medicina

MDPI

Article

MCG-Net: Multiscale Residual Context Learning for Retinal OCT Classification with External Model Comparison

Denizcan Ozizmirliler [1], Hasan Can Küçük [2], Shahid Mohammad Ganie [3,*], Mehmet Burukanli [4], Merve Ersoy [5], Faruk Ozger [6] and Ishak Pacal [6,7,8]

1 Department of Ophthalmology, Bakırköy Dr. Sadi Konuk Training and Research Hospital, Istanbul 34000, Türkiye; denizcanozizmirliler@gmail.com
2 Department of Ophthalmology, Uşak Training and Research Hospital, Uşak 64000, Türkiye; hsncnkck@gmail.com
3 Department of Health Information Management and Technology, College of Applied Medical Sciences, King Faisal University, Al-Ahsa 31982, Saudi Arabia
4 Department of Computer Engineering, Faculty of Engineering and Architecture, Bitlis Eren University, Bitlis 13000, Türkiye; mburukanli@beu.edu.tr
5 Department of Software Engineering, Faculty of Engineering and Natural Sciences, Istanbul Topkapi University, Istanbul 34000, Türkiye; merveersoy@topkapi.edu.tr
6 Department of Computer Engineering, Faculty of Engineering, Igdir University, Igdir 76000, Türkiye; faruk.ozger@igdir.edu.tr (F.O.); ishakpacal@igdir.edu.tr (I.P.)
7 Department of Electronics and Information Technologies, Faculty of Architecture and Engineering, Nakhchivan State University, Nakhchivan AZ 7012, Azerbaijan
8 Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Fenerbahce University, Istanbul 34000, Türkiye
* Correspondence: sganie@kfu.edu.sa

Abstract

Background and Objectives: Retinal optical coherence tomography (OCT) classification requires recognition of subtle lesions within their anatomical surroundings, yet expanding contextual coverage can obscure local structural information. MCG-Net integrates local morphology and multiscale context within a gated convolutional hierarchy. *Methods:* A retained depthwise local pathway is augmented by two compressed separable context branches with bounded channel-wise residual scales. This study distinguishes choroidal neovascularization, diabetic macular edema, drusen, and normal retina. OCT-C8 provides development images; independent CAVRI-H5 and OCTDL datasets provide 384 external observations. Both external datasets remain excluded from parameter fitting, development validation, and checkpoint selection, but their ablation results inform selection of the Full 11/23 architecture. Paired comparisons, source-specific assessment, and visual error analysis characterize this external model comparison. *Results:* MCG-Net achieved 92.01 $\pm$ 0.60% accuracy, 91.99 $\pm$ 0.61% macro-F1, and 89.47 $\pm$ 0.80% Matthews correlation coefficient. Combined context improved macro-F1 over the local baseline by 2.18 percentage points, with a paired 98% interval of 0.50–4.04 and adjusted $p = 0.019$. Exploratory comparisons identified differences against six alternatives, while advantages over the strongest comparators remained uncertain. Source-specific accuracy was 90.86% on CAVRI-H5 and 95.49% on OCTDL; drusen remained the most difficult category. The model contains 26.82 million parameters and has 4.28 ms median inference latency on an RTX 5090. *Conclusions:* Compressed residual context strengthens gated convolutional OCT classification without replacing local processing. Reuse of the external cohort for architecture selection limits confirmatory interpretation. Untouched external validation, complete-development patient metadata, and representative clinical sampling are needed before clinical use.

Check for updates

Academic Editor: Brent Siesky

Received: 7 September 2026
Revised: 13 October 2026
Accepted: 16 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the Lithuanian University of Health Sciences. 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.

Keywords: artificial intelligence; deep learning; optical coherence tomography; retinal disease classification; medical diagnosis; explainable artificial intelligence; computer-assisted diagnosis

1. Introduction

Optical coherence tomography (OCT) provides noninvasive cross-sectional visualization of retinal microstructure [1]. Its depiction of the neurosensory retina, retinal pigment epithelium (RPE), and adjacent tissue supports the assessment and management of macular disease [2]. High-resolution retinal sections reveal fluid compartments, layer deformation, and focal tissue changes that are difficult to characterize from en face appearance alone [3]. Interpretation requires relating local abnormalities to their anatomical position and surrounding retinal architecture [4].

The present task distinguishes choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal retinal appearance [5,6]. Intraretinal cysts, retinal thickening, and fluid distribution are characteristic OCT findings used to assess DME [7,8]. Drusen appear as focal elevations of the RPE and may require differentiation from shallow structural changes in otherwise normal scans [9,10]. CNV and drusen can occur within the clinical spectrum of age-related macular degeneration (AMD), so repository-defined image categories should not be treated as mutually exclusive clinical diagnoses [11].

Deep learning learns image representations directly from data and has supported OCT disease classification and referral assessment. These developments connect local lesion recognition with broader advances in medical image analysis [12–14]. Beyond ophthalmology, recent studies have examined adaptive saliency networks and explainable vision Transformers for brain MRI classification, together with boundary-aware brain tumor segmentation. Related work on colonoscopy polyp detection, breast ultrasound classification, and scale- and boundary-aware multi-finding detection in panoramic dental radiographs further demonstrates the applicability of deep learning across medical imaging modalities [15–20]. Such successes motivate anatomically appropriate OCT representations rather than direct assumptions that an architecture transfers unchanged between modalities. Convolutional feature extraction remains relevant because it preserves useful local inductive biases for boundaries, fluid compartments, and texture [21].

The principal architectural difficulty is to broaden contextual coverage while retaining access to subtle local abnormalities. Transformer, hybrid, and self-supervised OCT models address this difficulty through different mechanisms and training resources [22–26]. Gated convolution offers an alternative spatial hierarchy in which a wider context can be introduced through residual convolutional paths. Its value must be assessed together with patient metadata, source composition, matched optimization, and paired external errors rather than inferred from an internal accuracy ranking [27–30].

MCG-Net addresses this representation problem by coupling a retained local depthwise pathway with two compressed context branches within a single gated hierarchy. Bounded channel-wise residual scales control the additional spatial responses before feature gating. CAVRI-H5 and OCTDL provide data from separate subjects and acquisition sources for external model comparison; their ablation results also informed selection of the contextual configuration. The principal contributions are as follows:

- MCG-Net integrates local 7×7 depthwise (DW7) processing with compressed separable context branches of sizes 11 and 23, extending spatial coverage within one gated feature hierarchy without introducing a second feature-extraction backbone.

- Bounded channel-wise residual scaling regulates each context branch before feature gating and provides explicit control over multiscale spatial adaptation.
- A matched A1/A2/A3/Full experiment isolates individual and combined context contributions while keeping shared initialization, gating, stage transitions, and the classification head consistent.
- Subject-level external model comparison on CAVRI-H5 and OCTDL evaluates models fitted on OCT-C8, with one selected B-scan per subject, explicit source-by-class accounting, and disclosure of external-data-informed architecture selection.
- Run-aware paired comparisons link repeated training runs to class-specific and source-specific errors, prevalence-sensitive predictive values, and measured inference costs, yielding an assessment that accounts for sampling dependence and external source composition.

The remainder of this paper is organized as follows. Section 2 reviews related OCT classification approaches. Section 3 describes the datasets, curation, external cohort, MCG-Net architecture, training protocol, and statistical analysis. Section 4 reports the external comparison, ablation, kernel, computational, prevalence, and visual analyses. Section 5 discusses the findings and limitations, and Section 6 presents the conclusions.

2. Related Work

Retinal OCT classification has developed along several overlapping lines: morphology-sensitive convolution, attention-based contextual modeling, hybrid feature extraction, and transferable representation learning [31]. Recent work also examines class imbalance, prediction reliability, and acquisition-source variation. Local morphology, spatial context, and acquisition-source variation jointly influence representation design and external evaluation.

2.1. CNN-Based and Morphology-Sensitive OCT Classification

Early OCT classification studies used retinal texture and geometry to distinguish disease categories. Srinivasan et al. investigated multiscale descriptors for DME and dry AMD, and Kermany et al. demonstrated large-scale image-based deep learning for retinal disease classification [5,32]. Convolutional neural networks (CNNs) provide a natural representation for fluid compartments, retinal boundaries, and RPE deformation because they reuse local filters across the image. General-purpose backbones such as EfficientNetV2 improve the efficiency of feature extraction through network scaling and training-aware design, although their original evaluations do not establish performance on the present OCT task [21].

OCT-specific methods adapt this processing to image noise and lesion morphology. Karthik and Mahadevappa combined an OCT-oriented input representation with adaptive convolution [33]. Pan et al. integrated lightweight local processing, global encoding, and channel attention [34]. These approaches illustrate how computational constraints influence the representation of subtle structural findings. Rahmani and Ghatee pursued a different efficiency objective through early-exit inference, allowing computation to depend on the difficulty of an individual image [35]. A fixed-path classifier and an early-exit system consequently require different interpretations of average inference cost.

Vijayan et al. added spatial-adaptive multiscale processing to ConvNeXt-Tiny and evaluated predictive calibration and robustness to image perturbations [36]. Leyba-Mesa and Barkana addressed class imbalance using an auxiliary classifier on intermediate features jointly optimized with inverse-frequency weighting [37]. The former emphasizes representation scale and reliability, while the latter examines minority-class learning. Together, they show why aggregate accuracy should be complemented by class-specific errors

and a description of the conditions under which predictions are evaluated. MCG-Net examines a distinct architectural variable: the effect of adding two compressed context scales to a retained local gated-convolutional backbone.

2.2. Transformers and Self-Supervised OCT Representation Learning

Vision Transformers model relationships among image tokens through self-attention, while Swin Transformer organizes attention within shifted windows and a hierarchical representation [22,23]. For OCT, contextual modeling is relevant when a focal abnormality must be interpreted relative to the foveal contour, layer continuity, or wider RPE configuration. Shen et al. introduced a structure-oriented Transformer for retinal disease grading; He et al. investigated an interpretable Transformer classifier; and Akca et al. evaluated Vision Transformer classification of CNV, DME, and drusen [24,38,39]. CRAT further refines attention through re-attention and class-re-attention to strengthen feature interaction and class-oriented representation [40].

Self-supervised learning addresses the related problem of acquiring transferable representations from incompletely labeled and heterogeneous data. OCT-SelfNet combines OCT images from multiple sources during SwinV2-based masked-autoencoder pretraining, followed by supervised adaptation for AMD-versus-normal classification [41]. Multi-OCT-SelfNet extends multi-source pretraining to multiclass downstream tasks and reports evaluation across datasets [26]. These frameworks examine both the pretraining objective and the diversity of the data used to learn it. Their contribution is complementary to a comparison of spatial mixers within a fixed pretrained backbone. Durmaz Engin et al. also illustrate the importance of task-specific model development in an expert-designed-versus-automated-AMD classification comparison [42].

Attention-based performance depends on spatial tokenization, pretraining, augmentation, and optimization as well as architecture. A comparison conducted with a common optimizer describes performance under that shared protocol; it does not establish each model's individually optimized operating point. Similarly, an improvement obtained through additional multi-source pretraining cannot be attributed solely to the downstream network architecture. These distinctions are necessary when comparing supervised convolutional, Transformer, and self-supervised OCT systems.

2.3. Hybrid Contextual Models and Gated Convolutional Architectures

Hybrid CNN and Transformer models integrate local feature extraction with attention-based context. HTC-Retina combines convolutional and Transformer representations, whereas HyReti-Net fuses local ResNet and global Swin features through a dual-branch design [25,43]. MergeNeXt continues this hybrid direction for OCT classification [44]. These models make the integration of local and broader information explicit, but the additional feature streams and fusion operations must be included in computational comparisons. At a different level of analysis, lesion-attentive fusion aggregates disease evidence across OCT B-scans [45]. Its volume-level endpoint differs from classification of a selected two-dimensional B-scan.

Gated convolutional hierarchies offer another route to contextual image processing. The gated CNN of Yu and Wang uses convolutional blocks for visual recognition without relying on a selective state-space mixer [27]. Large-kernel designs, including Large Separable Kernel Attention and the Visual Attention Network, also motivate efficient spatial aggregation through convolutional operations [46,47]. This study provides context for multiscale spatial processing, but the architectural operations must be distinguished precisely. MCG-Net implements two compressed residual context branches alongside the native DW7 path and retains the backbone's Gaussian error linear unit (GELU) gate, convolutional

transitions, and average-pooling head. The branch-specific residual scales control how much additional context enters the shared feature hierarchy.

Table 1 integrates architectural strategy with selected quantitative results reported in the original publications. Class definitions, sampling, patient grouping, training resources, and evaluation protocols differ, so those results contextualize the literature rather than provide a controlled ranking or a cross-study estimate of computational advantage.

Table 1. Representative recent OCT methods: Architectural strategy, dataset, evaluation scope, published performance, and reported computational cost. Multiple outcomes follow the listed dataset order; protocols differ and do not support a controlled cross-study ranking.

Study	Strategy	Dataset/Classes	Evaluation Scope	Accuracy/F1 (%)	M Params	GFLOPs
He et al. (2023) [38]	Interpretable Transformer	OCT2017/4; OCT-C8/8	Within-dataset benchmarks	99.80/99.80; 97.12/97.10	–	–
HTC-Retina (2024) [25]	CNN–Transformer hybrid	OCT2017/4; OCT-C8/8; OCT2014/3	Separate dataset benchmarks	99.40/–; 97.00/–; 99.77/–	–	–
Pan et al. (2025) [34]	Local/global encoding	OCT-C8/8; OCT2017/4	Within-dataset benchmarks	98.00/–; 99.85/99.70	–	–
HyReti-Net (2025) [43]	ResNet–Swin fusion	OCT2017/4; OCT2014/3; OCT-C8/8	Image-wise 70/20/10 split	99.06/99.05; 99.44/99.16; 98.76/98.33	–	–
OCT-SelfNet (2025) [41]	Multi-source self-supervision	Kermany/2	SwinV2, on-domain binary task	96.00/84.00	34.27	3.34
Multi-OCT-SelfNet (2026) [26]	Multiclass self-supervision	Kermany/4	SwinV2, on-domain DS1; multi-source SSL	90.00/90.00	34.27	3.34
ConvNeXt-SAMC (2026) [36]	Spatial multiscale adaptation	OCT-C8/8	Within-dataset test	98.39/98.39	–	–
PhysioSpeck-Net (2026) [48]	Physics-guided separation	OCT2017/4; OCT-C8/4	Test 1; external Test 2 without fitting	99.00/99.00; 97.36/97.35	8.80	–
MCG-Net (this study)	Gated local/residual context	OCT-C8/4; external/4	Independent CAVRI-H5/OCTDL; 384 subjects	92.01/91.99	26.82	9.97 *

–: Comparative data unavailable. F1 uses each publication's averaging definition (PhysioSpeck-Net: weighted F1). SSL: Self-supervised learning. Published FLOP conventions, input sizes, and hardware differ; no normalization is attempted. * MCG-Net uses twice the THOP-counted multiply–accumulate operations (MACs), excluding functional attention matrix products. The MCG-Net parameter count uses its four-output configuration; other counts follow the source publications.

2.4. Multi-Source Transfer, Robustness, and Prediction Reliability

OCT repositories differ in acquisition characteristics, image processing, disease composition, reference labels, and patient metadata. OCTDL provides multiclass images with patient information, while CAVRI-H5 contributes a separate acquisition source with one image per documented patient [28,29]. Their availability supports external assessment, but combining sources requires explicit label harmonization. Related clinical diagnoses can overlap in their OCT appearance, requiring explicit reference-label mapping. When a disease category occurs in only one source, a classifier may exploit source characteristics as well as disease morphology; source-stratified reporting can reveal heterogeneity but cannot eliminate this confounding.

