Article

Not peer-reviewed version

A Hierarchical Deep Learning Architecture for OCT Diagnostics of Retinal Diseases and Their Complications Capable of Performing Cross-Modal OCT to Fundus Translation in the Absence of Paired Data

Ekaterina A. Lopukhova *, Gulnaz M. Idrisova, Timur R. Mukhamadeev, Grigory S. Voronkov, Ruslan V. Kutluyarov, Elizaveta P. Topolskaya

Posted Date: 5 December 2025

doi: 10.20944/preprints202512.0450.v1

Keywords: optical coherence tomography; diabetic retinopathy; age-related macular degeneration; diabetic macular edema; deep learning; hierarchical neural networks; cross-modal learning; multilabel classification; contrastive learning; medical imaging; ophthalmology; computer-aided diagnosis; probability calibration; domain adaptation

Article

A Hierarchical Deep Learning Architecture for OCT Diagnostics of Retinal Diseases and Their Complications Capable of Performing Cross-Modal OCT to Fundus Translation in the Absence of Paired Data

Ekaterina A. Lopukhova [1,*,†], Gulnaz M. Idrisova [2], Timur R. Mukhamadeev [2], Grigory S. Voronkov [1], Ruslan V. Kutluyarov [1] and Elizaveta P. Topolskaya [1]

[1] Research Laboratory "Sensor Systems Based on Integrated Photonics Devices", Ufa University of Science and Technology, 32 Z. Validi Street, 450076 Ufa, Russia
[2] Department of Ophthalmology, Bashkir State Medical University, 3 Lenin Street, 450008 Ufa, Russia
* Correspondence: lopukhova.ea@ugatu.su

Abstract

The paper presents a solution to the limited accuracy of automated diagnostics for retinal pathologies, such as diabetic retinopathy and age-related macular degeneration. These challenges arise from difficulties in modeling comorbidities, a reliance on paired multimodal data, and issues related to class imbalance. The proposed solution features a novel hierarchical deep learning architecture designed for multi-label classification of optical coherence tomography (OCT) data. This architecture facilitates cross-modal knowledge transfer from fundus images without the need for paired fundus images. It was accomplished through the modular specialization of the architecture and the application of contrast equalization, which creates a latent "bridge" between the OCT and fundus data. The results demonstrate that the proposed approach achieves high accuracy (macro-F1 score of 0.989) and good calibration (Expected Calibration Error of 2.1%) in classification and staging tasks. Notably, it eliminates the need for fundus images for diabetic retinopathy staging in 96.1% of cases and surpasses traditional monolithic architectures on the macro-AUROC metric.

Keywords: optical coherence tomography; diabetic retinopathy; age-related macular degeneration; diabetic macular edema; deep learning; hierarchical neural networks; cross-modal learning; multi-label classification; contrastive learning; medical imaging; ophthalmology; computer-aided diagnosis; probability calibration; domain adaptation

1. Introduction

Diabetic retinopathy (DR), including diabetic macular edema (DME), and age-related macular degeneration (AMD) are the leading causes of blindness and irreversible vision loss worldwide [1–5]. DR currently affects over 100 million people globally, with DME developing in 6–7% of individuals with diabetes mellitus (DM) [6,7]. As of 2021, the global prevalence of DR among patients with DM is 22.27% [6].

In turn, AMD is the leading cause of central vision loss in individuals aged 50 and older in developed countries [8,9]. Additionally, clinical observations reveal a high incidence of co-occurring diseases; for instance, the presence of DR significantly increases the risk of developing AMD [10]. DME can occur at any stage of diabetic retinopathy, including the proliferative stage (PDR). In specialized tertiary-level clinical cohorts, the prevalence of DME can reach approximately 20–30% [11,12].

In modern ophthalmology, Optical Coherence Tomography (OCT) is a crucial method for the morphological assessment of macular structures, providing micrometer resolution and enabling visu-

 doi:10.20944/preprints202512.0450.v1

alization of intraretinal architecture [13–16]. Meanwhile, color fundus photography remains the "gold standard" for staging DR according to the international classifications, such as the Early Treatment Diabetic Retinopathy Study (ETDRS) and the International Clinical Classification of Diabetic Retinopathy (ICDR) [17–20]. However, this specialization presents a practical challenge: for a comprehensive diagnosis, both OCT and fundus images of the same eye are necessary. In real clinical practice, obtaining both can be difficult due to technical, logistical, and economic limitations [20,21].

