanced and reliable classification across all classes.

In contrast, the Vision Transformer-based models showed significantly lower performance in this task. ViT-iBOT obtained an accuracy of 0.5284 and a weighted F1-score of 0.512, while ViT-DINO v2 achieved near-random performance, with an accuracy of 0.50 and a weighted F1-score of 0.497. These results suggest that, for limited dataset sizes, CNN-based architectures pretrained with self-supervised learning may be more effective than transformer-based approaches.

2) SPECIFIC TASK 2

The second use case, table 4 addressed a more complex six-class classification problem involving multimodal retinal imaging (OCT and fundus images). The dataset comprised 11,220 training samples, 2,401 validation samples, and 2,411 test samples. The classes included AMD, Diabetic Retinopathy (DR), Glaucoma, Maculopathy, Normal, and Retinal Detachment.

In this scenario, all models exhibited improved performance compared to Use Case 1, highlighting the positive impact of a larger and more diverse dataset. ResNet-DINO

again achieved the highest performance, with an accuracy of 0.9353, a macro F1-score of 0.9464, and a weighted F1-score of 0.9354. Precision and recall values (both 0.94) further confirm the robustness of this model across classes.

ViT-iBOT demonstrated competitive performance, achieving an accuracy of 0.9096 and a macro F1-score of 0.9236, suggesting that transformer-based models benefit substantially from increased data availability. However, ViT-DINO v2 lagged behind, with an accuracy of 0.7648 and a macro F1-score of 0.7740, indicating limited generalization capability in this multimodal setting.

C. COMPARATIVE ANALYSIS

This study presents a comparative evaluation of self-supervised pretrained convolutional and transformer-based architectures across two retinal disease classification tasks of increasing complexity. In response to RQ2 and RQ3, our results consistently highlight a divergence in performance dictated by architectural inductive biases and dataset scale. In the first use case (4)-class AMD classification), the ResNet-DINO model significantly outperformed both Vision Transformer-based models. This substantial performance gap suggests that *convolutional inductive biases* specifically locality and translation invariance are indispensable when training data are scarce ($n < 1,000$). In contrast, ViT-based models, particularly ViT-DINO v2, exhibited near-random performance. This indicates an insufficient adaptation of global self-attention mechanisms under limited supervision, where the model fails to converge on discriminative features without a critical mass of medical samples.The second use case (6-class multimodal task) revealed a pivotal shift in this trend. By utilizing a large-scale dataset combining OCT and fundus images, the performance gap between ResNet-DINO and ViT-iBOT narrowed considerably. This observation supports the hypothesis that transformer-based architectures leverage long-range dependencies and global context more effectively as the dataset size grows, eventually challenging the dominance of CNNs.Interestingly, the consistent underperformance of ViT-DINO v2 compared to ViT-iBOT across both tasks suggests that *pre-training objectives* are as critical as architecture. While DINO focuses on discriminative contrastive learning, iBOT's Masked Image Modeling (MIM) objective seems to force a more robust internalization of retinal structural context, facilitating better transferability to downstream medical tasks.Overall, these findings indicate that model selection for retinal image classification must be strategically guided by **data availability and task dimensionality**. While CNN-based SSL models offer superior stability in low-data regimes, transformer-based models represent a more scalable solution for large-scale, multimodal clinical integration where global pathological context is paramount.

D. RELATED STUDIES COMPARISONS

When contextualized within prior ophthalmic AI research, the trends observed in our experiments both align with and extend prevailing expectations regarding model capacity and data dependency. Supervised convolutional approaches have historically defined the benchmark for retinal disease classification; for instance, Hauri-Rosales et al. [16] and Sunija et al. [20] reported near-ceiling accuracies (up to 99.9%) on 2D OCT datasets. However, these models depend heavily on extensive pre-labeled data and supervised transfer learning from ImageNet, which often constrains their generalizability in heterogeneous clinical settings.