Ajmal et al. examined external OCT evaluation, emphasizing the distinction between internal discrimination and transfer across acquisition sources [30]. Multi-source pretraining addresses representation transfer, as illustrated by OCT-SelfNet and Multi-OCT-SelfNet, but the source composition of pretraining, development, and evaluation data must still be disclosed [26,41]. Patient- or volume-level grouping is preferable when identifiers are available because related B-scans are correlated. Where identifiers are absent, image-level separation cannot be described as verified patient-level independence. The present study compares OCT-C8-trained gated CNN configurations on CAVRI-H5, published in 2026, alongside OCTDL, with the external observations excluded from parameter fitting and checkpoint selection [28,29]. This design provides evidence on a recent acquisition source without treating results from earlier, differently constructed benchmarks as directly comparable.

Reliability also concerns how confidence and explanation relate to classification errors. DAVis-Net combines attention and deep supervision with Monte Carlo dropout, calibration, split conformal prediction, and quantitative attribution analysis [49]. Its reported reliability estimates are internal, illustrating why a broad reliability framework and external validation answer different questions. PhysioSpeck-Net examines acquisition-related variation by separating feature branches according to OCT image-formation effects and testing cross-dataset behavior [48]. This physics-guided approach differs from learning contextual responses directly from image features. Neither an external discrimination estimate nor qualitative visualization alone establishes calibrated probabilities, reliable abstention, or clinical safety.

Balanced cohorts support comparison across classes but do not reproduce clinical disease prevalence. Precision and predictive value may therefore change even if sensitivity and specificity are stable. Grad-CAM and Grad-CAM++ can help inspect regions associated with a prediction [50,51], but plausible heatmaps do not validate lesion segmentation or causal explanation. Class-specific errors, confidence measures, source effects, and the evaluation unit must be considered together when interpreting an OCT classifier.

The contribution of compressed spatial context can be evaluated within a fixed backbone and a common training protocol. A1 retains native DW7; A2 adds context 11; A3 adds context 23; and Full combines both context branches. Shared gating, transitions, and classifier and matched initialization allow this comparison to focus on the additional context paths. Development-only checkpoint selection, repeated training runs, and paired external predictions connect the architectural comparison to a defined assessment of transfer. External outcomes informed configuration selection, so the same cohort provides exploratory model comparison rather than untouched confirmation of a previously fixed architecture.

3. Materials and Methods

3.1. Study Design and Classification Scope

This study addresses four-class classification of retinal OCT B-scans: CNV, DME, DRUSEN, and NORMAL. OCT-C8 supplies model-fitting and development-validation data [6], while CAVRI-H5 and OCTDL provide separate subject-level data for external model comparison [28,29]. Dataset-provided clinical labels define the reference standard. The official OCT-C8 test partition is unused, and external observations are excluded from optimization and checkpoint selection.

The four-category development dataset contains 9024 training and 1272 validation images, with 2256 training and 318 validation images per category. The external dataset contains 384 B-scans, with 96 per category: 288 from CAVRI-H5 and 96 from OCTDL. CAVRI-H5 and OCTDL are independent external datasets, and each selected observation represents one distinct subject. Neither dataset contributes to parameter fitting, develop-

ment validation, or checkpoint selection. The external ablation and kernel outcomes did, however, contribute to selection of the final architecture; independence of the data sources therefore does not imply independence of the final configuration from those outcomes. OCT-C8 patient and OCT-volume identifiers are unavailable, so development partitions cannot be described as verified patient-disjoint splits.

All models were trained for 300 epochs with AdamW and batch size 16 in three separate runs using seeds 0, 1, and 2. The same four-category output mapping was used for the principal comparison, architectural ablations, and kernel experiments. Numerical summaries report the mean and sample standard deviation (SD) across these three runs.

3.2. OCT-C8 Dataset and Curation

The OCT-C8 release provides eight repository categories; this study uses its CNV, DME, DRUSEN, and NORMAL images, with 2300 training and 350 validation images per selected category [6]. Native image dimensions are heterogeneous (512–1536 pixels wide and 496–512 pixels high). Images are assigned to development partitions before augmentation. Integrity checks use SHA-256 on decoded pixel arrays to detect exact matches and perceptual hashes to flag structural similarity; candidate assessment also uses dHash, structural similarity, and normalized root mean square error. All flagged cross-partition perceptual candidates follow the same conservative retention rule: the validation-side image is excluded, while the training-side image is retained unless a cross-class conflict requires its exclusion. Hash equality and perceptual similarity are distinguished because layered OCT anatomy can yield similar perceptual hashes for different pixel arrays.

For the four retained categories, 304 of 10,600 images are removed: 12 cross-class exact conflicts, 89 within-partition exact duplicates, 26 validation copies of cross-split exact duplicates, one high-confidence cross-split near duplicate, 41 additional validation-side perceptual review candidates, and 135 images through class-balance downsampling. Each removed image is assigned to its first applicable exclusion criterion; overlapping flags are not counted twice. Within a partition, one deterministic representative of an exact duplicate is retained; for cross-split duplicates, the validation copy is removed. Balancing uses seed 20260929 and randomly downsamples each category to a common size within each partition: 2256 training images, the size of the smallest remaining training category, and 318 validation images, one fewer than the smallest remaining validation category (DRUSEN, 319); balancing therefore also removes one DRUSEN validation image. Table 2 gives the resulting counts. The resulting sequence is 10,600 source images, 10,431 images after integrity cleaning, and 10,296 images after balancing; the latter comprise 9024 training and 1272 validation images. These counts describe image-level curation, not exclusion of a documented number of patients.

Table 2. OCT-C8 counts for the four study categories before curation, after integrity cleaning, and after class balancing.

	Original		Cleaned		Retained Balanced	
Class	**Train**	**Val.**	**Train**	**Val.**	**Train**	**Val.**
CNV	2300	350	2269	332	2256	318
DME	2300	350	2282	337	2256	318
DRUSEN	2300	350	2256	319	2256	318
NORMAL	2300	350	2298	338	2256	318
Total	9200	1400	9105	1326	9024	1272

The final development dataset contains 10,296 images. Decoded-pixel hashing identified no retained exact cross-split overlap. Images implicated in the 36 cross-partition

perceptual-similarity candidate pairs were excluded from the retained development partitions. Candidate-pair counts are distinguished from numbers of removed images because a single image may participate in more than one pair. The separate development-to-external check found no shared decoded-pixel hashes and a minimum pHash distance of six bits. These checks address image overlap but cannot establish patient independence when subject identifiers are unavailable.

3.3. External Cohort and Class Distribution by Source

CAVRI-H5 supplies 609 B-scans, with one documented patient per image, and original counts of 96 CNV, 99 DME, 137 DRUSEN, 142 NORMAL, and 135 vitreomacular adhesion/traction (VMA–VMT) images [29]. Its official training, validation, and test partitions are pooled solely for external assessment; none is used to fit a model or select a checkpoint. The selected subset retains all 96 CNV images, 48 DME, 96 DRUSEN, and 48 NORMAL images, totaling 288 images from 288 distinct patients. Decoded-pixel hashing identified no exact duplicates in the pooled release. Original split counts and selected support appear in Table 3. VMA–VMT is outside the evaluation taxonomy. The four-category cohort size was set by the smallest available category, CNV, with 96 subjects. All 96 CAVRI-H5 CNV subjects were retained. CAVRI-H5 supplied all 96 selected DRUSEN subjects, while DME and NORMAL were each allocated 48 subjects from CAVRI-H5 and 48 from OCTDL. These quotas produced 96 subjects per category in the combined cohort and equal source contributions for the two shared categories. Within the eligible CAVRI-H5 categories exceeding their quotas, subjects were sampled randomly without replacement before model predictions or performance results were examined. Sampling did not use predicted labels, confidence, or correctness. The random-number seed for CAVRI-H5 sampling was not retained; the fixed image-selection manifest identifies the realized subset but does not allow exact regeneration of the random draw from a seed. Original source partition membership did not determine eligibility because all source partitions were pooled solely for external evaluation.

Table 3. Independent external datasets: available source images, selected observations, original partition support, and final class distribution. CAVRI-H5 and OCTDL are entirely excluded from model fitting, validation, and checkpoint selection.

Class	CAVRI-H5 Available	CAVRI-H5 Selected (Orig. Train/Val/Test)	OCTDL Available Images	OCTDL Selected (Distinct Patients)	Final *n*
CNV	96	96 (76/10/10)	–	0	96
DME	99	48 (37/6/5)	147	48 (48)	96
DRUSEN	137	96 (85/7/4)	–	0	96
NORMAL	142	48 (43/4/1)	332 (NO)	48 (48)	96
Total study-category support	474	288	479	96 (96)	384

CAVRI-H5 official partitions are pooled only for assessment; no model fitting uses them. Selected split triplets retain original train/validation/test labels. OCTDL patient selection uses seed 20260929 and one image per recorded patient globally. CAVRI-H5 also contains 135 VMA–VMT images outside the study taxonomy. A dash denotes no selected study-category support, not absence of other diagnoses in OCTDL. All selected external observations are independent subjects; original source partition names describe repository provenance only.

OCTDL provides image-level metadata with patient identifiers [28]. Metadata rows are matched to image files, and one image per patient is selected globally using selection seed 20260929. Its contribution to this study comprises 48 DME and 48 NO images, yielding 96 observations with distinct within-source identifiers. NO is mapped to NORMAL. Other repository diagnoses are outside the study taxonomy, and ancillary condition metadata are not used to create an additional DRUSEN category.

The external cohort contains 384 independent subject-level observations from two independent datasets. The selected OCTDL subset contains no repeated patient identifiers or decoded-pixel hashes. External-cohort membership was fixed before predictions were examined. Neither external dataset was used for parameter fitting, development validation, or checkpoint selection; external outcomes were subsequently used for architectural comparison and configuration selection. The development-to-external pixel-hash check found no exact image overlap.

Source–class association remains: CNV and DRUSEN occur only in CAVRI-H5, while DME and NORMAL have 48 observations from each source. This composition creates potential source confounding. Source-specific summaries specify their reference-label support and macro-averaging set. DME and NORMAL permit direct source-stratified comparisons, with the full four-category prediction rule maintained in both sources. The OCTDL subset alone does not provide external evidence for CNV or DRUSEN recognition.

3.4. Reference Labels and Clinical OCT Characteristics

Dataset-provided labels form the reference standard. The retained OCT-C8 and CAVRI-H5 categories are CNV, DME, DRUSEN, and NORMAL; OCTDL contributes DME and NO, with NO renamed NORMAL. This mapping aligns names without establishing identical diagnostic criteria across repositories. CNV and DRUSEN remain clinically related to age-related macular degeneration, and the four labels should be interpreted as repository-defined image categories rather than an exhaustive, mutually exclusive clinical disease taxonomy.

Two specialist ophthalmologists jointly assessed the repository labels in relation to visible OCT morphology through a shared review. Review was conducted together rather than as two independent readings. Separate reader–label sets were therefore not generated, and Cohen's kappa and initial disagreement counts were not calculated. The reference categories reflect repository labels supported by this joint morphological review; the procedure does not establish a diagnosis based on a complete clinical examination or multimodal imaging. Figure 1 presents annotated CNV, DME, and drusen examples together with normal B-scans. These illustrations are separate from the images used for model inference.

3.4.1. Choroidal Neovascularization

CNV denotes pathological vessel growth from the choroid into the sub-RPE and/or subretinal space and is commonly associated with neovascular AMD. Leakage can disturb the outer retina and produce exudation [4]. The CNV source label is retained without subdivision into neovascular subtypes or attribution of every case to a single etiology.

Structural OCT may show irregular RPE elevation, pigment epithelial detachment (PED), subretinal fluid, hyper-reflective material, and sometimes intraretinal fluid. Figure 1 illustrates RPE- and fluid-related changes. PED and subretinal fluid are not specific to CNV; their relationship to the RPE and surrounding retinal tissue is more informative than either finding in isolation [4,52].

3.4.2. Diabetic Macular Edema

DME involves fluid accumulation in the macula in diabetes-associated retinal vascular disease. Vascular leakage can produce diffuse thickening, cystoid spaces, and serous detachment separately or together [7,8]. OCT depicts the anatomical distribution of fluid and the associated deformation of the retinal layers.

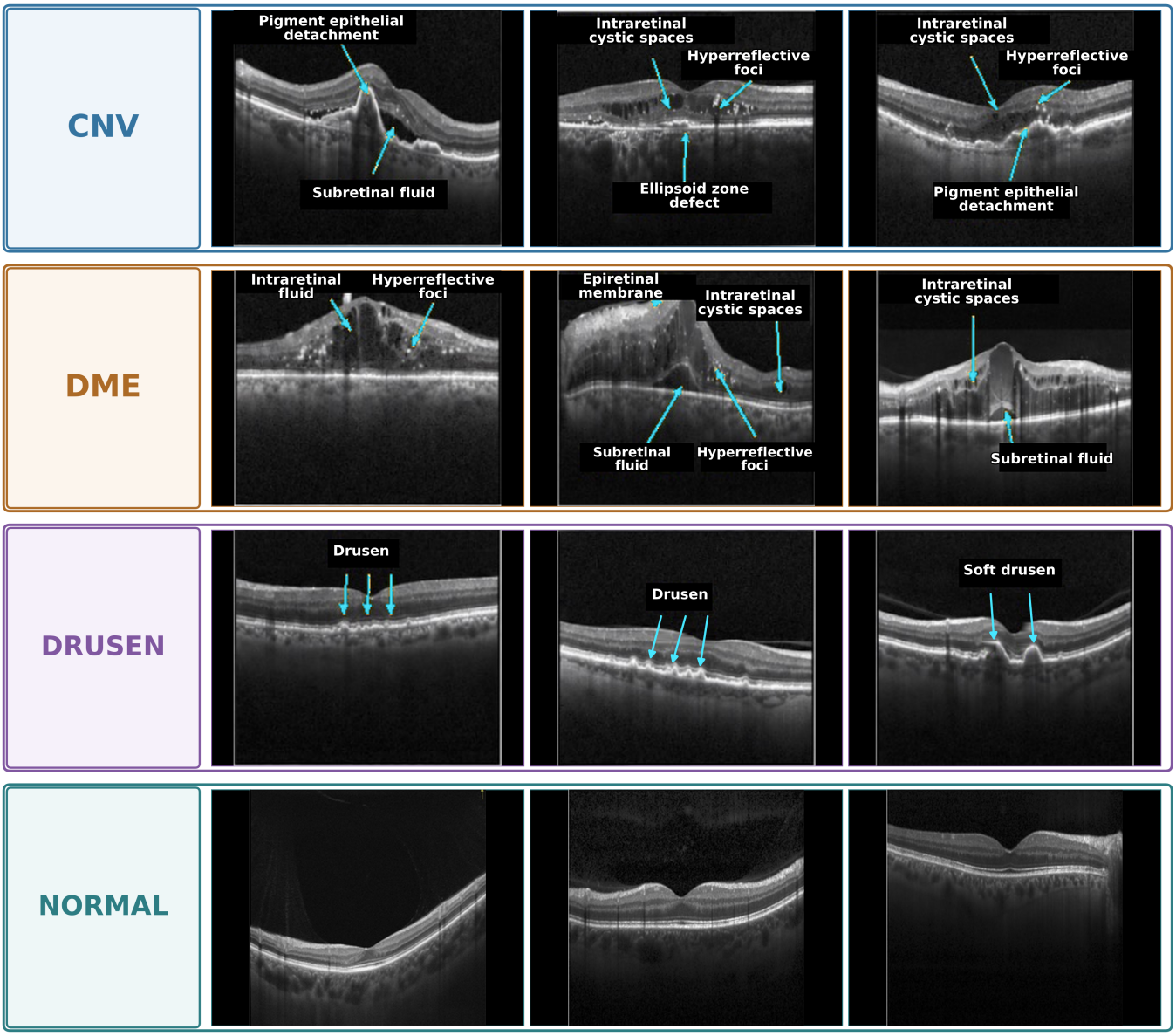


Figure 1. Representative OCT B-scans of the four study categories, with three examples per row. From top to bottom: CNV, DME, DRUSEN, and NORMAL. Annotations indicate pigment epithelial detachment (PED) and fluid-related changes in CNV, intraretinal cystic spaces and hyper-reflective foci in DME, and focal retinal pigment epithelium (RPE) elevations in drusen. Normal examples are unannotated. These illustrative images are separate from the evaluation cohort. OCT, optical coherence tomography.