Current methods for automatically diagnosing retinal diseases using various imaging modalities encounter several significant limitations. For instance, traditional classification techniques that rely on the softmax function impose an artificial mutual exclusivity among classes, which is insufficient for accurately representing the comorbidity of DR, including DME and its combinations with AMD [22,23]. An alternative approach is to use a multi-target or multi-label formulation with independent sigmoid activations, which allows for the accurate representation of multiple diseases occurring simultaneously. However, this approach necessitates specialized loss functions and calibration [24–27]. Moreover, most modern methods depend on the availability of strictly paired OCT and fundus images for both training and testing [28–30]. Existing cross-modal methods exhibit significant performance degradation when the corresponding images are unavailable [31–35].

One additional limitation is the class imbalance and domain shifts present in medical data, which result in the systematic underrepresentation of rare but clinically significant conditions [36–38]. At the same time, variations in scanning parameters across different equipment manufacturers, such as Optovue, Zeiss, and Heidelberg, lead to domain shifts that can degrade algorithm performance when switching between scanners [39–42].

Monolithic Convolutional Neural Network (CNN) architectures often show inadequate probability calibration [43,44]. It implies that, even when the overall accuracy, measured by the Brier Score, reaches clinically significant levels due to specific calibration efforts, the Expected Calibration Error—a metric that assesses how well probabilities align with actual outcomes—can still be high. In fact, it may fluctuate by as much as 4–6% or more [43–45].

Existing solutions struggle to address the complex and interconnected challenges of ophthalmological diagnostics. Traditional architectures that rely on single modalities, such as convolutional neural networks (e.g., ResNet, EfficientNet, ConvNeXt) and Vision Transformers (e.g., ViT, Swin Transformer) [46–49], are limited because they cannot utilize complementary information from other modalities. They also tend to exhibit unstable performance when faced with domain shifts [50–55]. Current cross-modal approaches typically use encoder-decoder architectures for direct image transformation between modalities or rely on joint learning with shared representations [56,57]. However, these methods heavily depend on the availability of strictly paired data [58–60]. Contrastive learning techniques, such as SimCLR, CLIP, and MoCo, have proven effective for cross-modal representation learning [61–63]. Nonetheless, standard implementations of SimCLR require large batch sizes (ranging from 512 to 4096 examples), leading to significant computational demands [63–65]. Although momentum-based methods that utilize queues of negatives partially mitigate this issue, the challenge of adapting these approaches to medical data with limited pairing remains unresolved [66,67].

Thus, there are no systems capable of: correctly modeling disease comorbidity through multi-label staging with suitable regularizations; performing consistent staging of DR using OCT data when fundus images are unavailable; and maintaining high performance in the face of domain shifts and significant class imbalances that are typical of real clinical data.

To address the identified issues, the paper proposes a new approach to automated diagnosis of retinal diseases that leverages a hierarchical modular architecture with a cross-modal latent bridge. The solution combines a multi-label parent model with clinical-logical regularizers and specialized child models to stage the identified diseases. To handle the challenge of joint data analysis (pairing) between OCT and fundus images, cross-modal knowledge transfer is implemented through a two-stage scheme. The first stage involves preliminary contrastive alignment of latent spaces using a

 doi:10.20944/preprints202512.0450.v1

moment encoder and negative example queues. The second stage focuses on training a latent bridge using a multi-component loss function that includes both geometric and informational components.

Additionally, during development, we introduced prototypical regularization into the child modules to enhance the geometry of the latent space and improve robustness to class imbalance. We also implemented a comprehensive calibration and post-processing system that includes optimizing class-specific thresholds and applying clinical-logical rules.

2. Methods

The study implements a hierarchical modular system (HMS). After analyzing clinical needs and the limitations of existing approaches, we propose the following testable hypotheses:

Hypothesis 1: A hierarchical architecture with specialized child staging models will achieve higher accuracy in classifying AMD and DR stages by decomposition of a complex task into more manageable subtasks.