In our previous work [54], we demonstrated that cascaded architectures combining ViT-extracted features with CNN and MLP classifiers could achieve high diagnostic precision (94.19% for AMD) by focusing on specific numerical patterns. While effective, that approach still relied on supervised feature extraction. In contrast, our current ResNet-DINO foundational model achieves comparable reliability (93.53% across six complex classes) without any supervised transfer learning. This underscores the evolution from task-specific cascaded models to label-efficient foundational paradigms that internalize retinal anatomy through massive self-supervised pre-training on 760,243 images.

Closely related, the hybrid frameworks of Kumar et al. [15] suggested that spatial integration is critical for 3D OCT tasks. Our findings partly resonate with this; however, we observe that the importance of global context modeling increases with task complexity. This is evidenced by the improvement of ViT-iBOT in our six-class scenario compared to the four-class task. Furthermore, Elsharkawy et al. [17] noted that domain-specific adaptation of Transformers (OCT-Trans) outperforms generic models. This mirrors our observation regarding the ''Semantic Gap'': the underperformance of ViT-DINOv2 in classification tasks suggests that architectural scaling alone cannot compensate for the need for domain-aligned pre-training objectives, even when the model exhibits superior emergent localization capabilities.

Conceptually, the benefits of multiscale integration highlighted by Ebrahimi et al. [18] align with the implicit sensitivity of our DINO-based framework, which captures fine-grained retinal textures without explicit fusion layers. While federated approaches like those of Lo et al. [19] focus on institutional privacy, our SSL approach prioritizes robustness and autonomous interpretability key strengths that may be more relevant than raw accuracy for clinical deployment.

Taken together, these comparisons reveal a converging narrative: while supervised CNNs and cascaded models (as seen in [54]) define reference performance on curated data, the next frontier represented by our Explainable Multimodal Foundational Model emphasizes the synergy between self-supervised representation learning and autonomous clinical biomarker discovery. This shift addresses the critical need for AI systems that are not only accurate but inherently transparent and adaptable to the fragmented landscape of global ophthalmic data.

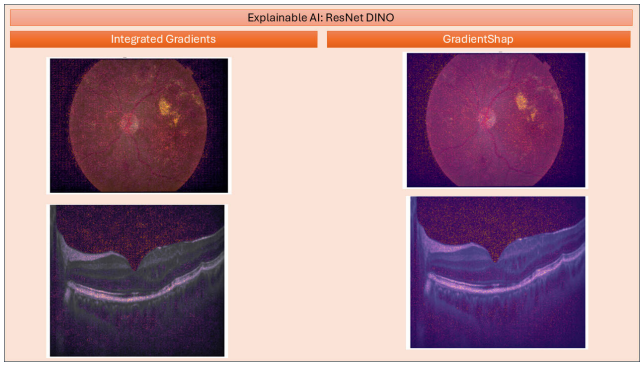

FIGURE 4. Explainability and interpretability of the ResNet DINO classifier. Visualizations generated from frozen self-supervised weights, demonstrating zero-shot pathological localization.

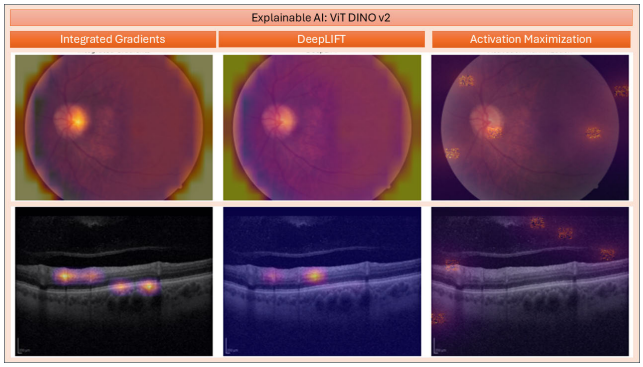

FIGURE 5. Explainability and interpretability of the ViT DINO v2 classifier. Visualizations generated from frozen self-supervised weights, demonstrating zero-shot pathological localization.