Representative appearances include retinal thickening and rounded or ovoid hyporeflective intraretinal cavities; cysts may coalesce, and subretinal fluid can coexist. The illustrative examples show cavities and an altered macular contour. The fluid compartment helps distinguish intraretinal edema from predominantly serous detachment, while continuity of the surrounding retina distinguishes a fluid-filled cavity from a full-thickness foveal defect [7,53].

3.4.3. Drusen

Drusen are extracellular deposits beneath the RPE, classically between its basal lamina and the inner collagenous layer of Bruch's membrane [9]. Their significance depends on size and associated changes: small drusen alone do not establish AMD, whereas larger deposits

and pigmentary abnormalities contribute to risk classification [11]. Quantitative OCT assessment has also been investigated in intermediate AMD [10]. The retained DRUSEN label denotes a source-defined OCT phenotype and does not cover every form or stage of non-neovascular AMD.

Drusen produce localized RPE elevations with underlying material of variable reflectivity. Subretinal drusenoid deposits lie above the RPE and represent a different anatomical lesion [9]. Figure 1 shows focal elevations along the RPE profile. Accompanying exudation, RPE continuity, and surrounding outer-retinal changes are relevant to distinguishing this appearance from neovascular pathology [4].

3.4.4. Normal Retina

NORMAL denotes the source repository's normal-appearing B-scan category. Preserved retinal-layer organization, without the fluid compartments, pathological RPE deformation, or foveal tissue defects described above, provides the anatomical reference. This image-level definition does not establish that the entire eye is disease-free or exclude abnormalities outside the sampled scan.

The normal examples in Figure 1 retain their original source labels and have no added lesion annotations. They illustrate continuous retinal layers and preserved foveal anatomy. Clinical arrows and text in the pathological illustrations are not introduced into development or external inference images.

3.4.5. Differential Diagnoses Outside the Retained Categories

Central serous chorioretinopathy (CSC) can cause serous separation of the neurosensory retina from the RPE, sometimes with PED. Choroidal vascular hyperpermeability and impaired RPE barrier function are implicated [54]. OCT can depict a hyporeflective subretinal compartment, but subretinal fluid alone cannot establish CSC or distinguish it conclusively from neovascular disease [4,52].

A full-thickness macular hole (MH) interrupts foveal neurosensory tissue and may have elevated margins and adjacent cystoid changes. OCT-based classification distinguishes this defect from vitreomacular traction, partial-thickness abnormalities, and cavities within continuous retina [53]. CSC and MH are differential diagnoses outside the four-category classification endpoint.

3.5. Image Preprocessing and Augmentation

Images are decoded as three-channel inputs for source-compatible pretrained models. Grayscale OCT intensity is represented in three channels rather than interpreted as independent color information. Network input size is 224×224 pixels. Validation and external transforms are deterministic, with bicubic interpolation and a crop percentage of 1.0. Mean and standard deviation match the training transform: ImageNet statistics are used for all the principal models except EfficientNetV2-M and ViT-Base/16, which use mean and standard deviation 0.5 in every channel. Evaluation applies the same deterministic transforms consistently. Normalization statistics are not estimated from external images. Harmonization is limited to the declared label mapping, input geometry, and model-compatible intensity normalization. Preprocessing includes no scanner-specific intensity calibration or learned domain adaptation, and image quality was not independently adjudicated. Accordingly, preprocessing does not imply that acquisition differences between repositories have been removed.

The training transform uses random resized cropping with area scale $[0.08, 1.0]$ and aspect ratio $[0.75, 4/3]$, horizontal flipping with probability 0.5, no vertical flipping, color jitter of 0.4, and random training interpolation. The effect of these augmentations on lesion preservation was not separately quantified. Cropping can reduce visible anatomical

extent and intensity jitter can alter low-contrast cues. Mixup, CutMix, random erasing, and label smoothing are disabled. External evaluation uses native, unannotated images and introduces no stochastic training augmentation.

3.6. MCG-Net Architecture

MCG-Net uses a gated convolutional hierarchy with stage depths $(3, 3, 9, 3)$ and widths $(96, 192, 384, 576)$ [27]. The stem contains two 3×3 stride-2 convolutions, with intermediate normalization, GELU activation, and a final channel-wise normalization. For a $224 \times 224 \times 3$ input, the stem yields $56 \times 56 \times 96$ features. Three native 3×3 stride-2 convolutional transitions produce the resolutions in Table 4. Each transition applies convolution followed by normalization. The model contains 18 gated blocks and a global-average-pooling classification head. Figure 2 shows the four-class network and its multiscale context mixer.

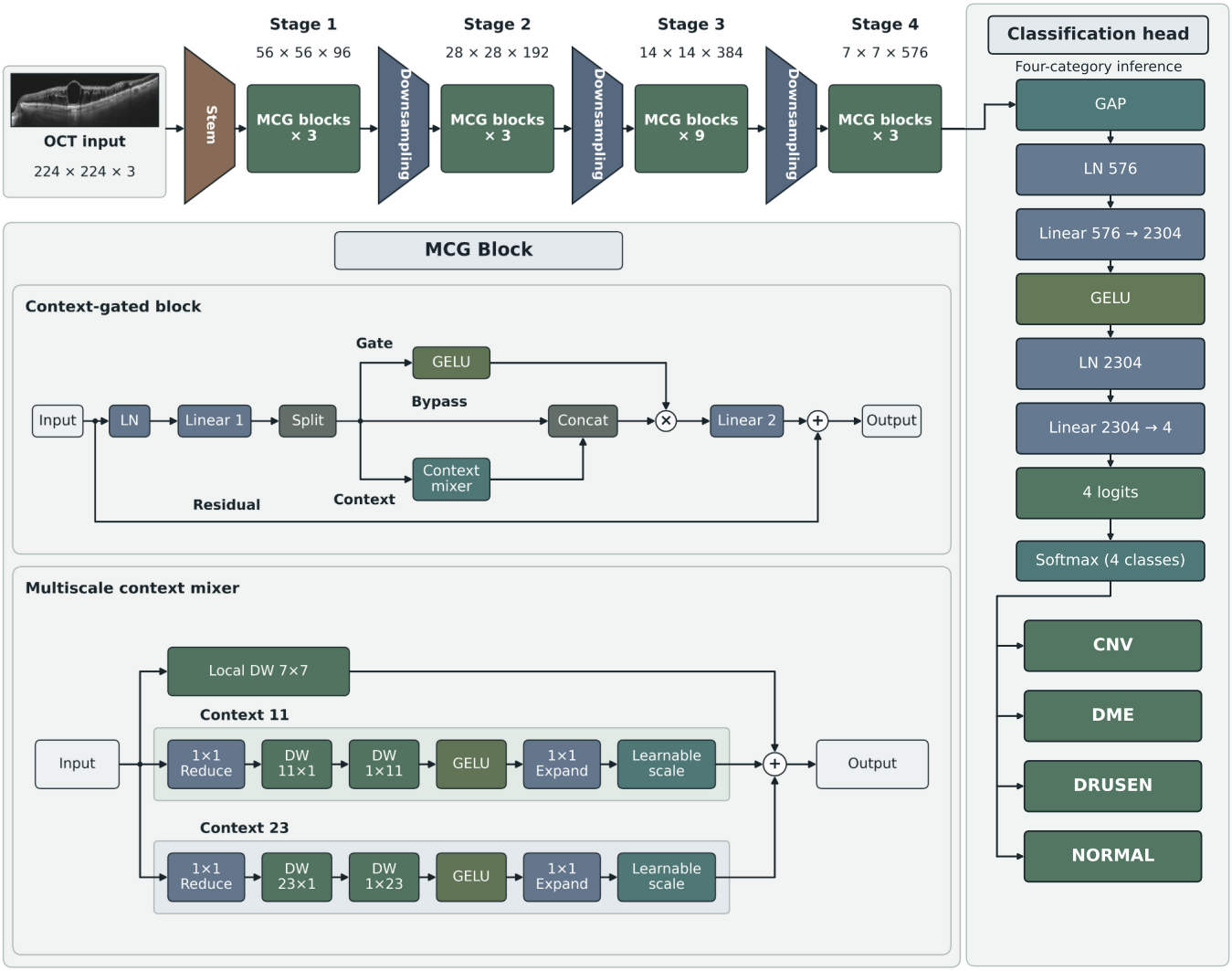


Figure 2. MCG-Net for four-class OCT classification. The gated CNN contains a convolutional stem, four feature stages, convolutional transitions, and a global-average-pooling classification head. Local DW7 processing is augmented by compressed separable context branches with kernels 11 and 23 and learned channel-wise residual scales. The final projection maps 2304 features to CNV, DME, DRUSEN, and NORMAL logits. LN, layer normalization; DW, depthwise convolution; GAP, global average pooling. Colors distinguish operation types (stem, downsampling, gated blocks, context branches, classification head) and carry no quantitative meaning.

Table 4. Stage geometry and four-output classification head of MCG-Net.

Component	Output Shape	Blocks	Operation
Stem	$56 \times 56 \times 96$	–	Two stride-2 3×3 convolutions
Stage 1	$56 \times 56 \times 96$	3	Gated blocks
Stage 2	$28 \times 28 \times 192$	3	Convolutional transition and gated blocks
Stage 3	$14 \times 14 \times 384$	9	Convolutional transition and gated blocks
Stage 4	$7 \times 7 \times 576$	3	Convolutional transition and gated blocks
Inference head	4 logits	–	Average pooling and $576 \to 2304 \to 4$ MLP

3.6.1. Native Gated Feature Processing

Let $X \in \mathbb{R}^{B \times H \times W \times C}$ denote a block input and $h = \lfloor 8C/3 \rfloor$. Normalization and a channel projection produce three feature partitions:

$$Z = \mathrm{LN}(X), \qquad [G, I, V] = \mathrm{Split}(W_1 Z + b_1), \tag{1}$$

where W_1 maps C channels to $2h$, and the gate, bypass, and spatial partitions have h, $h - C$, and C channels, respectively. Spatial operations act on V after conversion to a channel-first layout. The processed spatial features are concatenated with the bypass partition, yielding h channels, as follows:

$$T = \mathrm{Concat}(I, \mathcal{M}(V)). \tag{2}$$

The block output is

$$Y = X + \mathrm{DropPath}(\mathrm{LayerScale}(W_2[\mathrm{GELU}(G) \odot T] + b_2)). \tag{3}$$

LayerScale and DropPath are configured as identities for the matched ablation. GELU, $x\Phi(x)$, is the native activation [55]. Its output is not a probability and need not lie in $[0, 1]$; the term gate refers to element-wise feature modulation rather than a calibrated lesion-attention score. The residual connection preserves a direct path for the block input.

3.6.2. Compressed Multiscale Context Mixer

The native local response is

$$L(V) = \mathrm{DW}_{7\times7}(V). \tag{4}$$

For $k \in \{11, 23\}$, the context branch first reduces channels to $r = \lceil C/4 \rceil$ with a bias-free 1×1 projection P_k. Separable depthwise operations preserve spatial dimensions using same padding, and a bias-free projection Q_k expands the result back to C channels, as follows:

$$R_k(V) = Q_k[\mathrm{GELU}(\mathrm{DW}_{1\times k}(\mathrm{DW}_{k\times1}(P_k V)))]. \tag{5}$$

The two responses are added to the local path through learned channel-wise scales $a_k \in \mathbb{R}^C$, as follows:

$$\mathcal{M}(V) = L(V) + \tanh(a_{11}) \odot R_{11}(V) + \tanh(a_{23}) \odot R_{23}(V). \tag{6}$$

The scales are broadcast across batch and spatial dimensions. The tanh bound permits positive or negative residual modulation; it is not a softmax competition between branches. This context-scale mechanism is separate from the native GELU gate in Equation (3).

Compression precedes the spatial convolutions. Ignoring bias, the additional parameter count of a branch, including its residual scale, is

$$N_k = 2Cr + 2kr + C. \tag{7}$$

The depthwise spatial part costs $2kr$ weights rather than k^2r for a dense two-dimensional depthwise kernel at the same reduced width. The 1×1 projections remain part of the cost; separability alone is not taken as evidence of lower total cost than any other comparator. At late stages, a kernel can extend beyond the feature-map dimensions, so its nominal support is not interpreted as additional observed retinal area outside the image.

3.6.3. Pretrained Initialization and Residual Adaptation

The gated CNN hierarchy is initialized from ImageNet-1K pretrained weights of the source architecture [27,56]. Its 1000-class source projection is replaced by a four-output task head in the training protocol. The local source parameters provide the shared initialization for the context-augmented network. The context down-projections initially select evenly distributed channels, and their up-projections are initialized from a zero-mean normal distribution with standard deviation 0.02. For each selected local depthwise kernel K_c, a rank-one singular-value approximation provides separable factors:

$$K_c = U_c \Sigma_c V_c^{\mathsf{T}}, \qquad v_c = \sqrt{\sigma_{c,1}}\, u_{c,1}, \quad h_c = \sqrt{\sigma_{c,1}}\, v_{c,1}^{\mathsf{T}}. \tag{8}$$

The seven-element factors are centered within the 11- or 23-element filters, with zero values outside their support. These are source-derived initializations of new filters, distinct from directly transferred backbone tensors. All matched variants use the same pretrained source state.

Both context scales are initialized to zero. Consequently,

$$\mathcal{M}(V)|_{a_{11}=a_{23}=0} = L(V). \tag{9}$$

This preserves the native block function at initialization when the shared backbone and classifier states match. It does not imply equivalence after fine-tuning. Initially, the residual scales can receive gradients while branch-weight gradients are suppressed by zero scaling; branch weights can adapt once the scales move away from zero.

3.6.4. Four-Class Classification Head

Global average pooling aggregates the final $7 \times 7 \times 576$ features. The classification head applies LayerNorm, a $576 \to 2304$ linear layer, GELU, LayerNorm, and a $2304 \to 4$ output projection. With feature vector $u \in \mathbb{R}^{2304}$, the logits are $z = Wu + b$. The class order is $\mathcal{C}_4 = \{\text{CNV}, \text{DME}, \text{DRUSEN}, \text{NORMAL}\}$. Probabilities and the predicted class are

$$q_c = \frac{\exp(z_c)}{\sum_{j \in \mathcal{C}_4} \exp(z_j)}, \qquad \hat{y} = \arg\max_{c \in \mathcal{C}_4} q_c. \tag{10}$$

Reference labels and output indices are matched by semantic class name. Softmax scores express relative model confidence within the four-category task; they are not assumed to be calibrated clinical probabilities.

3.7. Optimization, Repeated Runs, and Checkpoint Selection

The training protocol comprised 300 epochs with batch size 16 and three independent runs per model using seeds 0, 1, and 2. These settings apply to comparators, architectural variants, and kernel variants (Table 5). AdamW uses initial learning rate 3×10^{-4}, weight decay 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ [57]. A five-epoch warm-up starts at 10^{-6} and is followed by cosine decay to 10^{-6}. All trainable layers are fine-tuned using four-class cross-entropy, as follows:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^{B} \log q_{i,y_i}. \tag{11}$$

Classifier dropout, stochastic depth, exponential moving averages, label smoothing, mixup, CutMix, and random erasing are disabled for the matched ablation. Evaluation and profiling use FP32 with AMP disabled.