Hypothesis 2: A loss function that employs class-balanced weighting is more effective than the standard Binary Cross-Entropy (BCE) for a multi-label task with class imbalance in OCT images.

Hypothesis 3: Using a contrastive loss function allows for the training of a cross-modal OCT to fundus image bridge under conditions where strictly paired data is limited, and it will outperform alternative loss functions in terms of cross-modal alignment quality.

Hypothesis 4: Calibrating decision thresholds using the F1-optimization method across classes will reduce the expected model calibration error and enhance clinical applicability compared to using a fixed threshold of 0.5.

Hypothesis 5: The HMS system exhibits high stability across different scanners.

2.1. Overview of the System Architecture

The HMS system features a three-level hierarchical architecture that decomposes complex diagnostic tasks into specialized subtasks and merges their results through cross-modal representation alignment [68,69].

The problem is formulated as a multi-label classification task: for each class, an independent probability is predicted using a sigmoid model and binary cross-entropy. This approach allows for the simultaneous presence or absence of multiple diseases to be expressed [70–72].

After establishing a general diagnosis, the algorithm proceeds through a series of more detailed decision-making stages for the identified diseases. The basic (parent) multi-objective model is responsible for the initial classification of pathologies as either absent or present. Following this, specialized components perform their specific tasks within a narrower context, based on the outputs of the parent model. This approach enhances the system's interpretability, aids in debugging, and improves its tolerance for errors [73,74].

Since OCT is widely regarded as the "gold standard" for diagnosing AMD and is essential for quantitatively assessing retinal thickness and DME [75,76], the AMD staging according to the AREDS is utilized in this work. In contrast, for staging DR, the most commonly used imaging modality is fundus photography, as the ETDRS and ICDR scales were developed explicitly for color retinography [77,78].

Following this, the architecture incorporates a fundus model for classifying DR stages. It also features a trainable latent "bridge" that translates characteristics from the OCT space into the fundus encoder's latent space. This setup enables DR assessments even when a fundus image is unavailable. Thus, the architectural structure of the HMS system comprises four key components: a parent model for multi-target classification of OCT scans, an AMD model for staging, a fundus classification model that determines DR stages based on fundus images, and a cross-modal OCT-to-Fundus bridge [79].

To ensure the proper functioning of the latent bridge, a contrastive alignment of the OCT and fundus latent spaces is performed in advance. This process maximizes the similarity of positive pairs while dispersing negative pairs, thereby reducing the "modal gap" between the encoders. Contrastive learning formalizes this objective using the InfoNCE loss function. Simultaneously, the moment

 doi:10.20944/preprints202512.0450.v1

encoder and negative queue (MoCo) variants build a comprehensive, consistent dynamic dictionary of negative examples [79,80].

By using alignment and a trained bridge, the system can perform DR staging without fundus images. It achieves this by projecting the OCT embedding into the fundus space and then applying the fundus classifier head, which ensures alignment with the gold standard.

The diagram in Figure 1 illustrates the structure and data flow between the components of the HMS system.