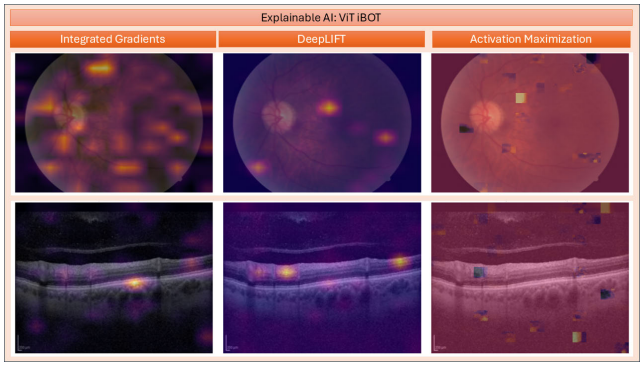

FIGURE 6. Explainability and interpretability of the ViT iBOT classifier. Visualizations generated from frozen self-supervised weights, demonstrating zero-shot pathological localization.

E. EXPLAINABILITY AND INTERPRETABILITY

The deployment of foundational models in this study reveals a critical trade-off between predictive robustness and autonomous clinical interpretability. Notably, the explainability analyses presented in Figs. 4–6 were conducted on the **frozen self-supervised weights prior to any supervised fine-tuning**, highlighting the models' inherent capacity to internalize ophthalmic structures.

As shown in Fig. 4, the ResNet-DINO architecture demonstrated superior diagnostic robustness ($Acc = 0.9205$). However, its decision-making process relies on a global integration of high-frequency textural features. While computationally efficient, this representation remains less intuitive for clinical validation as it lacks localized anatomical grounding.

Conversely, the ViT-DINOv2 model (see Fig. 5) exhibited remarkable *emergent segmentation* capabilities. Without explicit supervision or label exposure, the model precisely localized pathological biomarkers, such as subretinal fluid and drusen, in OCT B-scans. We identify this phenomenon as a "Semantic Gap": while foundational SSL pre-training excels at identifying and isolating anatomical anomalies through structural contrast, the subsequent categorical differentiation (e.g., specific AMD staging) requires a more discriminative latent space mapping provided by the specific training phase.

Furthermore, the performance of ViT-iBOT (Fig. 6) underscores the efficacy of the Masked Image Modeling (MIM) objective. By utilizing an optimized mask ratio of 0.64, the model was forced to internalize the global structural context of the retina, resulting in regional attention maps that spontaneously align with clinical pathological zones.

In conclusion, leveraging foundation models trained on massive datasets provides a superior "visual prior" that drastically reduces the dependency on large-scale labeled medical archives. However, the choice between architectures remains strategic: ResNet-based models prioritize raw diagnostic accuracy, whereas Transformer-based models offer unprecedented clinical transparency through unsupervised feature localization.

VI. LIMITATIONS

While this study demonstrates the potential of foundational models, several structural and technical constraints must be acknowledged. A primary limitation arises from **institutional fragmentation**: the use of heterogeneous imaging equipment and varying acquisition protocols (e.g., different OCT signal-to-noise ratios and fundus camera optics) hinders the establishment of standardized interoperability. This landscape complicates the deployment of federated frameworks and may constrain the generalizability of representations learned in isolated institutional silos.

Model-level limitations also define the current boundaries of ViTs in clinical settings. The sensitivity of architectures like ViT-DINOv2 to dataset volume was evident in Task 1, where accuracies near 0.50 confirmed that without substantial training data, Transformers lack the *local inductive bias* and *translation invariance* inherent in CNNs.

Furthermore, we identify a critical "Semantic Gap": although ViT-DINOv2 exhibited *emergent segmentation* autonomously identifying morphological biomarkers like subretinal fluid it lacked the discriminative mapping necessary for precise disease staging. Finally, the high computational overhead of hybrid models (e.g., ViT-Large+GRU)

and the requirement for meticulous fine-tuning remain barriers for practical adoption in resource-constrained hospital environments.

VII. CONCLUSION AND FUTURE DIRECTIONS

This research demonstrates that the future of ophthalmic AI lies in balancing predictive robustness with autonomous clinical interpretability. By training foundational models on a massive corpus of **760,243 images across 15+ datasets**, we have shown that self-supervised learning (SSL) can internalize complex retinal priors without the need for exhaustive expert labeling.