For each 300-epoch run, the checkpoint with the highest development validation Top-1 accuracy was selected for evaluation. External images were not used for epoch selection.

Table 5. Training and evaluation settings shared by all models.

Setting	Configuration
Task/checkpoint criterion	CNV, DME, DRUSEN, NORMAL/best validation Top-1
Epochs/batch size	300/16
Runs/seeds per model	Three/0, 1, 2
Optimizer	AdamW; LR 0.0003; weight decay 0.05
Warm-up/schedule	5 epochs from 0.000001; cosine minimum 0.000001
Input/external interpolation	224 × 224/deterministic bicubic
Evaluation precision	FP32; AMP and test-time augmentation disabled
Disabled augmentation	Mixup, CutMix, random erasing, label smoothing

3.8. Comparators and Matched Architectural Ablation

The comparison includes 12 alternatives: DenseNet-169 [58], RDNet-Base [59], EfficientNetV2-M [21], StarNet-S3 [60], ConvNeXt-Base [61], PiT-B [62], InceptionNeXt-Base [63], Swin-Base [23], RegNetY-16GF [64], SHViT-S3 [65], DeiT III-Base [66], and ViT-Base/16 [22]. Models share the development partitions, 300-epoch AdamW schedule, and external observations, with compatible pretrained initialization and model-specific training normalization. Using one optimizer, schedule, training budget, and seed set controls these factors across architectures. The comparison therefore evaluates performance under a common protocol rather than each model's individually optimized maximum. Model-specific optimizer and schedule searches were not performed. All 13 models, including MCG-Net, have three evaluated runs, yielding 39 checkpoint evaluations in FP32.

A1 contains native DW7; A2 adds context 11; A3 adds context 23; and Full contains both context branches (Table 6). Within a seed, the common initialization begins with Full and prunes unused branches, preserving shared backbone, classifier, and surviving context tensors. The random state after common construction is restored to align subsequent stochastic initialization behavior. A2 and A3 isolate the two individual context scales, with native GELU gating and downsampling held constant.

Table 6. The four architectural variants. Native GELU gating, stem, stage transitions, and classifier construction are shared; the task head has four outputs in every variant.

Variant	Native DW7	Context 11	Context 23
A1	Yes	No	No
A2	Yes	Yes	No
A3	Yes	No	Yes
Full	Yes	Yes	Yes

The four-variant experiment comprises 12 runs: seeds 0, 1, and 2 for each of A1, A2, A3, and Full. A further kernel comparison fixes the first branch at 11 and sets the second to 7, 15, or 23. The 11/23 configuration is Full. The kernel experiment evaluates these three configurations. These experiments retain the native gate, stage transitions, and pooling, so their effect is attributed to added residual context rather than removal or replacement of those operations. Full with kernels 11 and 23 was selected from these external ablation and kernel results because it had the highest observed aggregate accuracy and macro-F1. It was then compared with the 12 alternative architectures on the same external cohort. Thus, the final configuration was not fixed before external outcomes were inspected. All

reported benchmark, ablation, and kernel endpoints are external test results; development validation was used only for checkpoint selection. No additional untouched cohort was reserved to confirm the selected configuration.

3.9. Four-Category Predictions and External Endpoints

The assessment uses the same ordered 384-sample list for each model and seed. Each prediction record associates a sample identifier with its source, reference label, model, run, and four-class output mapping.

The principal descriptive endpoints are accuracy and macro-F1, with macro-precision, macro-recall, multiclass Matthews correlation coefficient (MCC), one-versus-rest macro-averaged area under the receiver operating characteristic curve (AUROC), and area under the precision–recall curve (AUPRC) as complementary measures. AUROC/AUPRC use the four-class probabilities; all four labels enter their macro-average. For class c, one-versus-rest counts define

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Sensitivity}_c = \frac{TP_c}{TP_c + FN_c}, \quad \text{Specificity}_c = \frac{TN_c}{TN_c + FP_c}, \tag{12}$$

and

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \qquad \text{Macro-F1} = \frac{1}{4}\sum_{c \in \mathcal{C}_4} F1_c. \tag{13}$$

as the macro-precision and macro-recall average across the same four labels. A zero denominator receives a zero contribution and is disclosed. Support-weighted recall equals accuracy in single-label multiclass classification. In the balanced 384-image cohort, macro-recall also equals accuracy because each reference class has equal support; macro-F1 need not equal either. Confusion matrices have four reference rows and four prediction columns.

Accuracy, class-specific metrics, and macro metrics are calculated separately for each run, followed by their arithmetic mean and sample standard deviation. These summaries describe repeat-fit variability rather than the uncertainty of a population estimate. Source-specific summaries retain four possible predictions but average class metrics over the reference labels represented in that source: four for CAVRI-H5 and DME/NORMAL for OCTDL. Predicting CNV or DRUSEN for an OCTDL DME/NORMAL observation remains an error. Source comparisons do not replace the classifier with a binary DME-versus-NORMAL decision rule.

3.10. Run-Aware Uncertainty and Paired Comparisons

Two-sided 98% percentile intervals use 10,000 paired bootstrap replicates with resampling seed 20261004. Each replicate samples 384 external indices with replacement and uses identical indices for all models and fitted runs. Because the independent external datasets contribute one B-scan per selected subject, resampling observation indices is subject-level resampling. For metric m,

$$\bar{m} = \frac{1}{3}\sum_{r=0}^{2} m_r, \qquad \bar{m}^{(b)} = \frac{1}{3}\sum_{r=0}^{2} m_r(\mathcal{I}^{(b)}). \tag{14}$$

Metrics are computed separately for each fit before averaging within a replicate; predictions are not pooled into an enlarged independent sample. The 0.01 and 0.99 quantiles estimate sampling uncertainty conditional on the available fits and complement the across-run SD. Neither captures all uncertainty due to training choices or incompletely documented development grouping. The bootstrap and Holm adjustments condition on the evaluated configurations and do not account for selecting Full 11/23 from the same external results. Paired intervals and adjusted tests are therefore exploratory; they do not provide selection-adjusted confirmatory evidence for the chosen architecture.

For comparator q, the paired differences are

$$\Delta_q^{(b)} = \frac{1}{3}\sum_{r=0}^{2}\left[m_{\mathrm{Full},r}(\mathcal{I}^{(b)}) - m_{q,r}(\mathcal{I}^{(b)})\right]. \tag{15}$$

Approximate two-sided bootstrap p-values compare absolute centered bootstrap differences with the absolute observed difference, with Monte Carlo resolution 10^{-4}. Holm adjustment is applied across the 12 principal comparators separately for accuracy, macro-F1, and MCC, at $\alpha = 0.02$. Ablation comparisons form a separate family of five alternatives per metric. Nominal 98% intervals are not simultaneous multiplicity-adjusted intervals; a positive nominal interval does not guarantee that a Holm-adjusted test meets the threshold. These paired comparisons are exploratory and apply to the evaluated cohort and repeated-run protocol.

Exact two-sided McNemar tests compare paired correctness separately for each matched run. With n_{10} correct only for Full and n_{01} correct only for the comparator, the null distribution is binomial with $n_{10} + n_{01}$ trials and probability 0.5. Holm adjustment across the 36 main model-and-run tests controls this distinct family at $\alpha = 0.02$. Repeated predictions from the same subject are never treated as independent participants. The context interaction, $m_{\mathrm{Full}} - m_{\mathrm{A2}} - m_{\mathrm{A3}} + m_{\mathrm{A1}}$, is reported descriptively, with an interval, to assess whether the joint improvement exceeds the sum of the individual additions.

Prevalence scenarios estimate the positive predictive value (PPV) from one-versus-rest sensitivity (Se) and specificity (Sp) derived from the aggregate confusion counts, as follows:

$$\mathrm{PPV}(\pi) = \frac{\mathrm{Se}\,\pi}{\mathrm{Se}\,\pi + (1-\mathrm{Sp})(1-\pi)}. \tag{16}$$

The calculation uses unrounded counts at hypothetical prevalences of 1%, 5%, 10%, and 25%. It assumes stable sensitivity and specificity as prevalence changes; it is not a calibrated probability model or an estimate from a clinical prevalence cohort.

3.11. Computational Efficiency and Visual Error Analysis

Computational reporting uses the instantiated four-output models. Parameter count, the counted operation proxy, and measured timing are reported separately because they describe different aspects of computational cost.

Input size is 224×224. Operation counts are obtained with THOP, counting one multiply-and-accumulate operation as two FLOPs. Functional attention matrix products are excluded by that implementation, so estimates describe its counting convention rather than universally complete hardware costs. The GPU measurement protocol uses FP32, disables AMP and TF32, and runs without gradients. Batch-1 latency uses 100 warm-up iterations followed by 500 measurements; batch-16 throughput uses 20 timing windows of 50 iterations after warm-up. CUDA synchronization bounds timed regions, and measurements use an NVIDIA RTX 5090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), Python 3.12.12 (Python Software Foundation, Wilmington, DE, USA), PyTorch 2.9.0 (The Linux Foundation, San Francisco, CA, USA), CUDA 12.8 and cuDNN 9.10 (NVIDIA Corporation, Santa Clara, CA, USA), THOP (open-source, https://github.com/Lyken17/pytorch-OpCounter, accessed on 6 September 2026), and timm 1.0.29 (open-source, https://github.com/huggingface/pytorch-image-models, accessed on 6 September 2026). Median latency is reported. These measurements concern model inference rather than acquisition or clinical workflow.

Qualitative figures show selected correct predictions and errors drawn from the three MCG-Net fits trained with seeds 0, 1, and 2. Each displayed probability is obtained from the fitted model associated with that example. Grad-CAM++ is applied to that model's predicted-class score at `stages.3.blocks.1` [50,51]. Maps are normalized per image and

displayed with the corresponding B-scan. The prediction and attribution panels contain different selected cases; displayed probabilities and maps are individual-run outputs, not averages or ensemble outputs. The CAM rendering path includes image resizing and can produce small probability differences from the metric inference path, so figure probabilities are not used to recompute benchmark endpoints. Attribution remains a post hoc localization without lesion-mask comparison, perturbation-based faithfulness testing, or a new clinical reader adjudication.

3.12. Reproducibility and Interpretation

OCT-C8 supplied training images and development-validation images for checkpoint selection. CAVRI-H5 and OCTDL supplied all reported external performance endpoints for the comparator, ablation, and kernel experiments. External observations never entered parameter fitting or epoch selection. Their outcomes nevertheless informed the selection of Full 11/23, and the selected model was compared with the alternatives using those same observations. This design is an external model comparison with subject-level separation from fitting, not untouched external validation of a prespecified final architecture.

The clinical scope of this study was limited to four OCT image categories: CNV, DME, DRUSEN, and NORMAL. The identification of other retinal diseases and the rejection of previously unseen conditions were beyond the scope of the proposed framework.

OCT-C8 patient-level independence could not be verified because patient and volume identifiers were unavailable. The selected external observations represent distinct subjects, but reuse of their outcomes for architecture selection limits confirmatory interpretation. Differences in imaging devices, clinical settings, and patient populations may also affect model generalizability. Accordingly, the findings should be interpreted within the defined four-class OCT classification task. Further validation using independent, multicenter datasets acquired with different imaging systems is necessary to establish broader clinical applicability.

4. Results

4.1. External Performance and Paired Comparisons

On the same 384-image external cohort used for architectural selection, MCG-Net achieved the highest mean accuracy ($92.01 \pm 0.60\%$), macro-F1 ($91.99 \pm 0.61\%$), and MCC ($89.47 \pm 0.80\%$) among the 13 evaluated models (Table 7). Its run accuracies were 92.71%, 91.67%, and 91.67%. DenseNet-169 and RDNet-Base followed at 90.71% and 90.62% mean accuracy. EfficientNetV2-M attained the highest macro-AUROC (98.94%) and macro-AUPRC (97.35%), showing that the ranking of probability-based discrimination differed from the argmax ranking.

The paired macro-F1 difference between MCG-Net and DenseNet-169 was 1.37 percentage points (98% interval, −0.48 to 3.35; Holm-adjusted $p = 0.174$). Corresponding differences against RDNet-Base and EfficientNetV2-M were 1.43 (−0.18 to 3.15; $p = 0.135$) and 1.57 (−0.59 to 3.76; $p = 0.174$). Differences against StarNet-S3, ConvNeXt-Base, and PiT-B also failed to meet the adjusted $\alpha = 0.02$ threshold. Six comparisons (InceptionNeXt-Base, Swin-Base, RegNetY-16GF, SHViT-S3, DeiT III-Base, and ViT-Base/16) met that threshold for accuracy, macro-F1, and MCC (Table 8). These exploratory paired comparisons identify differences against part of the comparator set. The leading mean and adjusted tests remain conditional on selection of MCG-Net using the same cohort; superiority to the strongest alternatives remains uncertain.

The separate per-run McNemar analysis was more conservative: four of the 36 model-and-run tests met the Holm-adjusted threshold, involving SHViT-S3 run 0, ViT-Base/16 runs 0 and 2, and RegNetY-16GF run 0 (Table 8, Panel B). These tests assess discrete correctness within one fit, whereas the bootstrap compares the average metric across

three fits; their distinct inferential targets and multiplicity families account for different significance patterns. MCG-Net macro-F1 was 92.69%, 91.62%, and 91.65% for seeds 0, 1, and 2, respectively; the corresponding accuracy values are reported above.

Table 7. Four-category external assessment ($n = 384$): mean ± sample SD over three fitted runs, in percent. Parameter counts describe four-output models. Macro-recall equals accuracy because reference classes are balanced.

Model	Params (M)	Accuracy	Macro-P	Macro-F1	MCC	AUROC	AUPRC
MCG-Net	26.82	92.01 ± 0.60	92.33 ± 0.63	91.99 ± 0.61	89.47 ± 0.80	98.57 ± 0.06	96.62 ± 0.24
DenseNet-169	12.49	90.71 ± 0.84	91.48 ± 0.64	90.62 ± 0.87	87.91 ± 1.04	98.72 ± 0.30	97.09 ± 0.37
RDNet-Base	85.70	90.62 ± 0.26	91.10 ± 0.25	90.55 ± 0.29	87.70 ± 0.33	97.88 ± 0.20	95.86 ± 0.26
EfficientNetV2-M	52.86	90.54 ± 0.60	91.01 ± 0.77	90.42 ± 0.59	87.60 ± 0.86	98.94 ± 0.06	97.35 ± 0.27
StarNet-S3	5.49	90.02 ± 0.84	90.50 ± 0.77	89.98 ± 0.83	86.87 ± 1.09	98.36 ± 0.38	95.99 ± 0.72
ConvNeXt-Base	87.57	89.76 ± 1.17	90.28 ± 1.01	89.69 ± 1.22	86.55 ± 1.49	98.20 ± 0.21	96.11 ± 0.45
PiT-B	72.74	89.41 ± 0.80	90.29 ± 0.94	89.29 ± 0.82	86.23 ± 1.10	98.41 ± 0.10	96.07 ± 0.21
InceptionNeXt-Base	83.61	89.24 ± 1.57	89.88 ± 1.25	89.18 ± 1.55	85.88 ± 1.99	98.42 ± 0.20	96.31 ± 0.60
Swin-Base	86.75	88.98 ± 0.54	89.86 ± 0.68	88.82 ± 0.60	85.66 ± 0.74	98.31 ± 0.28	96.26 ± 0.42
RegNetY-16GF	80.58	88.72 ± 1.92	90.03 ± 1.50	88.54 ± 1.98	85.48 ± 2.39	98.34 ± 0.57	96.34 ± 1.35
SHViT-S3	13.80	88.19 ± 1.67	89.22 ± 1.58	88.11 ± 1.70	84.64 ± 2.20	98.24 ± 0.34	95.83 ± 0.79
DeiT III-Base	85.82	88.19 ± 0.84	89.29 ± 0.92	87.95 ± 0.81	84.73 ± 1.13	97.62 ± 0.13	94.14 ± 0.35
ViT-Base/16	85.80	86.46 ± 1.14	87.75 ± 0.62	86.18 ± 1.15	82.49 ± 1.34	97.58 ± 0.41	94.29 ± 0.96

AUROC and AUPRC are macro one-versus-rest values from the four-class probabilities. MCG-Net has the highest mean accuracy, macro-F1, and MCC; EfficientNetV2-M has the highest AUROC and AUPRC. Ranking alone does not establish a significant difference.