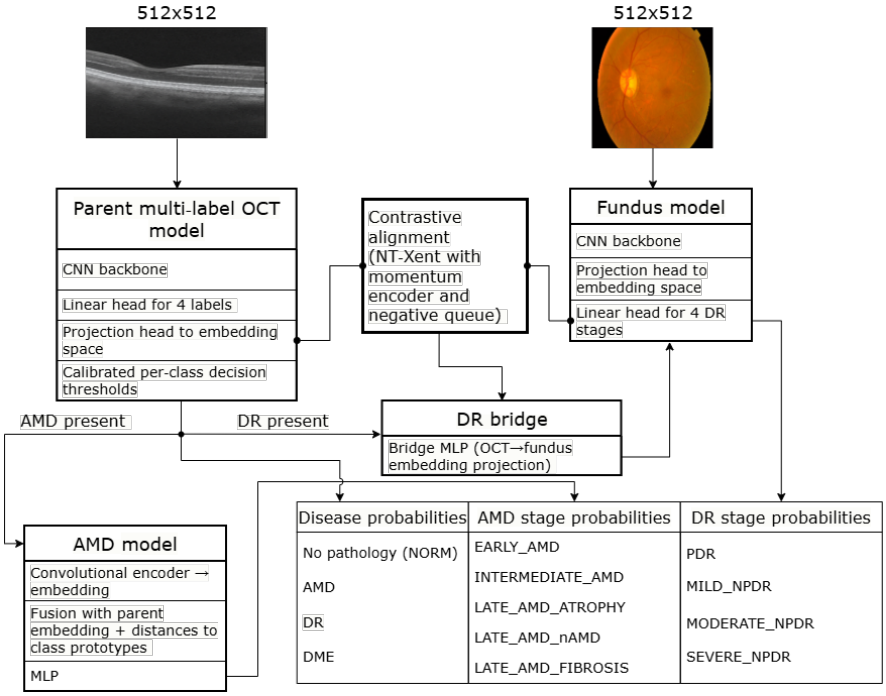

Figure 1. An overview of the general architecture of the HMS system, highlighting data flows and specialized modules.

2.2. Operation of HMS Components

For multi-label or multi-objective classification, BCE with focal boosting is used. In the multi-objective setting, each label is treated as an independent binary problem with logits and a sigmoid activation function [81].

The focal loss function reduces the contribution of well-classified examples and puts greater emphasis on complex and rare cases. It does this through the modifying factor $(1-p_t)^\gamma$, which effectively shifts the learning focus towards the less-frequent instances in the distribution. This approach enhances the model's robustness to class imbalance [82]. In practical situations, imbalance can also be addressed by using class weights α_c [83].

For a batch of N examples and $C = 4$ labels {No diseases (NORM), AMD, DR, DME}, with logits $s_{n,c}$ and probabilities $p_{n,c} = \sigma(s_{n,c})$, where $\sigma(\cdot)$ is the sigmoid function, the focal-weighted BCE is expressed as:

$$\mathcal{L}_{\text{Focal-BCE}} = \frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C}\alpha_c\,(1-p_{t,nc})^{\gamma}\,(-\log p_{t,nc}),\quad p_{t,nc} = \begin{cases} p_{n,c}, & y_{n,c}=1,\\ 1-p_{n,c}, & y_{n,c}=0,\end{cases}$$

which is the standard form for multi-objective problems and naturally generalizes BCE to the case with a focusing factor γ [82].

To identify the most effective backbone architecture, a thorough study was conducted involving ResNet18, ResNet34, EfficientNet-B0, and ConvNeXt-Tiny. The findings are detailed in the Results section. All architectures were pretrained on ImageNet, facilitating the transfer of essential low- and

 doi:10.20944/preprints202512.0450.v1

mid-level features and thereby enhancing convergence on medical images [84]. The pretrained weights for grayscale input were obtained by averaging the first convolutional layer's weights across the RGB channels.

2.2.1. The Parent Model

The parent model is designed as a multi-label classifier. In this approach, the softmax function is replaced with independent sigmoid activations, using binary cross-entropy for each label. The sigmoid function, defined as $\sigma(x) = 1/(1 + e^{-x})$, converts the logit (the neural network's raw output) into a probability ranging from 0 to 1 for each class independently. This method effectively models disease comorbidity while avoiding the artificial exclusivity characteristic of softmax approaches [68–72].

The penalty, $\mathcal{R}_{co} = \mathbb{E}[p_{AMD} \cdot p_{DR}]$, is designed to reduce the co-activation of AMD and DR. This measure was introduced as a heuristic specific to the dataset because there are few or no clear examples of cases where AMD and DR are co-labeled in the original sample. Its purpose is to avoid inconsistent decisions during the deployment phase of the current version of the system.

This setup facilitates the joint modeling of multiple states while implementing clinical-logical regularizations that ensure consistency between predictions of "normal" and "pathological" conditions. To leverage the benefits of binary cross-entropy for multi-objective problems, we use Focal-BCE with label smoothing. This approach includes a focusing factor that enhances the contribution of "difficult" examples, addressing issues related to class imbalance.