The ResNet-DINO model established a benchmark for stability, achieving a multimodal accuracy of 0.9353, proving that SSL strategies can effectively replace traditional ImageNet-based transfer learning. Simultaneously, our findings reveal a strategic trade-off: whereas CNNs prioritize raw diagnostic reliability, Transformers facilitate an unprecedented alignment with clinical reasoning through emergent attention maps.

Future research will focus on three key pillars:

1) **Architectural Evolution:** Developing 3D architectures that integrate ViT-Large with recurrent modules (e.g., GRUs) to capture the volumetric continuity of OCT B-scans, a dimension often lost in 2D approaches.
2) **Privacy-Preserving Scaling:** Implementing Federated Learning (FL) to scale models across institutional boundaries while maintaining data sovereignty and achieving target AUROC values above 0.95.
3) **Computational Efficiency:** Exploring lightweight alternatives like OctNET to achieve performance parity with larger backbones using a fraction of the parameters, ensuring viability for real-time clinical decision support.

Ultimately, narrowing the "Semantic Gap" through domain-specific SSL loss functions and multiscale fusion (e.g., integrating OCTA plexus layers) remains the most promising frontier for realizing trustworthy, interpretable, and clinically reliable AI in the ophthalmic field.

DECLARATIONS

- **Conflict of Interest**: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Funding**: This work was supported by the Basque Government through the Hazitek 2024 program, Spain, within the framework of the IRUD-IA project: "Medical Image Analysis Technologies with Artificial Intelligence for the Development of Medical Devices," project code ZE-2024/00030.
- **Intellectual Property**: The authors confirm that due consideration has been given to the protection of intellectual property associated with this work and that there are no impediments to publication, following the regulations of the involved institutions.
- **Authorship**: The manuscript has been read and approved by all named authors. The order of authors has been approved by all signatories.
- **Data Availability**: Detailed information regarding the datasets used in this study is provided in the Materials and Methods section.
- **Compliance with Ethical Standards**: This study was conducted using retrospective, de-identified data from public and private repositories. It does not contain any new studies with human participants or animals performed by any of the authors.
- **Consent to Participate and Publish**: Not applicable, as no individual human participant data are reported.

REFERENCES

[1] N. Shi, J. Li, M. Shang, W. Zhang, K. Xu, Y. Li, and L. Liang, "Detection and management of geographic atrophy secondary to age-related macular degeneration using noninvasive retinal images and artificial intelligence: Systematic review," *J. Med. Internet Res.*, vol. 27, Nov. 2025, Art. no. e81328, doi: 10.2196/81328.

[2] K. A. Bokhary, "Narrative review of artificial intelligence in ophthalmic disease detection," *Galen Med. J.*, vol. 14, p. e3979, Sep. 2025.

[3] L. F. C. Vilela, N. O. Cabral, A. C. Destefani, and V. C. Destefani, "Harnessing the power of artificial intelligence for early detection and management of diabetic retinopathy, age-related macular degeneration, and glaucoma: A narrative review of deep learning applications in ophthalmology," *Revista Ibero-Americana de Humanidades, Ciencias e Educação*, vol. 10, no. 8, pp. 3311–3320, Aug. 2024, doi: 10.51891/rease.v10i8.15395.

[4] H. Hashemian, T. Peto, R. Ambrósio Jr., I. Lengyel, R. Kafieh, A. M. Noori, and M. Khorrami-Nezhad, "Application of artificial intelligence in ophthalmology: An updated comprehensive review," *J. Ophthalmic Vis. Res.*, vol. 19, no. 3, pp. 354–367, Sep. 2024, doi: 10.18502/jovr.v19i3.15893.

[5] H. S. Ahmed and C. J. Thrishulamurthy, "Advancing diabetic retinopathy diagnosis: Leveraging optical coherence tomography imaging with convolutional neural networks," *Romanian J. Ophthalmol.*, vol. 67, no. 4, pp. 326–336, Dec. 2023. [Online]. Available: https://rjo.ro/advancing-diabetic-retinopathy-diagnosis-leveraging-optical-coherence-tomography-imaging-with-convolutional-neural-networks/

[6] A. Hayati, M. R. Abdol Homayuni, R. Sadeghi, H. Asadigandomani, M. Dashtkoohi, S. Eslami, and M. Soleimani, "Advancing diabetic retinopathy screening: A systematic review of artificial intelligence and optical coherence tomography angiography innovations," *Diagnostics*, vol. 15, no. 6, p. 737, Mar. 2025, doi: 10.3390/diagnostics15060737.