Table 8. Paired bootstrap differences, MCG-Net minus comparator, in percentage points. Brackets are nominal 98% intervals; p is Holm-adjusted separately across 12 comparisons per metric.

Panel A: Paired Bootstrap Differences and Adjusted Tests.

Comparator	ΔAccuracy	p (Holm)	ΔMacro-F1	p (Holm)	ΔMCC	p (Holm)
DenseNet-169	+1.30 [−0.52, 3.21]	0.199	+1.37 [−0.48, 3.35]	0.174	+1.56 [−0.83, 3.98]	0.229
RDNet-Base	+1.39 [−0.26, 3.12]	0.150	+1.43 [−0.18, 3.15]	0.135	+1.77 [−0.34, 3.96]	0.157
EfficientNetV2-M	+1.48 [−0.69, 3.65]	0.199	+1.57 [−0.59, 3.76]	0.174	+1.87 [−0.92, 4.64]	0.229
StarNet-S3	+2.00 [0.17, 3.82]	0.040	+2.00 [0.20, 3.86]	0.040	+2.60 [0.23, 5.03]	0.042
ConvNeXt-Base	+2.26 [0.52, 4.08]	0.023	+2.30 [0.54, 4.20]	0.022	+2.92 [0.65, 5.28]	0.023
PiT-B	+2.60 [0.69, 4.77]	0.023	+2.70 [0.77, 4.90]	0.022	+3.24 [0.77, 5.99]	0.027
InceptionNeXt-Base	+2.78 [1.04, 4.69]	0.004	+2.81 [1.07, 4.73]	0.004	+3.59 [1.31, 6.06]	0.004
Swin-Base	+3.04 [0.95, 5.21]	0.009	+3.17 [1.02, 5.43]	0.009	+3.81 [1.09, 6.59]	0.011
RegNetY-16GF	+3.30 [1.22, 5.56]	0.006	+3.45 [1.36, 5.70]	0.003	+3.99 [1.39, 6.75]	0.007
SHViT-S3	+3.82 [1.82, 6.08]	0.001	+3.88 [1.89, 6.17]	0.001	+4.83 [2.29, 7.70]	0.001
DeiT III-Base	+3.82 [1.39, 6.42]	0.004	+4.04 [1.63, 6.70]	0.004	+4.74 [1.73, 7.90]	0.003
ViT-Base/16	+5.56 [2.86, 8.51]	0.001	+5.81 [3.07, 8.86]	0.001	+6.98 [3.54, 10.64]	0.001

Panel B: Exact McNemar Tests by Seed (MCG-Net-Only Correct/Comparator-Only Correct; Unadjusted p).

Comparator	Run 0	Run 1	Run 2
DenseNet-169	17/8 (0.108)	12/6 (0.238)	14/14 (1.000)
RDNet-Base	15/6 (0.078)	12/9 (0.664)	10/6 (0.454)
EfficientNetV2-M	16/9 (0.230)	14/11 (0.690)	18/11 (0.265)
StarNet-S3	16/7 (0.093)	17/7 (0.064)	13/9 (0.523)
ConvNeXt-Base	16/9 (0.230)	16/9 (0.230)	21/9 (0.043)
PiT-B	24/8 (0.007)	16/10 (0.327)	16/8 (0.152)
InceptionNeXt-Base	26/7 (0.001)	13/3 (0.021)	14/11 (0.690)
Swin-Base	21/9 (0.043)	19/7 (0.029)	20/9 (0.061)
RegNetY-16GF	28/7 (<0.001)	11/8 (0.648)	26/12 (0.034)
SHViT-S3	27/5 (<0.001)	18/12 (0.362)	25/9 (0.009)
DeiT III-Base	26/11 (0.020)	22/5 (0.002)	24/12 (0.065)
ViT-Base/16	27/6 (<0.001)	30/12 (0.008)	36/11 (<0.001)

Panel A: 10,000 paired replicates; resampling seed 20261004; adjusted significance threshold 0.02. The intervals are not simultaneous adjusted intervals. Accuracy, macro-F1, and MCC each form a separate exploratory family. Panel B: Holm adjustment across all 36 tests at 0.02 leaves SHViT-S3 run 0 ($p_{\text{Holm}} = 0.004$), ViT-Base/16 runs 0 and 2 (0.011 and 0.012), and RegNetY-16GF run 0 (0.017) significant. Counts refer to paired correctness on the same 384 subjects per fit.

4.2. Class-Specific Performance and Error Structure

CNV sensitivity was highest at 96.5%, with precision of 87.2% (Table 9). DME precision and sensitivity were 96.0% and 92.0%; NORMAL sensitivity was 95.1%. DRUSEN had the lowest sensitivity, 84.4% (98% interval, 76.6–91.4), despite 95.1% precision. This asymmetry indicates that true drusen cases were more often missed than non-drusen cases were assigned to DRUSEN. Class-specific AUROC ranged from 97.38% for DRUSEN to 99.46% for NORMAL.

Table 9. Class-specific one-versus-rest endpoints for MCG-Net, averaged over three fitted runs. Percentages in parentheses are nominal two-sided 98% subject-index bootstrap intervals; AUROC is the point estimate.

Class	Precision	Sensitivity	Specificity	F1-Score	AUROC
CNV	87.2 (79.8–93.6)	96.5 (92.6–99.4)	95.3 (92.4–97.7)	91.6 (87.1–95.4)	98.55
DME	96.0 (91.6–99.2)	92.0 (85.6–97.0)	98.7 (97.3–99.8)	94.0 (90.0–97.1)	98.88
DRUSEN	95.1 (90.8–98.4)	84.4 (76.6–91.4)	98.5 (97.2–99.5)	89.3 (84.2–93.6)	97.38
NORMAL	91.1 (84.8–96.3)	95.1 (91.0–98.5)	96.9 (94.7–98.7)	93.0 (89.1–96.3)	99.46

The aggregate confusion matrix summarizes 1152 predictions from the same 384 observations across three fits (Figure 3); these are not 1152 independent subjects. Of 288 repeated DRUSEN predictions, 25 were assigned to CNV and 20 to NORMAL. DME-to-CNV errors accounted for 15 of 288 DME predictions, compared with seven CNV-to-DME errors. The row-normalized matrix places these error types beside class sensitivity and supports inspection of the selected visual examples.

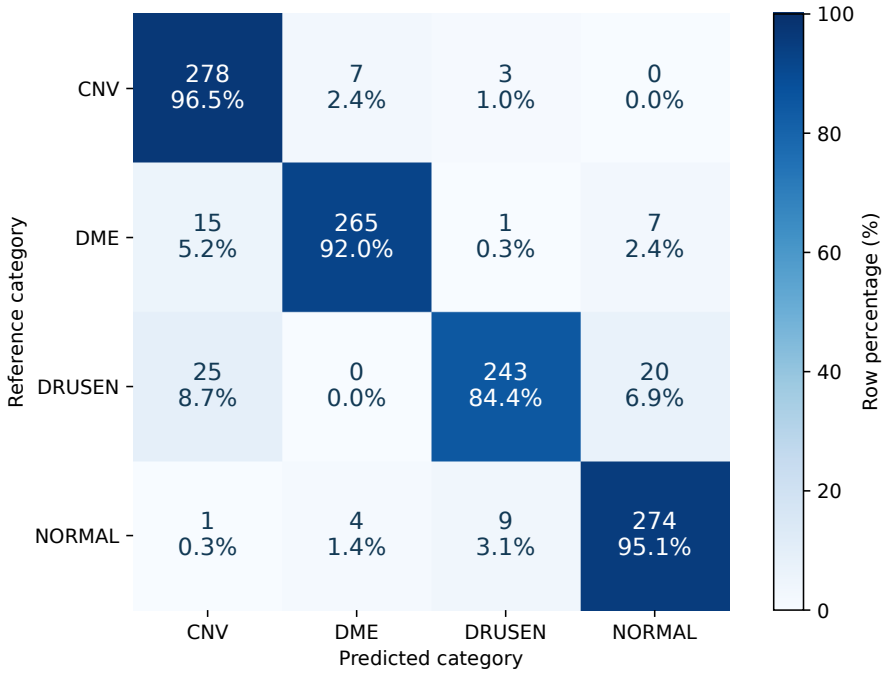


Figure 3. MCG-Net confusion matrix summed over three fitted runs. Each reference row represents 96 unique B-scans evaluated three times (288 predictions). Cells show repeated prediction counts and row percentages. The independent external sampling unit remains one of 384 represented subjects, not one of 1152 predictions.

4.3. Source-Specific Assessment

MCG-Net accuracy was 90.86 ± 0.20% on CAVRI-H5 ($n = 288$) and 95.49 ± 2.17% on OCTDL ($n = 96$). Macro-F1 was 90.69% across the four represented CAVRI-H5 classes and 96.81% across the DME/NORMAL reference support in OCTDL (Table 10). All four predictions remained possible in both sources. The two source-level summaries are not like-for-like four-class comparisons because OCTDL supplies no retained CNV or DRUSEN references.

Within the shared categories, MCG-Net DME sensitivity was 91.7% in CAVRI-H5 and 92.4% in OCTDL; NORMAL sensitivity was 91.7% and 98.6%, respectively. These estimates describe the observed source strata but do not isolate scanner, population, or acquisition effects. EfficientNetV2-M had the highest OCTDL accuracy among the evaluated models (96.53%), whereas MCG-Net retained the highest aggregate and CAVRI-H5 accuracy. Source-specific outcomes therefore add information beyond the pooled model ranking.

Table 10. Source-specific external endpoints (percent), mean ± SD across three fits. C/O denotes CAVRI-H5/OCTDL; Se, sensitivity.

Model	CAVRI Acc	CAVRI F1	OCTDL Acc	OCTDL F1	DME Se C/O	NORMAL Se C/O
MCG-Net	90.86 ± 0.20	90.69 ± 0.26	95.49 ± 2.17	96.81 ± 1.85	91.7/92.4	91.7/98.6
DenseNet-169	89.00 ± 0.80	89.42 ± 0.64	95.83 ± 1.04	97.14 ± 0.83	90.3/91.7	97.2/100.0
RDNet-Base	88.89 ± 0.35	89.38 ± 0.26	95.83 ± 0.00	96.81 ± 0.00	93.8/91.7	91.7/100.0
EfficientNetV2-M	88.54 ± 0.60	89.10 ± 0.91	96.53 ± 0.60	97.03 ± 1.08	95.1/98.6	94.4/94.4
PiT-B	87.85 ± 0.35	88.69 ± 0.69	94.10 ± 2.17	95.53 ± 1.89	93.1/88.9	95.8/99.3
StarNet-S3	88.19 ± 0.92	88.60 ± 0.65	95.49 ± 0.60	96.98 ± 0.61	91.0/93.1	93.1/97.9
ConvNeXt-Base	88.08 ± 1.75	88.24 ± 1.58	94.79 ± 1.04	96.60 ± 0.82	92.4/91.0	90.3/98.6
InceptionNeXt-Base	87.96 ± 1.12	87.94 ± 1.00	93.06 ± 3.18	95.38 ± 1.83	93.1/86.1	91.7/100.0
RegNetY-16GF	86.23 ± 2.36	86.99 ± 2.38	96.18 ± 3.18	97.30 ± 2.51	88.2/92.4	95.8/100.0
Swin-Base	86.81 ± 0.60	86.96 ± 0.92	95.49 ± 0.60	97.14 ± 0.79	90.3/91.7	93.8/99.3
DeiT III-Base	86.11 ± 0.92	86.59 ± 0.79	94.44 ± 0.60	95.41 ± 0.28	91.7/91.0	93.8/97.9
SHViT-S3	85.88 ± 1.45	86.29 ± 1.44	95.14 ± 3.18	96.92 ± 2.23	90.3/91.0	89.6/99.3
ViT-Base/16	83.56 ± 0.87	83.51 ± 0.91	95.14 ± 2.62	96.62 ± 1.66	85.4/93.1	89.6/97.2

CAVRI-H5: $n = 288$, four reference classes. OCTDL: $n = 96$, DME/NORMAL reference classes. OCTDL macro-F1 averages only the two represented reference labels, with all four predictions allowed. Aggregate source scores have different class support and are not a matched four-class source contrast.

4.4. Contribution of the Residual Context Branches

The native A1 backbone achieved 89.93 ± 0.60% accuracy and 89.81 ± 0.61% macro-F1. Adding context 11 (A2) increased these means to 90.45% and 90.33%; context 23 alone (A3) reached 90.80% and 90.71%. Full reached 92.01% accuracy and 91.99% macro-F1 (Table 11, Figure 4). Relative to A1, the two context branches add approximately 2.58 million parameters, or 10.6%. All four variants were evaluated using distinct seeds 0, 1, and 2.

Table 11. Architectural ablation on the external cohort. Mean ± SD over three fitted runs (percent), using the same four-class prediction rule. A1: Native DW7; A2: DW7 + context 11; A3: DW7 + context 23; Full: DW7 + both context branches.

Variant	Params (M)	Accuracy	Macro-P	Macro-F1	MCC	AUROC	AUPRC
A1 (local baseline)	∼24.25	89.93 ± 0.60	90.81 ± 0.68	89.81 ± 0.61	86.91 ± 0.78	98.29 ± 0.34	96.14 ± 0.71
A2	∼25.52	90.45 ± 0.15	91.30 ± 0.41	90.33 ± 0.17	87.61 ± 0.23	97.85 ± 0.22	95.21 ± 0.99
A3	∼25.56	90.80 ± 0.54	91.24 ± 0.62	90.71 ± 0.58	87.92 ± 0.73	98.56 ± 0.05	96.85 ± 0.31
Full	26.82	92.01 ± 0.60	92.33 ± 0.63	91.99 ± 0.61	89.47 ± 0.80	98.57 ± 0.06	96.62 ± 0.24

Parameter counts describe the four-output specification. Each variant uses seeds 0, 1, and 2; gating, transitions, and pooling remain shared.

The paired Full-minus-A1 macro-F1 difference was 2.18 points (98% interval, 0.50–4.04; adjusted $p = 0.019$). The accuracy difference was 2.08 points (0.43–3.91; $p = 0.042$), which did not meet $\alpha = 0.02$. Full-minus-A2 and Full-minus-A3 differences likewise did not meet the adjusted threshold (Table 12). The macro-F1 interaction estimate was 0.75 points (−2.00 to 3.58); therefore, the experiment does not establish a supra-additive interaction between branches.

Table 12. Paired bootstrap differences, Full minus architectural alternative, in percentage points with nominal 98% intervals and Holm-adjusted p within each five-comparison metric family.

Variant	ΔAccuracy	ΔMacro-F1	ΔMCC
A1	+2.08 [0.43, 3.91], $p = 0.042$	+2.18 [0.50, 4.04], $p = 0.019$	+2.56 [0.39, 4.95], $p = 0.045$
A2	+1.56 [−0.17, 3.47], $p = 0.200$	+1.66 [−0.12, 3.57], $p = 0.151$	+1.86 [−0.36, 4.23], $p = 0.237$
A3	+1.22 [−0.43, 2.95], $p = 0.251$	+1.27 [−0.38, 3.07], $p = 0.197$	+1.55 [−0.57, 3.82], $p = 0.255$
K7	+1.04 [−0.26, 2.43], $p = 0.251$	+1.06 [−0.23, 2.46], $p = 0.197$	+1.32 [−0.38, 3.17], $p = 0.255$
K15	+1.30 [−0.43, 3.13], $p = 0.251$	+1.37 [−0.32, 3.24], $p = 0.197$	+1.56 [−0.63, 3.93], $p = 0.255$

Only Full-minus-A1 macro-F1 meets the adjusted 0.02 threshold. Context interaction: Accuracy +0.69 [−2.00, 3.47]; macro-F1 +0.75 [−2.00, 3.58]. The interaction intervals include zero. The separate ablation McNemar family has no significant adjusted result.