To enhance clinical reliability, decision threshold calibration based on per-class F1-optimization was implemented [85]. This approach yielded a set of asymmetric thresholds that achieved an optimal balance between precision and recall for each diagnosis.

The detailed operating principle, routing logic to specialized modules, and interpretability requirements are outlined in the Supplementary Materials (SM Section 4.1). Detailed class metrics, confidence intervals, and the evaluation protocol are provided in Table SM-5.

2.2.2. The Child Model

To refine the diagnosis of AMD into five stages, a hybrid feature fusion approach is employed. This method combines local features extracted from a CNN with global context from the parent model and the prototypical geometry of the latent space. This approach aligns with multi-modal and multi-source feature-fusion practices, which consistently enhance quality by effectively integrating diverse features [86].

The central innovation involves transferring knowledge of DR stages from fundus images to OCT. Initially, the feature-extraction models, referred to as the parent model and the Fundus model, are trained separately. The Fundus model classifies DR stages according to international standards into four categories: MILD_NPDR, MODERATE_NPDR, SEVERE_NPDR, and PDR. Its architecture uses a convolutional encoder similar to that of the parent model, but is specifically adapted for grayscale images.

The cross-modal bridge from OCT to Fundus is trained in two stages. First, a contrastive alignment of the latent spaces is performed to narrow the modal gap. It is followed by a small regression projector that ensures a consistent projection into the fundus space, allowing for staging of DR even when paired data is not available. This modular approach enhances the solution's clinical explainability, manageability, and scalability, enabling independent improvements to individual components without disrupting the overall architecture.

The detailed operating principles of the child modules and the bridge are outlined in the Supplementary Material (SM Section 4.2). Additionally, the staging metrics, an analysis of common errors, and the impact of prototypal regularization are presented in Tables SM 6 and 7.

3. Results

This chapter outlines the experimental program established to assess the HMS system. It encompasses the development of a multimodal dataset simulating clinical conditions, a comparison of

 doi:10.20944/preprints202512.0450.v1

baseline convolutional architectures, threshold calibration, and an evaluation of the parent model and specialized staging modules. Additionally, it includes a cross-modal bridge analysis that investigates probability calibration, computational efficiency, and verification across different scanners.

3.1. Creating a Data Set

The experimental dataset is a comprehensive multimodal collection of medical images comprising 8,159 images across two primary ophthalmic imaging modalities. The dataset mirrors clinical practice, featuring 4,047 OCT images with detailed multi-label annotations and 4,112 fundus images. A significant aspect of this dataset is the limited yet crucial component of paired OCT and fundus images: there are only 128 pairs, representing 3.1% of the total dataset.

To ensure comprehensive coverage of various pathological conditions, the dataset was compiled from three distinct sources. The first source is an in-house clinical dataset obtained from the Optimed Laser Vision Restoration Center in Ufa, Russia, which constitutes the majority of the OCT images, totaling 2,185 images. The second source is the publicly available Optical Coherence Tomography Image Database (OCTID), developed by the University of Waterloo [87]. The third source is the OCT-AND-EYE-FUNDUS-DATASET, created specifically for the study of DME and DR. This collection includes 1,548 fundus images and 1,113 macular OCT images [88].

The final dataset, including class distribution across modalities, imbalance metrics, source descriptions, and a five-fold cross-validation strategy among patients, is presented in the Supplementary Materials (SM Section 4.3, Tables SM-1 and SM-2).

A methodologically key feature is the use of strict patient-identifier separation, which prevents information leakage between the training and test sets and ensures a fair assessment of the system's generalization ability on new patient data.

3.2. A Comparison of Backbone Architectures for Choosing a Base Classifier Model

To systematically evaluate and justify the selection of a backbone convolutional architecture, a comprehensive comparative study was conducted on four modern architectures: ResNet18, ResNet34, EfficientNet-B0, and ConvNeXt-Tiny. These architectures were chosen for their common use in medical imaging and for the balance they offer among accuracy, computational efficiency, and the number of trainable parameters [89,90].

To maintain methodological rigor, all models were trained using a uniform experimental protocol, with the same optimization hyperparameters and data augmentation strategies applied consistently across all models. The optimizer used was AdamW, set with a learning rate of 3×10^{-4} and a cosine annealing scheduler [91,92].