[7] P. Riazi Esfahani, A. J. Reddy, N. Nawathey, M. S. Ghauri, M. Min, H. Wagh, N. Tak, and R. Patel, "Deep learning classification of drusen, choroidal neovascularization, and diabetic macular edema in optical coherence tomography (OCT) images," *Cureus*, p. 41615, Jul. 2023, doi: 10.7759/cureus.41615.

[8] E. Hassan, S. Elmougy, M. R. Ibraheem, M. S. Hossain, K. AlMutib, A. Ghoneim, S. A. AlQahtani, and F. M. Talaat, "Enhanced deep learning model for classification of retinal optical coherence tomography images," *Sensors*, vol. 23, no. 12, p. 5393, Jun. 2023, doi: 10.3390/s23125393.

[9] A. K. Dadzie, S. P. Iddir, S. Ganesh, B. Ebrahimi, M. Rahimi, M. Abtahi, T. Son, M. J. Heiferman, and X. Yao, "Artificial intelligence in the diagnosis of uveal melanoma: Advances and applications," *Experim. Biol. Med.*, vol. 250, p. 10444, Feb. 2025, doi: 10.3389/ebm.2025.10444.

[10] H. Wang, K. K. L. Chong, Z. Lin, X. Yu, and Y. Pan, "An explainable artificial intelligence-based robustness optimization approach for age-related macular degeneration detection based on medical IoT systems," *Electronics*, vol. 12, no. 12, p. 2697, Jan. 2023, doi: 10.3390/electronics12122697.

[11] D. Benet and O. J. Pellicer-Valero, "Artificial intelligence: The unstoppable revolution in ophthalmology," *Surv. Ophthalmol.*, vol. 67, no. 1, pp. 252–270, Jan. 2022, doi: 10.1016/j.survophthal.2021.03.003.

[12] H. N. Tukur, O. Uwishema, H. Akbay, D. Sheikhah, and I. F. S. Correia, "AI-assisted ophthalmic imaging for early detection of neurodegenerative diseases," *Int. J. Emergency Med.*, vol. 18, no. 1, p. 90, May 2025, doi: 10.1186/s12245-025-00870-y.

[13] A. Admin, "Through the eyes into the brain, using artificial intelligence," *Ann. Singap.*, vol. 52, no. 2, pp. 88–95. [Online]. Available: https://annals.edu.sg/through-the-eyes-into-the-brain-using-artificial-intelligence/

[14] J. D. Akkara, "Retinal revelations: Seeing beyond the eye with artificial intelligence," *Kerala J. Ophthalmol.*, vol. 36, no. 3, pp. 295–298, Dec. 2024, doi: 10.4103/kjo.kjo_124_24.

[15] S. Kumar, V. Dixit, and M. Gupta, "Deep learning approaches for retinal disease diagnosis: Insights from fundus and OCT analysis," in *Proc. EPJ Web Conf.*, vol. 328, 2025, p. 01041, doi: 10.1051/epjconf/202532801041.

[16] W. Hauri-Rosales, O. Pérez, M. Garcia-Roa, E. López-Star, and U. Olivares-Pinto, "Optimizing ocular pathology classification with CNNs and OCT imaging: A systematic and performance review," *medRxiv*, Jun. 2024, doi: 10.1101/2024.06.18.24309070.

[17] M. Elsharkawy, I. Abdelhalim, M. Ghazal, A. Mahmoud, H. S. Sandhu, A. Thanos, and A. El-Baz, "OCT-trans: A novel transformer backbone with multimodal feature extraction in OCT-based retinal disease classification," in *Proc. IEEE 22nd Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2025, pp. 1–4, doi: 10.1109/ISBI60581.2025.10980790.

[18] B. Ebrahimi, D. Le, M. Abtahi, A. K. Dadzie, J. I. Lim, R. V. P. Chan, and X. Yao, "Optimizing the OCTA layer fusion option for deep learning classification of diabetic retinopathy," *Biomed. Opt. Exp.*, vol. 14, no. 9, pp. 4713–4724, Sep. 2023, doi: 10.1364/boe.495999.