Class-specific changes were largest for drusen sensitivity, from 75.3 ± 2.2% in A1 to 84.4 ± 2.1% in Full (a descriptive increase of 9.1 points). CNV F1 increased from 87.0% to 91.6%. NORMAL F1 decreased from 94.3% to 93.0%, indicating that the aggregate improvement was not uniform across categories.

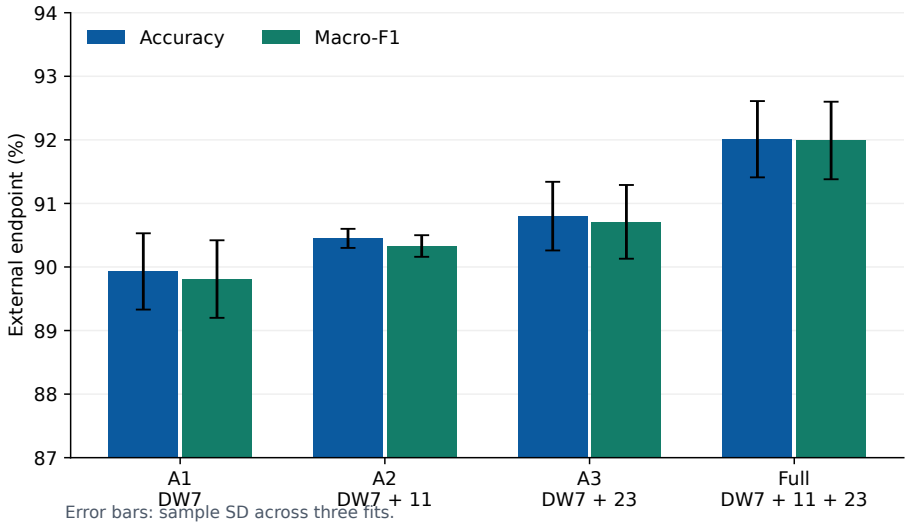


Figure 4. Accuracy and macro-F1 for native and context-augmented variants. Error bars show the sample SD across three fitted runs, not confidence intervals. Full-minus-A1 macro-F1 meets the paired adjusted threshold; numerical ordering of the other variants does not establish significant differences.

4.5. Second-Kernel Comparison

With context 11 fixed, second kernels 7, 15, and 23 produced mean macro-F1 scores of 90.93%, 90.62%, and 91.99%, respectively (Table 13, Figure 5). Kernel 23 had the highest observed external mean, and the Full 11/23 combination was selected using these outcomes together with the architectural ablation. Differences from kernels 7 and 15 remained uncertain after paired adjustment, and the external trend was not monotonic. This numerical selection does not establish an optimal kernel size or a significant advantage over every tested configuration. Confirmation of the selected combination requires a further cohort untouched by architectural selection.

Table 13. Second-kernel comparison with the first context branch fixed at 11. External endpoints describe CNV, DME, DRUSEN, and NORMAL. Means ± SD are percentages across three fitted runs.

Second Kernel	Params (M)	Accuracy	Macro-F1	MCC	AUROC
7	26.78	90.97 ± 0.54	90.93 ± 0.59	88.15 ± 0.65	98.42 ± 0.25
15	26.80	90.71 ± 0.84	90.62 ± 0.83	87.91 ± 1.08	98.06 ± 0.38
23 (Full)	26.82	92.01 ± 0.60	91.99 ± 0.61	89.47 ± 0.80	98.57 ± 0.06

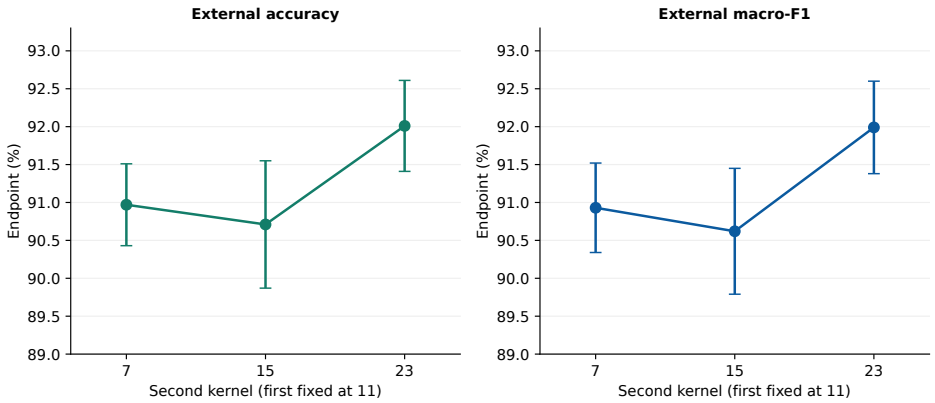


Figure 5. Second-kernel comparison with context 11 fixed. Accuracy and macro-F1 describe the same four-class external assessment. Error bars show across-run SD; the largest kernel has the highest observed external mean without a significant adjusted advantage over the alternatives.

4.6. Computational Profile

The instantiated four-output MCG-Net has approximately 26.82 million parameters, 9.97 GFLOPs under the declared THOP counting convention, 4.28 ms median batch-1 latency, and 1396 images/s batch-16 throughput on an RTX 5090 (Table 14). DenseNet-169 uses fewer parameters and a lower counted operation total, while MCG-Net has lower batch-1 latency. StarNet-S3 and SHViT-S3 are substantially smaller and provide higher throughput. ConvNeXt-Base has more parameters but lower batch-1 latency. The model consequently occupies a performance and resource trade-off rather than achieving minimum cost or maximum speed on every measure (Figure 6).

Table 14. Matched inference profile at 224 × 224 input on an RTX 5090. Macro-F1 is the external mean (percent); parameter counts use the instantiated four-output models. Latency is median batch-1 inference; throughput uses batch 16.

Model	Macro-F1	Params (M)	Counted GFLOPs	Latency (ms)	Throughput (img/s)
MCG-Net	91.99	26.82	9.97	4.28	1396
DenseNet-169	90.62	12.49	6.72	5.24	1521
RDNet-Base	90.55	85.70	30.71	16.28	597
EfficientNetV2-M	90.42	52.86	10.44	28.23	752
StarNet-S3	89.98	5.49	1.52	1.35	5365
ConvNeXt-Base	89.69	87.57	30.71	3.37	1136
PiT-B	89.29	72.74	21.11	2.51	1505
InceptionNeXt-Base	89.18	83.61	29.72	4.91	828
Swin-Base	88.82	86.75	30.34	4.79	1055
RegNetY-16GF	88.54	80.58	31.82	7.14	785
SHViT-S3	88.11	13.80	1.22	2.66	5975
DeiT III-Base	87.95	85.82	33.70	2.98	1162
ViT-Base/16	86.18	85.80	33.70	2.93	1181

FP32, AMP/TF32 disabled, no gradients. Counted GFLOPs equal two times the THOP MACs and exclude functional attention matrix products; they are not complete cross-family operation totals. Hardware timing concerns model execution, not OCT acquisition or clinical workflow.

The instantiated four-output MCG-Net contains 26,824,900 parameters. Parameter count, the partial operation-count proxy, and measured timing are reported separately because their rankings differ and their attention operators are incompletely covered by the profiler.

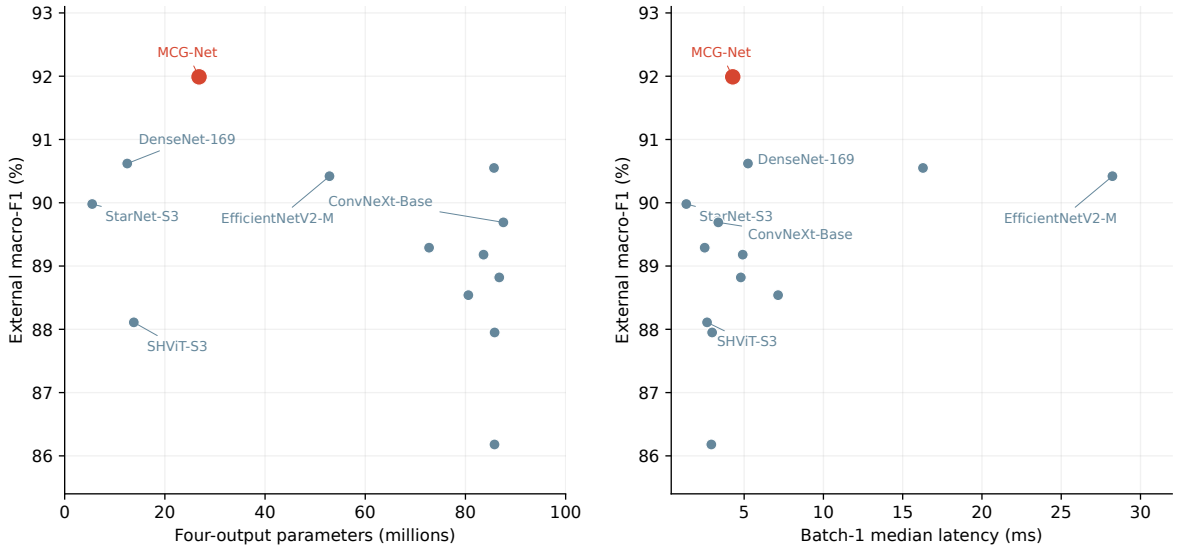


Figure 6. External macro-F1 versus parameter count (**left**) and measured batch-1 latency (**right**). Points summarize the evaluated models, and the labeled models illustrate distinct resource and performance choices. Timing uses the matched RTX 5090 FP32 protocol; parameter count alone does not determine latency. MCG-Net is shown in red; all comparator models are shown in blue.

4.7. Hypothetical Prevalence and Predictive Value

At the artificially balanced 25% class prevalence, the count-derived one-versus-rest PPV is 87.1% for CNV and 96.0% for DME. Under a hypothetical 1% prevalence, these values decrease to 17.0% and 42.2%, respectively (Figure 7). At 1% prevalence, DRUSEN and NORMAL PPV are 36.2% and 23.5%, compared with 94.9% and 91.0% at the balanced 25% prevalence. These scenarios illustrate the consequence of nonzero false-positive rates; they are neither empirical clinical PPV estimates nor a validation of probability calibration.

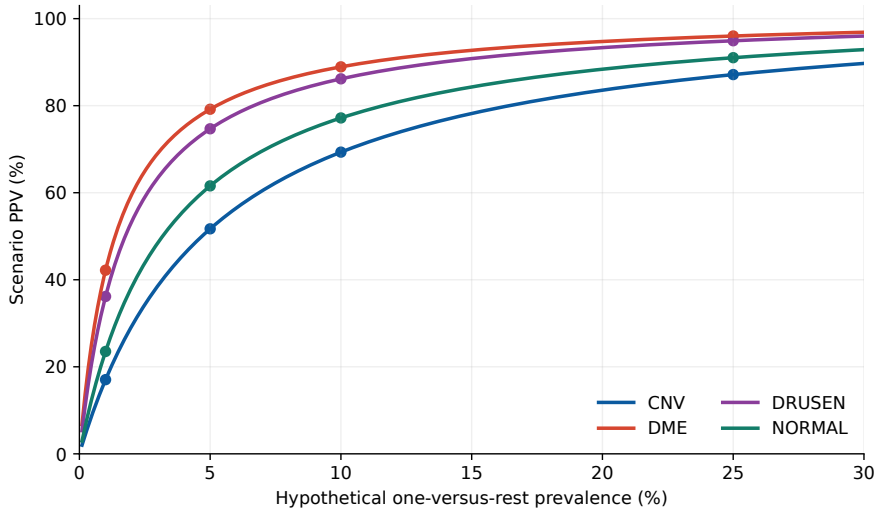


Figure 7. Hypothetical one-versus-rest positive predictive value as prevalence changes. Curves use unrounded aggregate confusion counts and assume stable sensitivity and specificity. Marked points correspond to 1%, 5%, 10%, and 25% prevalence. These are scenario calculations, not observed clinical prevalences or calibrated model probabilities.

4.8. Selected Predictions and Grad-CAM++ Localizations

Figure 8 presents four correctly classified B-scans, one per category, and four errors with four-class probabilities. Examples were selected from the fits trained with seeds 0, 1, and 2. The illustrated errors include DRUSEN assigned to CNV or NORMAL and DME assigned to CNV. All four errors have a predicted-class probability of at least 92.6%. High softmax confidence therefore does not exclude error, and this figure cannot be interpreted as evidence of calibrated uncertainty.

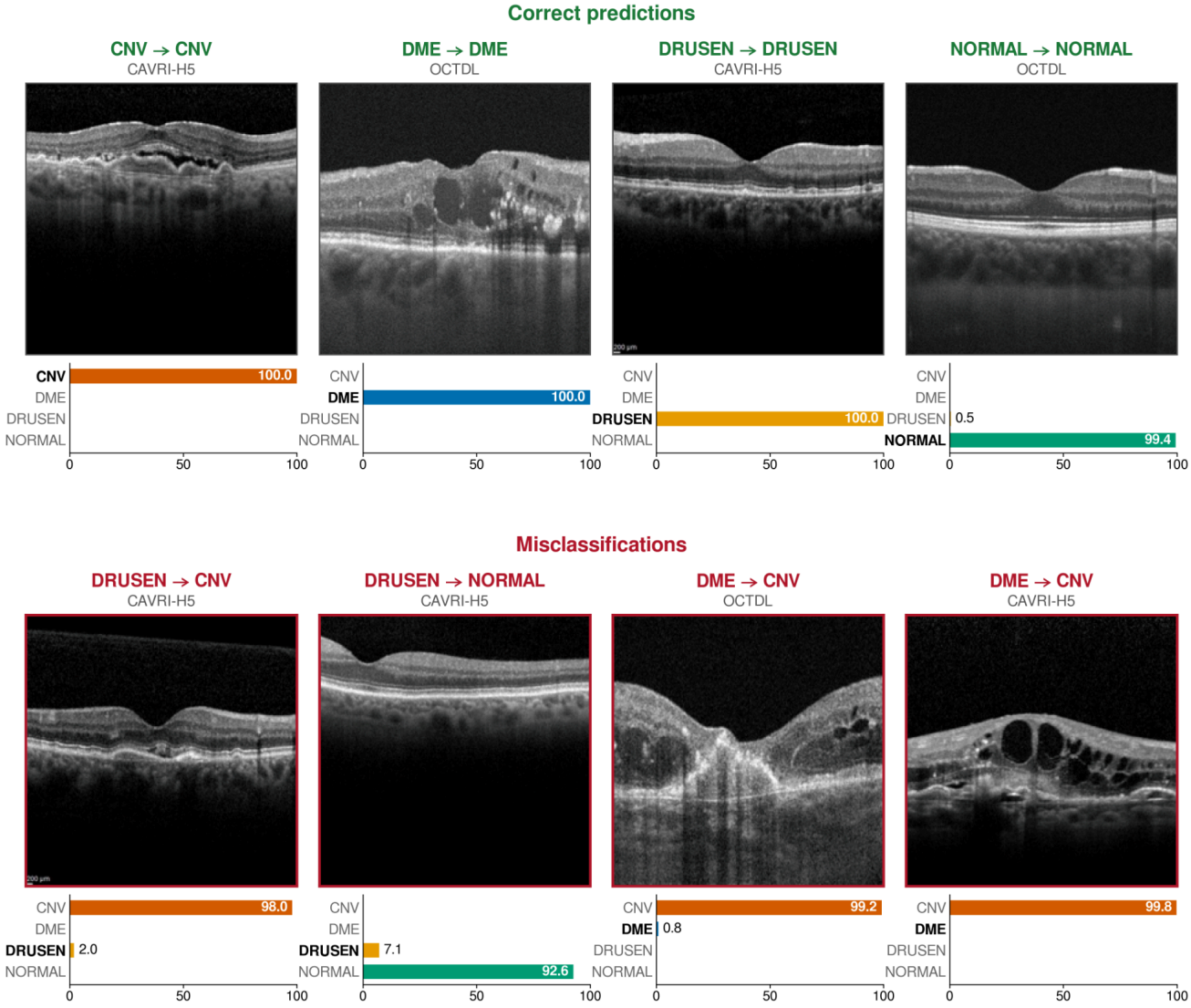


Figure 8. Selected external MCG-Net predictions drawn from the fits trained with seeds 0, 1, and 2. **Top row**: One correct prediction per category. **Bottom row**: Four misclassifications. Labels show reference and predicted category, source, and four-category probability bars. Each probability vector belongs to the fitted model used for that example; vectors are not averaged across seeds and are not calibrated diagnostic risk estimates.