The mini-batch size consisted of 64 images. To enhance the model's robustness, augmentation techniques were applied, including random horizontal flips, $\pm 10^{\circ}$ rotations, and brightness and contrast adjustments within ± 0.1 [93–95].

For comparison, we utilized unified macro-F1 and micro-F1 metrics, which are standard aggregates for multi-class classification. These metrics are calculated using micro- and macro-averaging of Precision, Recall, and F1 scores across different classes [96]. Additionally, we employed multi-class ROC-AUC with binarization (where $p > 0.5$) [97]. The dataset was split into training, validation, and test sets at 80/10/10. We selected BCEWithLogitsLoss (binary cross-entropy with logits) as the loss function [98,99]. The details, which include Macro/Micro-F1 metrics, Hamming loss, Jaccard index, and computational efficiency metrics, are provided in the Supplementary Materials (SM Section 4.4, Table SM-3).

As a result, architectures designed for compactness and efficient scaling, specifically the EfficientNet family, outperformed deep residual networks in terms of discrimination performance while using significantly fewer parameters. In clinical applications, it is crucial to balance classification accuracy with computational demands, as the selected architecture impacts inference latency. A methodologically significant finding is that the advantages of compact architectures were consistent across both the validation set and cross-validation, demonstrating their stability in generalization. The

 doi:10.20944/preprints202512.0450.v1

chosen backbone architecture served as the foundation for developing all subsequent components of the hierarchical system.

3.3. Calibrating Thresholds to Compensate for Class Imbalance

To address class imbalance in a multi-label classification setting, it is crucial to optimize class-specific thresholds [100]. The standard threshold of 0.5 is often inadequate, particularly in cases of severe class imbalance. To tackle this limitation, a systematic comparison of four adaptive threshold calibration methods was conducted. Details can be found in the Supplementary Materials, specifically in Section 4.5, Table SM-4.

The F1-optimization methodology, applied independently to each class, demonstrated superior performance to other approaches, providing an optimal balance between sensitivity and specificity across disease prevalences. This strategy employs an aggressive detection threshold of 0.15 for AMD to minimize the risk of missing cases in the late stages of the disease, when the possibility of vision loss is exceptionally high. In contrast, a more conservative approach for DME is used, with a threshold of 0.78, to prevent unnecessary interventions in patients with significant comorbidities [101]. The optimal thresholds obtained were as follows: NORM = 0.29, AMD = 0.15, DR = 0.67, and DME = 0.78. These thresholds reflect adjustments for class imbalance in the dataset, where AMD accounts for over 50% of the OCT samples, while DME accounts for only 8.6% of the cases.

Methodologically, this approach utilizes the principle of cost-sensitive learning. The implementation of individually calibrated thresholds resulted in a notable performance improvement compared to the baseline threshold of 0.5.

3.4. Outcomes of the Parent Model Operation

The parent multi-label classification model demonstrated high performance on the stratified test set, effectively identifying multiple comorbid pathologies within a single diagnostic cycle. Detailed metrics are available in the Supplementary Materials, specifically in Section 4.6, Table SM-5.

The AMD class exhibited ideal performance, with precision, recall, and F1-score values of 1.00. It is primarily attributed to its prevalence in the dataset and the prominent morphological features visible on OCT. The Normal class (NORM) achieved a recall of 1.00 and a precision of 0.993, indicating only one false-positive result out of 136 cases. It aligns with a conservative screening strategy that emphasizes minimizing the risk of overdiagnosis. The DME class recorded a recall of 0.978, corresponding to one missed case out of 50. It may be due to borderline cases involving minimal intraretinal fluid that are on the threshold of the clinical criteria for DME with central involvement [102]. Meanwhile, the DR class achieved a recall of 0.990, with two false-negative results, both associated with MILD_NPDR and presenting a single microaneurysm. It reflects the limited information value of OCT for staging DR [103,104].

The performance distribution emphasizes the importance of adaptive calibration tailored to each specific diagnosis.