[19] J. Lo, T. T. Yu, D. Ma, P. Zang, J. P. Owen, Q. Zhang, R. K. Wang, M. F. Beg, A. Y. Lee, Y. Jia, and M. V. Sarunic, "Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data," *Ophthalmol. Sci.*, vol. 1, no. 4, Dec. 2021, Art. no. 100069, doi: 10.1016/j.xops.2021.100069.

[20] A. P. Sunija, S. Kar, S. Gayathri, V. P. Gopi, and P. Palanisamy, "OctNET: A lightweight CNN for retinal disease classification from optical coherence tomography images," *Comput. Methods Programs Biomed.*, vol. 200, Mar. 2021, Art. no. 105877, doi: 10.1016/j.cmpb.2020.105877.

[21] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[22] D. Kermany, "Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images," *Mendeley Data*, vol. 3, Jun. 2018, doi: 10.17632/rscbjbr9sj.3.

[23] (2026). *OCT Retina Disease Dataset*. Accessed: Jan. 14, 2026. [Online]. Available: https://www.kaggle.com/datasets/mohaimenulshawon/oct-ratina-disease

[24] S. Sotoudeh-Paima, "Labeled retinal optical coherence tomography dataset for classification of normal, drusen, and CNV cases," *Mendeley Data*, vol. 1, Oct. 2021, doi: 10.17632/8kt969dhx6.1.

[25] C. de Vente et al., "AIROGS: Artificial intelligence for RObust glaucoma screening challenge," 2023, *arXiv:2302.01738*.

[26] *Diabetic Retinopathy Detection Dataset*. Accessed: Jan. 14, 2026. [Online]. Available: https://kaggle.com/diabetic-retinopathy-detection

[27] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, Oct. 2019, doi: 10.1016/j.ins.2019.06.011.

[28] *Retinal OCT Image Classification–C8 Dataset*. Accessed: Jan. 14, 2026. [Online]. Available: https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8

[29] *OIMHS Dataset*, Figshare, London, U.K., Oct. 2023, doi: 10.6084/m9.figshare.23508453.v1.

[30] T. Hassan, "A composite retinal fundus and OCT dataset with detailed clinical markings of retinal layers and retinal lesions to grade macular and glaucomatous disorders," *Mendeley Data*, vol. 4, Sep. 2021, doi: 10.17632/trghs22fpg.4.

[31] U. Venugopal and A. Agur. (Jan. 2026). *Glaucoma Detection Dataset*. [Online]. Available: https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection

[32] *OCTDL: Optical Coherence Tomography Dataset*. Accessed: Jan. 14, 2026. [Online]. Available: https://www.kaggle.com/datasets/orvile/octdl-optical-coherence-tomography-dataset

[33] U. S. Kim, "Machine learn for glaucoma," *Harvard Dataverse*, Nov. 2018, doi: 10.7910/dvn/1yrrac.

[34] Y. Chen. (Jan. 7, 2026). *Yiweichen04/retinadataset*. Accessed: Jan. 14, 2026. [Online]. Available: https://github.com/yiweichen04/retinadataset

[35] *REFUGE2 Dataset*. Accessed: Jan. 14, 2026. [Online]. Available: https://www.kaggle.com/datasets/victorlemosml/refuge2

[36] M. R. Rashid, S. Sharmin, T. Khatun, M. Z. Hasan, and M. S. Uddin, "Eye disease image dataset," *Mendeley Data*, vol. 1, Apr. 2024, doi: 10.17632/s9bfhswzjb.1.

[37] *Comparison of Age-Related Macular Degeneration Treatments Trials (CATT)*. Accessed: Jan. 14, 2026. [Online]. Available: https://www.med.upenn.edu/cpob/catt

[38] A. Hussien, A. Elkhateb, M. Saeed, N. M. Elsabawy, A. E. Elnakeeb, and N. Elrashidy, "Explainable self-supervised learning for medical image diagnosis based on DINO V2 model and semantic search," *Sci. Rep.*, vol. 15, no. 1, p. 32174, Sep. 2025, doi: 10.1038/s41598-025-15604-6.