Figure 9 shows different selected B-scans and Grad-CAM++ overlays for the predicted class. In correct cases, highlighted regions overlap the RPE elevation and fluid-related CNV appearance, intraretinal spaces in DME, focal RPE elevations in DRUSEN, and organized parafoveal retina in NORMAL. In the DRUSEN-to-CNV error, localization includes irregular RPE elevations; in the DRUSEN-to-NORMAL error, it is more diffuse around subtle deposits. The DME-to-CNV and CNV-to-DME examples show emphasis near the outer

retina and RPE–choroid band. These appearances are compatible with morphological overlap between categories, but they do not establish which clinical finding caused a prediction. No lesion-mask agreement or attribution-faithfulness endpoint was measured.

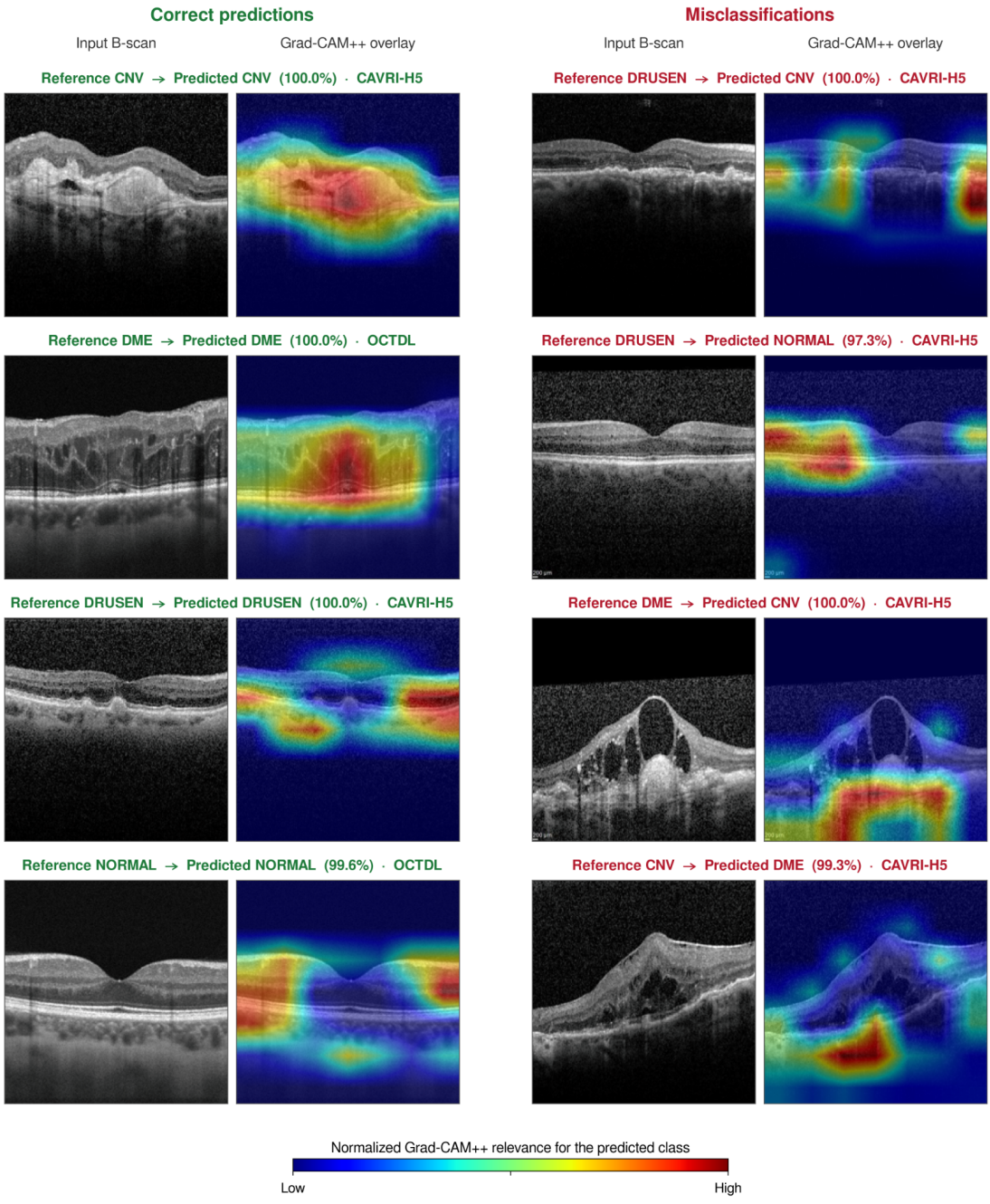


Figure 9. Grad-CAM++ localizations for external B-scans selected from MCG-Net fits trained with seeds 0, 1, and 2. Each overlay is generated from the corresponding fitted model at target layer `stages.3.blocks.1`. **Left**: Correct predictions, one per category. **Right**: Misclassifications selected across the three fits. Each example pairs an input with its predicted-class overlay and gives its reference-to-prediction label, displayed probability, and source. Maps are normalized per image and are post hoc localizations rather than lesion segmentations. The cases differ from Figure 8; CAM-path resizing can change displayed probabilities relative to benchmark inference.

5. Discussion

The external comparison suggests that compressed multiscale residual context can improve the observed OCT classification mean under the assessed protocol. Be-

cause these outcomes also informed architecture selection, the magnitude of improvement may be optimistic for a future independent cohort. MCG-Net attains the leading mean accuracy and macro-F1 in the principal comparison, and the Full-minus-native macro-F1 difference meets the adjusted threshold. Differences against DenseNet-169, RDNet-Base, and EfficientNetV2-M remain compatible with sampling uncertainty, while EfficientNetV2-M has stronger probability-based discrimination. The Full-minus-native macro-F1 difference is the clearest exploratory architectural result, but its interval and adjusted test do not correct for external-data-informed configuration selection.

5.1. Spatial Context and the Limits of the Architectural Evidence

The context branches supplement an already effective local path. Compression controls their parameter growth, separable kernels expand spatial support, and bounded residual scales regulate how the new responses enter the native GELU-gated hierarchy. The contribution lies in bounded additive context within the retained gated hierarchy, distinct from multiplicative attention-map construction and separate CNN–Transformer streams [25,43,46,47].

The ablation supports the combined context extension most clearly for macro-F1. Single-branch variants improve the observed mean, but their adjusted differences from Full are uncertain. The interaction interval includes zero, so synergy is not established. Similarly, kernel 23 has the highest external mean without a significant advantage over the tested smaller kernels. These outcomes informed the selection of Full 11/23 and require the cautious interpretation of that selected configuration. A further external cohort excluded from architecture selection is needed to confirm the improvement and assess whether the apparent ordering persists.

5.2. Class-Specific Errors and Clinical Interpretation

Drusen are the main residual weakness. Their shallow RPE elevations can resemble normal contours or elements of neovascular disease, particularly after spatial resizing. The shift in drusen sensitivity from A1 to Full suggests that wider context may help distinguish these appearances, but the remaining 84.4% sensitivity leaves a substantial miss rate. CNV has high sensitivity with lower precision, consistent with drusen and edema images being assigned to it. Intraretinal and subretinal fluid, RPE changes, and surrounding-layer continuity are relevant anatomical distinctions, yet repository labels cannot determine whether a network uses these distinctions appropriately in each image [4,7,9].

The attribution examples make this limitation visible. Highlighting near an RPE elevation may be anatomically plausible in a correct CNV case and in a DRUSEN-to-CNV error. A plausible map therefore does not validate a diagnosis. The selected errors also show that probabilities close to one can accompany incorrect classifications. Grad-CAM++ provides localization of score-associated features; it supplies neither a segmentation of disease tissue nor causal evidence of biomarker reliance [50,51]. Quantitative lesion agreement, perturbation tests, probability calibration, and a prospectively selected referral threshold remain necessary for a clinically interpretable system.

5.3. External Assessment and Prevalence

Using external repositories separates model fitting from the assessment images, and within-source subject identifiers support paired subject-index uncertainty estimates. However, CNV and DRUSEN references appear only in CAVRI-H5. The better OCTDL aggregate score cannot be interpreted as uniformly better transfer because that subset contains only DME and NORMAL references. The shared-category sensitivities offer a more appropriate source contrast but do not separate patient mix from acquisition effects.

Finally, equal category counts facilitate comparison but differ from most clinical populations. The PPV scenarios demonstrate how the same sensitivity and specificity can yield many false-positive referrals at low prevalence. They assume a fixed operating behavior and cannot model all changes in disease spectrum, acquisition, and diagnostic practice. The benchmark consequently supports comparison of retrospective classifiers, not a population-screening claim or a deployment-ready triage pathway.

5.4. Relationship to Recent OCT Studies and Resource Requirements

Published OCT systems report high within-dataset results and increasingly examine external transfer, calibration, and uncertainty [26,36,41,48,49]. Table 1 places selected original results beside their task and evaluation scope. Differences in category definitions, pretraining sources, subject grouping, and external selection make raw percentages unsuitable for cross-study ranking. In particular, binary AMD-versus-normal self-supervised evaluation, four-category classification, and eight-category OCT-C8 assessment address different endpoints. Here, CAVRI-H5 is an external resource released in 2026 that is excluded from parameter fitting but contributes to configuration selection through the reported external comparisons [29]. Published results on other OCT cohorts are therefore used for methodological context; the repeated comparator experiment establishes performance on the same external observations.

MCG-Net's contribution is assessed most directly through its matched native-backbone ablation and shared external comparator experiment. It uses fewer parameters than several large Transformer and modern convolutional alternatives, but compact StarNet and SHViT models are faster and smaller. DenseNet-169 is also smaller while remaining statistically close in external performance. Measured latency is therefore necessary alongside parameter count, and the incomplete THOP attention coverage prevents an unqualified cross-family FLOP-efficiency claim. These measurements identify an operating trade-off on the specified GPU; they do not establish throughput on a clinical scanner, mobile device, or CPU workflow.

5.5. Limitations and Future Work

OCT-C8 was used only for training and validation, but its missing patient and volume identifiers prevent verification of patient-disjoint development. Exact-match and perceptual-similarity exclusions address image overlap without resolving unidentified within-patient dependence. CAVRI-H5 and OCTDL are separate external data sources with one selected observation per subject. Their outcomes informed the selection of Full 11/23 before the comparator analysis on the same cohort, introducing a risk of selection-related optimism. No additional untouched external cohort was used, and the reported bootstrap intervals and multiplicity adjustments do not correct for configuration selection. Accordingly, the numerical ranking and paired differences require confirmation on new external subjects. Cohort selection preceded predictions and used random subject sampling within fixed class and source quotas; however, the CAVRI-H5 sampling seed was not retained. Its manifest preserves the selected membership, while class balancing and incomplete random-draw provenance limit representativeness and exact regeneration. The two ophthalmologist reviewers reviewed images jointly, so inter-reader reliability cannot be estimated from independent ratings. Source and class remain associated, and conditions outside the four-category task have no validated rejection mechanism.

MCG-Net classifies individual B-scans and does not integrate information from adjacent slices within an OCT volume. This may limit detection of small or eccentric abnormalities that are poorly represented in a selected scan, so volume-level aggregation requires separate evaluation before the model can support comprehensive clinical assessment.

The three-seed runs characterize a limited range of training variability. Exact reproduction of the training environment is limited by the absence of documented training hardware and numerical precision. The common AdamW protocol supports comparison in a controlled optimization setting, but it does not establish the best achievable configuration for every architecture. The present analysis also does not establish calibrated uncertainty, prospective scanner robustness, attribution faithfulness, or reader benefit. The external sample size and category balance limit prevalence-sensitive conclusions. Further validation should use identifiable patient/volume-disjoint development, independent cross-center reference assessment, prospectively defined operating thresholds, and representative clinical sampling.

Future work should use prospectively assembled cohorts with complete patient and volume identifiers, independent blinded readers, and representative disease prevalence. Validation-only model-specific optimizer and schedule selection should be followed by repeated evaluation on untouched external cohorts, alongside controlled retraining of recent OCT-specific methods. Volume-level aggregation, calibrated confidence and abstention, lesion-level attribution validation, and acquisition-stratified assessment should be evaluated before clinical deployment.

6. Conclusions

MCG-Net augments a gated CNN with compressed residual context at kernels 11 and 23. On the 384-image external cohort comprising CNV, DME, DRUSEN, and NORMAL, the selected configuration achieved 92.01% mean accuracy and 91.99% macro-F1 across three fitted runs. The exploratory Full-minus-native macro-F1 difference was 2.18 points, while advantages over the strongest comparator models remained uncertain. This cohort also informed configuration selection, so these results do not constitute untouched validation of the final architecture. Source-specific results, prevalence scenarios, and visual errors delimit this finding. The evidence supports further investigation of multiscale residual context for OCT representation; it does not yet establish unrestricted, calibrated clinical diagnosis or prospective deployment.

Author Contributions: Conceptualization, D.O., H.C.K., S.M.G. and I.P.; methodology, S.M.G., M.B., M.E., F.O. and I.P.; software, M.B. and I.P.; validation, D.O., H.C.K., S.M.G., M.E. and F.O.; formal analysis, M.B., F.O. and I.P.; investigation, D.O., H.C.K., S.M.G. and I.P.; data curation, D.O., H.C.K., M.E. and I.P.; writing—original draft preparation, D.O. and I.P.; writing—review and editing, H.C.K., S.M.G., M.B., M.E., F.O. and I.P.; visualization, M.B. and I.P.; supervision, S.M.G., F.O. and I.P.; project administration, S.M.G. and I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No: KFU266184].

Institutional Review Board Statement: This retrospective secondary analysis used publicly available de-identified OCT images. No new participants were recruited, contacted, or examined, and no new clinical data were collected. The original dataset publications and repository records describe their source-specific ethics procedures. No new institutional approval number or formal waiver is claimed in this manuscript.

Informed Consent Statement: Not applicable to this secondary analysis of public de-identified images. Consent arrangements for the original collections are governed by the source-dataset records. No new participant consent was obtained for the present analysis.

Data Availability Statement: The development resource is Retinal OCT-C8 [6]; external resources are CAVRI-H5 [29] and OCTDL [28]. Access and reuse are subject to the original providers' terms. Aggregate results and cohort accounting are included in the main text and tables. Curated split

manifests, image-selection and exclusion records, per-image predictions, and experiment metadata are available upon reasonable request with permission from the corresponding author, subject to the original data providers' terms. The training and evaluation implementation, configuration files, analysis scripts, and analysis records, including the tabulated analytical inputs, plotting code, and prevalence-scenario calculation, are available upon reasonable request with permission from the corresponding author.

Acknowledgments: The authors thank the clinicians who provided the illustrative OCT annotations and the maintainers of the public OCT datasets. The authors gratefully acknowledge the support provided by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia. An LLM assisted with language editing, clarity, and organization of the manuscript. It was not used to design the model, train the models, acquire or generate primary OCT data, assign clinical reference labels, or generate the reported model predictions. The authors reviewed the text and remain responsible for verifying the scientific content, references, numerical results, and final declarations.

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

References

1. Huang, D; Swanson, E.A.; Lin, C.P.; Schuman, J.S.; Stinson, W.G.; Chang, W.; Hee, M.R.; Flotte, T.; Gregory, K.; Puliafito, C.A.; et al. Optical coherence tomography. *Science* **1991**, *254*, 1178–1181. https://doi.org/10.1126/science.1957169.
2. Jaffe, G.J.; Caprioli, J. Optical coherence tomography to detect and manage retinal disease and glaucoma. *Am. J. Ophthalmol.* **2004**, *137*, 156–169. https://doi.org/10.1016/S0002-9394(03)00792-X.
3. Fujimoto, J.; Swanson, E. The development, commercialization, and impact of optical coherence tomography. *Investig. Ophthalmol. Vis. Sci.* **2016**, *57*, OCT1–OCT13. https://doi.org/10.1167/iovs.16-19963.
4. Spaide, R.F.; Jaffe, G.J.; Sarraf, D.; Freund, K.B.; Sadda, S.R.; Staurenghi, G.; Waheed, N.K.; Chakravarthy, U.; Rosenfeld, P.J.; Holz, F.G.; et al. Consensus Nomenclature for Reporting Neovascular Age-Related Macular Degeneration Data: Consensus on Neovascular Age-Related Macular Degeneration Nomenclature Study Group. *Ophthalmology* **2020**, *127*, 616–636. https://doi.org/10.1016/j.ophtha.2019.11.004. Correction in *Ophthalmology* **2020**, *127*, 1434–1435. https://doi.org/10.1016/j.ophtha.2020.07.019.
5. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.
6. Naren, O.S. Retinal OCT Image Classification—C8 [Dataset]; Kaggle: San Francisco, CA, USA, 2021. Available online: https://doi.org/10.34740/KAGGLE/DSV/2736749 (accessed on 22 August 2026).
7. Otani, T.; Kishi, S.; Maruyama, Y. Patterns of diabetic macular edema with optical coherence tomography. *Am. J. Ophthalmol.* **1999**, *127*, 688–693. https://doi.org/10.1016/S0002-9394(99)00033-1.
8. Acan, D.; Karahan, E.; Kocak, N.; Kaynak, S. Evaluation of systemic risk factors in different optical coherence tomographic patterns of diabetic macular edema. *Int. J. Ophthalmol.* **2018**, *11*, 1204–1209. https://doi.org/10.18240/ijo.2018.07.21.
9. Spaide, R.F.; Curcio, C.A. Drusen characterization with multimodal imaging. *Retina* **2010**, *30*, 1441–1454. https://doi.org/10.1097/IAE.0b013e3181ee5ce8.
10. Farsiu, S.; Chiu, S.J.; O'Connell, R.V.; Folgar, F.A.; Yuan, E.; Izatt, J.A.; Toth, C.A. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology* **2014**, *121*, 162–172. https://doi.org/10.1016/j.ophtha.2013.07.013.
11. Ferris, F.L., III; Wilkinson, C.P.; Bird, A.; Chakravarthy, U.; Chew, E.; Csaky, K.; Sadda, S.R.; Beckman Initiative for Macular Research Classification Committee. Clinical classification of age-related macular degeneration. *Ophthalmology* **2013**, *120*, 844–851. https://doi.org/10.1016/j.ophtha.2012.10.036.
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. https://doi.org/10.1038/nature14539.
13. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005.
14. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **2018**, *24*, 1342–1350. https://doi.org/10.1038/s41591-018-0107-6.
15. Ganie, S.M.; Pacal, I. Adaptive saliency cascade network for efficient multiclass classification of multi-source brain MRI data. *Chemom. Intell. Lab. Syst.* **2026**, *279*, 105910.

16. Pacal, I. Modern Vision Transformer Models for Accurate and Explainable Brain Tumor MRI Classification. *OBM Neurobiol.* **2026**, *10*, 352. https://doi.org/10.21926/obm.neurobiol.2603352.
17. Yurdakul, M.; Ersoy, M.; Ozger, F.; Pacal, I. Bezier-YOLO enables real-time brain tumor segmentation in multi-plane MRI through boundary-aware parametric contour regression. *Sci. Rep.* **2026**, *advance online publication*. https://doi.org/10.1038/s41598-026-71696-8.
18. Pacal, I. Real-time You Only Look Once-based colorectal polyp detection for cancer screening across independent colonoscopy datasets. *Eurasian J. Med. Oncol.* **2026**, 026220238. https://doi.org/10.36922/EJMO026220238.
19. Naral, S.; Cakmak, Y.; Pacal, I. Comparative evaluation of vision transformers and convolutional networks for breast ultrasound image classification. *Explor. Med.* **2026**, *7*, 1001382. https://doi.org/10.37349/emed.2026.1001382.
20. Karatas, F.; Aslan, O.; Burukanli, M.; Ganie, S.M.; Ersoy, M.; Ozger, F.; Pacal, I. SBRF-Det: Scale-Boundary Representative Fusion for Multi-Finding Detection in Panoramic Dental Radiographs. *Diagnostics* **2026**, *16*, 3259. https://doi.org/10.3390/diagnostics16193259.
21. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; pp. 10096–10106.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. https://arxiv.org/abs/2010.11929.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.
24. Shen, J.; Hu, Y.; Zhang, X.; Gong, Y.; Kawasaki, R.; Liu, J. Structure-Oriented Transformer for retinal diseases grading from OCT images. *Comput. Biol. Med.* **2023**, *152*, 106445. https://doi.org/10.1016/j.compbiomed.2022.106445.
25. Laouarem, A.; Kara-Mohamed, C.; Bourennane, E.-B.; Hamdi-Cherif, A. HTC-retina: A hybrid retinal diseases classification model using transformer-convolutional neural network from optical coherence tomography images. *Comput. Biol. Med.* **2024**, *178*, 108726. https://doi.org/10.1016/j.compbiomed.2024.108726.
26. Jannat, F.E.; Gholami, S.; Lim, J.I.; Leng, T.; Alam, M.N.; Tabkhi, H. Multi-OCT-SelfNet: Integrating self-supervised learning with multi-source data fusion for enhanced multi-class retinal disease classification. *Front. Syst. Biol.* **2026**, *6*, 1717398. https://doi.org/10.3389/fsysb.2026.1717398.
27. Yu, W.; Wang, X. MambaOut: Do We Really Need Mamba for Vision? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2025; pp. 4484–4496. https://doi.org/10.1109/CVPR52734.2025.00423.
28. Kulyabin, M.; Zhdanov, A.; Nikiforova, A.; Stepichev, A.; Kuznetsova, A.; Ronkin, M.; Borisov, V.; Bogachev, A.; Korotkich, S.; Constable, P.A.; Maier, A. OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods. *Sci. Data* **2024**, *11*, 365. https://doi.org/10.1038/s41597-024-03182-7.
29. Ignatowicz, A.; Marciniak, T.; Stankiewicz, A.; Stopa, M.; Marciniak, E. CAVRI-H5: A Homogeneous OCT Dataset for Retinal Classification Featuring VMA-VMT Pathologies. *Sci. Data* **2026**. https://doi.org/10.1038/s41597-026-08009-1.
30. Ajmal, M.M.; Mumtaz, R.; Mumtaz, S.; Temdee, P.; Asif, M.D.A.; Khan, M.A.; Khan, S. Deep learning based eye disease classification using Optical Coherence Tomography (OCT) images. *Exp. Eye Res.* **2026**, *268*, 111017. https://doi.org/10.1016/j.exer.2026.111017.
31. Abdi, A.S.; Abdulazeez, A.M. A comprehensive review of deep learning in OCT image segmentation and classification. *Med. Nov. Technol. Devices* **2025**, *28*, 100396. https://doi.org/10.1016/j.medntd.2025.100396.
32. Srinivasan, P.P.; Kim, L.A.; Mettu, P.S.; Cousins, S.W.; Comer, G.M.; Izatt, J.A.; Farsiu, S. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. Opt. Express* **2014**, *5*, 3568–3577. https://doi.org/10.1364/BOE.5.003568.
33. Karthik, K.; Mahadevappa, M. Deep learning with adaptive convolutions for classification of retinal diseases via optical coherence tomography. *Image Vis. Comput.* **2024**, *146*, 105044. https://doi.org/10.1016/j.imavis.2024.105044.
34. Pan, H.; Miao, J.; Yu, J.; Dong, J.; Zhang, M.; Wang, X.; Feng, J. A lightweight model for the retinal disease classification using optical coherence tomography. *Biomed. Signal Process. Control* **2025**, *101*, 107146. https://doi.org/10.1016/j.bspc.2024.107146.
35. Rahmani, A.; Ghatee, M. Fast inference in classification of optical coherence tomography (OCT) images for real-time retinal disease diagnosis. *Biomed. Signal Process. Control* **2025**, *103*, 107365. https://doi.org/10.1016/j.bspc.2024.107365.
36. Vijayan, M.; Veerapu, G.; Nisha, J.S. A spatial adaptive multi-scale ConvNeXt framework for robust, calibrated, and statistically validated multi-class retinal OCT classification. *Front. Artif. Intell.* **2026**, *9*, 1898978. https://doi.org/10.3389/frai.2026.1898978.
37. Leyba-Mesa, M.V.; Barkana, B.D. Interpretable deeply supervised networks for class-imbalanced OCT classification. *AI* **2026**, *7*, 164. https://doi.org/10.3390/ai7050164.
38. He, J.; Wang, J.; Han, Z.; Ma, J.; Wang, C.; Qi, M. An interpretable transformer network for the retinal disease classification using optical coherence tomography. *Sci. Rep.* **2023**, *13*, 3637. https://doi.org/10.1038/s41598-023-30853-z.

39. Akça, S.; Garip, Z.; Ekinci, E.; Atban, F. Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal OCT images using vision transformers. *Lasers Med. Sci.* **2024**, *39*, 140. https://doi.org/10.1007/s10103-024-04089-w.
40. Yang, M.; Du, J.; Lv, R. CRAT: Advanced transformer-based deep learning algorithms in OCT image classification. *Biomed. Signal Process. Control* **2025**, *104*, 107544. https://doi.org/10.1016/j.bspc.2025.107544.
41. Jannat, F.-E.; Gholami, S.; Alam, M.N.; Tabkhi, H. OCT-SelfNet: A self-supervised framework with multi-source datasets for generalized retinal disease detection. *Front. Big Data* **2025**, *8*, 1609124. https://doi.org/10.3389/fdata.2025.1609124.
42. Durmaz Engin, C.; Beşenk, U.; Özizmirliler, D.; Selver, M.A. Comparative Analysis of Automated vs. Expert-Designed Machine Learning Models in Age-Related Macular Degeneration Detection and Classification. *Turk. J. Ophthalmol.* **2025**, *55*, 120–126. https://doi.org/10.4274/tjo.galenos.2025.74780.
43. Yang, J.; Hsu, C.; Wang, J.; Wu, B.; Lu, Y.; Ding, Y.; Zhao, Z.; Tang, K.; Lu, F.; Ma, L. HyReti-Net: Hybrid retinal diseases classification and diagnosis network using optical coherence tomography. *Front. Med.* **2025**, *12*, 1660920. https://doi.org/10.3389/fmed.2025.1660920.
44. Liu, J.; Yuan, Y.; Xue, S.; Li, Z.; Li, X.; Jiao, J.; Ma, X.; Chang, K.; Cao, Z.; Fan, Y.; Liu, D. MergeNeXt: A hybrid CNN–transformer model for retinal OCT image classification. *Sci. Rep.* **2026**. https://doi.org/10.1038/s41598-026-70044-0.
45. Ko, J.; Han, J.; Yoon, J.; Choi, S.; Won, H.; Yang, M.; Park, J.; Kim, J.-E.; Hwang, J.S.; Han, J.M.; Yoo, T.K.; Park, J.I.; Hwang, D.D.-J. Retinal disease classification from spectral domain optical coherence tomography using lesion attentive fusion. *Sci. Rep.* **2026**. https://doi.org/10.1038/s41598-026-64770-8.
46. Lau, K.W.; Po, L.-M.; Rehman, Y.A.U. Large separable kernel attention: Rethinking the large kernel attention design in CNN. *Expert Syst. Appl.* **2024**, *236*, 121352. https://doi.org/10.1016/j.eswa.2023.121352.
47. Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; Hu, S.-M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. https://doi.org/10.1007/s41095-023-0364-2.
48. Fahim, T.A.; Alam, F.B.; Ahmed, M.S.; Jia, Y. PhysioSpeck-Net: Physics-guided feature separation for parameter-efficient and uncertainty-aware retinal OCT classification under cross-dataset shift. *Computers* **2026**, *15*, 630. https://doi.org/10.3390/computers15090630.
49. Kaza, V.; Chattu, P.; Kumar, C.K.; Pandraju, T.K.S.; Sirisha, U. DAVis-Net: A dual-attention deep supervision framework for reliable retinal OCT image classification with integrated explainability and uncertainty quantification. *Math. Comput. Appl.* **2026**, *31*, 190. https://doi.org/10.3390/mca31050190.
50. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. https://doi.org/10.1007/s11263-019-01228-7.
51. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2018; pp. 839–847. https://doi.org/10.1109/WACV.2018.00097.
52. Iida, T.; Hagimura, N.; Sato, T.; Kishi, S. Evaluation of central serous chorioretinopathy with optical coherence tomography. *Am. J. Ophthalmol.* **2000**, *129*, 16–20. https://doi.org/10.1016/S0002-9394(99)00272-X.
53. Duker, J.S.; Kaiser, P.K.; Binder, S.; de Smet, M.D.; Gaudric, A.; Reichel, E.; Sadda, S.R.; Sebag, J.; Spaide, R.F.; Stalmans, P. The International Vitreomacular Traction Study Group classification of vitreomacular adhesion, traction, and macular hole. *Ophthalmology* **2013**, *120*, 2611–2619. https://doi.org/10.1016/j.ophtha.2013.07.042.
54. Guyer, D.R.; Yannuzzi, L.A.; Slakter, J.S.; Sorenson, J.A.; Ho, A.; Orlock, D. Digital indocyanine green videoangiography of central serous chorioretinopathy. *Arch. Ophthalmol.* **1994**, *112*, 1057–1062. https://doi.org/10.1001/archopht.1994.01090200063023.
55. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415. https://doi.org/10.48550/arXiv.1606.08415.
56. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
57. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019. https://doi.org/10.48550/arXiv.1711.05101.
58. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
59. Kim, D.; Heo, B.; Han, D. DenseNets Reloaded: Paradigm Shift Beyond ResNets and ViTs. *arXiv* **2024**, arXiv:2403.19588. https://doi.org/10.48550/arXiv.2403.19588.
60. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the Stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 5694–5703.

61. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022; pp. 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167.
62. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking Spatial Dimensions of Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 11936–11945.
63. Yu, W.; Zhou, P.; Yan, S.; Wang, X. InceptionNeXt: When Inception Meets ConvNeXt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 5672–5683.
64. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing Network Design Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 10425–10433. https://doi.org/10.1109/CVPR42600.2020.01044.
65. Yun, S.; Ro, Y. SHViT: Single-Head Vision Transformer with Memory Efficient Macro Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 5756–5767.
66. Touvron, H.; Cord, M.; Jégou, H. DeiT III: Revenge of the ViT. In *Computer Vision—ECCV 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; pp. 516–533. https://doi.org/10.1007/978-3-031-20053-3_30.

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.