[39] R.-A. Bourceanu, N. De La Fuente, J. Grimm, A. Jardan, A. Manucharyan, C. Weiss, D. Cremers, and R. Pflugfelder, "Foundations and models in modern computer vision: Key building blocks in landmark architectures," 2025, *arXiv:2507.23357*.

[40] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. Camara, A. M. Kain, C. Saillard, and J.-B. Schiratti, "Scaling self-supervised learning for histopathology with masked image modeling," *medRxiv*, Dec. 2024, doi: 10.1101/2023.07.21.23292757.

[41] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "IBOT: Image BERT pre-training with online tokenizer," 2021, *arXiv:2111.07832*.

[42] A. Filiot, P. Jacob, A. Mac Kain, and C. Saillard, "Phikon-v2, a large and public feature extractor for biomarker prediction," 2024, *arXiv:2409.09173*.

[43] G. Ravi, "Content-based image retrieval using deep feature extraction with ResNet-50," *IJSREM*, vol. 9, no. 5, pp. 1–9, Jan. 2025. [Online]. Available: https://ijsrem.com/download/content-based-image-retrieval-using-deep-feature-extraction-with-resnet-50/

[44] H. Yan, V. Mubonanyikuzo, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Hybrid-RViT: Hybridizing ResNet-50 and vision transformer for enhanced Alzheimer's disease detection," *PLoS ONE*, vol. 20, no. 2, Feb. 2025, Art. no. e0318998, doi: 10.1371/journal.pone.0318998.

[45] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629, doi: 10.1109/ICCV48922.2021.00950.

[46] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Wozniak, and M. F. Ijaz, "Implementing vision transformer for classifying 2D biomedical images," *Sci. Rep.*, vol. 14, no. 1, p. 12567, May 2024, doi: 10.1038/s41598-024-63094-9.

[47] A. Vanyan, A. Barseghyan, H. Tamazyan, V. Huroyan, H. Khachatrian, and M. Danelljan, "Analyzing local representations of self-supervised vision transformers," 2023, *arXiv:2401.00463*.

[48] J. Zhang, J. Wang, Z. Sun, J. Zou, and R. Balestriero, "FastDINOv2: Frequency based curriculum learning improves robustness and training speed," 2025, *arXiv:2507.03779*.

[49] C. Vairetti, S. Maldonado, L. Cuitino, and C. A. Urzua, "Interpretable multimodal classification for age-related macular degeneration diagnosis," *PLoS ONE*, vol. 19, no. 11, Nov. 2024, Art. no. e0311811, doi: 10.1371/journal.pone.0311811.

[50] D. T. Huff, A. J. Weisman, and R. Jeraj, "Interpretation and visualization techniques for deep learning models in medical imaging," *Phys. Med. Biol.*, vol. 66, no. 4, Feb. 2021, Art. no. 04TR01, doi: 10.1088/1361-6560/abcd17.

[51] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2017, *arXiv:1704.02685*.

[52] J. M. Metsch and A.-C. Hauschild, "BenchXAI: Comprehensive benchmarking of post-hoc explainable AI methods on multi-modal biomedical data," *Comput. Biol. Med.*, vol. 191, Jun. 2025, Art. no. 110124, doi: 10.1016/j.compbiomed.2025.110124.

[53] M. W. Shinkle and M. D. Lescroart, "Visualizing and controlling cortical responses using voxel-weighted activation maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2025, pp. 4825–4829, doi: 10.1109/CVPRW67362.2025.00471.

[54] A. Osa-Sanchez, H. M. Balaha, A. Mahmoud, A. Sewelam, M. Ghazal, B. Garcia-Zapirain, and A. El-Baz, "Explainable AI-based approach for age-related macular degeneration (AMD) detection via fundus imaging," *IEEE Access*, vol. 13, pp. 341–360, 2025, doi: 10.1109/ACCESS.2024.3522862.

AINHOA OSA-SANCHEZ received the B.S. degree in industrial electronics and automation engineering from the University of Deusto, Bilbao, Spain, in 2021, and the M.S. degree in industry 4.0 from the International University of La Rioja, in 2022. She is currently pursuing the Ph.D. degree with the eVIDA Research Group, University of Deusto. Her doctoral research has evolved from wearable-based pain detection using EEG and NIR signals to the development of explainable, large-scale foundational AI models for ophthalmic diagnosis. Since June 2024, she has been a Visiting Researcher with the Department of Bioengineering, University of Louisville, USA, leading the international research project on Multimodal Age-Related Macular Degeneration (AMD) stratification. Her current work focuses on self-supervised learning, vision transformers, and the autonomous discovery of clinical biomarkers. She has co-authored several high-impact publications in international journals and conferences, specializing in the intersection of deep learning and medical imaging.

AYMAN EL-BAZ (Fellow, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering, in 1997 and 2001, respectively, and the Ph.D. degree in electrical engineering from the University of Louisville, in 2006. He is currently a Professor with University Scholar, and the Chair of the Department of Bioengineering, University of Louisville. He is also a Distinguished Professor with the University of Louisville and Alamein International University (UofL-AIU), New Alamein City, Egypt. He has two decades of hands-on experience in the fields of bio-imaging modeling and non-invasive computer-assisted diagnosis systems. He has authored or co-authored more than 700 technical articles (168 journals, 44 books, 85 book chapters, 255 refereed-conference papers, 196 abstracts, and 36 U.S. patents). He was named as a fellow for Coulter, AIMBE, and NAI for his contributions to the field of artificial intelligence in medicine and biomedical translational research.

IBON OLEAGORDIA-RUIZ received the degree in physics (specializing in electronics and automation), in 1999, the degree in electronics engineering from the University of the Basque Country, Bilbao, Spain, in 2001, and the Ph.D. degree in computer science from the University of Deusto, Bilbao. He joined the Faculty Member of Engineering, University of Deusto, as co-responsible for the electronics laboratories. He is currently a Lecturer in charge of the Ph.D. Program with the Department of Mechanics, Design, and Industrial Organization. His teaching activity focuses on mathematics, calculus, and differential equations, holding the Label two seal of teaching accreditation. He is a Researcher with the eVIDA Research Group, which has been recognized as a Type-A Group by the Basque Government and as a European Living Laboratory (ENoLL). From 2008 to 2018, he was a Researcher with the Deusto Tech-Life Unit. Since 2003, he has participated in more than 120 national and international research projects. He has published more than 30 articles in international journals, authored several book chapters, registered multiple patents, and presented more than 60 papers at international conferences. He is currently a member of the Kronikgune Research Center of Excellence. In 2015, he defended the Ph.D. Thesis in computer science (with European mention) with the University of Deusto. He holds a recognized six-year research period (sexenio) from CNEAI. He has received several awards with the eVIDA Team, including the UD-Santander Research Award, in 2007 and 2015, and the ONCE Euskadi-Solidarios Award. He was a finalist for European Award for Social Innovation in Ageing, in 2014. He serves on the Program Committee for the ISSPIT Congress. He is a reviewer for various international scientific journals and conferences.

BEGONYA GARCIA-ZAPIRAIN was born in San Sebastian, Spain. She received the degree in communications engineering, specializing in telematics from Basque Country University, Bilbao, Spain, in 1994, and the Ph.D. degree in biomedical signal processing, in 2004. After four years of working for ZIV Company, in 1997, she joined as a Lecturer in signal theory and electronics with the Engineering School, University of Deusto. She was the Head of the Telecommunication Department, University of Deusto, from 2002 to 2008, and also the Head of the DeustoTech-LIFE Department, from 2008 to 2018. In 2001, she created the eVIDA Research Group (evida.deusto.es). She has participated in more than 100 research projects on international, national and regional levels, published more than 70 articles in international scientific ISI indexed journals, and presented more than 160 articles in international and national scientific conferences. She has been supervising more than ten theses and has ongoing five Ph.D. students. She has strong collaborations with research labs in USA, France, Ireland, and The Netherlands among others. In 2004, she defended the Ph.D. Thesis in biomedical signal processing research area. She has five research awards at national level.