3.5. Results from the Specialized AMD Staging Module

The specialized module developed for diagnosing AMD showed impressive diagnostic performance, achieving an overall accuracy of $98.3 \pm 1.4\%$. Detailed results can be found in the Supplementary Materials, specifically in SM Section 4.7, Table SM-6. A significant observation is that late-stage AMD, characterized by clear morphological changes such as atrophy, subretinal neovascularization, and fibrosis, is classified with nearly perfect accuracy. However, distinguishing between the early and intermediate stages poses a systematic challenge due to the continuous nature of disease progression and the subjective boundaries between these stages, even among experts. This challenge highlights the inherent uncertainty of the AREDS clinical classification scheme.

A key finding of this study is that the model shows high Precision at all stages, which is crucial for reducing false-positive diagnoses of late-stage conditions that require aggressive treatment.

 doi:10.20944/preprints202512.0450.v1

The performance achieved surpasses previously published results for five-class AMD staging using OCT images.

3.6. The Results from the Specialized Module for Staging DR

The developed model for classifying DR stages using fundus images achieved an overall accuracy of $94.8 \pm 0.9\%$. A significant methodological advancement is that the model can stage DR according to the international ICDR classification based solely on synthesized fundus representations derived from OCT scans. It demonstrates a successful cross-modal transfer of diagnostically significant features.

The performance distribution by stage reveals a distinct pattern: intermediate stages of nonproliferative retinopathy can be classified with high accuracy, driven by clear morphological features such as microaneurysms and hemorrhages. However, borderline cases between stages require expert verification because they rely on subjective clinical criteria. Importantly, the performance achieved is comparable to that of models trained directly on real fundus images. For detailed metrics, please refer to the Supplementary Materials, specifically Section 4.8, Table SM-7.

3.7. The Cross-Modal Bridge and Analysis of Cross-Modal Inconsistencies

The contrastive alignment of OCT and fundus images was achieved using the NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss function, along with a pulse encoder set to m=0.999 and a negative queue containing 512 samples [105]. The training process exhibited a two-phase dynamic typical of contrastive learning. During the first 20 epochs, the loss function decreased rapidly on the training set. Subsequently, the Recall@1 metric, which measures the accuracy of retrieving the first nearest neighbor, gradually improved, reaching its highest value at the 54th epoch. The peak value of Recall@1=0.411 indicates a successful alignment of the latent-space geometries of the two modalities, facilitating reliable cross-modal image retrieval [106].

The cross-modal bridge between OCT and fundus images was trained using a multi-component loss function that combined seven regularization components: mean squared error (MSE), cosine closeness (Cosine), Kullback-Leibler divergence (KL-divergence), InfoNCE contrastive loss, prototype loss (Prototype), maximum mean discrepancy (MMD), and correlation alignment (CORAL) [107–109]. The training process is described in detail in the Supplementary Material, specifically in SM Section 4.9, Table SM-10. A significant improvement in the fundus consistency score, which increased from 0.815 to 0.984, occurred between epochs 18 and 30. It indicates a qualitative shift in the bridge's ability to generate semantically consistent representations. These results demonstrate that the multi-component learning strategy not only facilitates geometric alignment of feature spaces but also achieves a high degree of semantic correspondence between OCT and fundus imaging modalities, which is crucial for subsequent cross-modal diagnostic tasks related to retinal diseases [110].

An ablative study (with detailed results available in the Supplementary Materials, SM Section 4.9, Table SM-10) highlights the importance of contrastive alignment. The main methodological finding is that information-theoretic components, such as InfoNCE and KL-divergence, significantly improve the quality of representation alignment. In contrast, geometric components such as MSE and Cosine play a vital yet less prominent role in maintaining metric consistency. Removing any of these components results in a statistically significant decline in performance, demonstrating the synergistic effect of optimizing multiple components and underscoring the need to balance the various aspects of representation alignment.

To evaluate the quality of cross-modality transfer between OCT and fundus representations, it is essential to identify cases in which the bridge model struggles to project OCT into the fundus representation space accurately. The main challenges include image quality artifacts, physiological variations at the boundaries of clinical staging criteria, and the anatomical limitations of each modality's field of view. It is important to understand that these discrepancies do not indicate algorithm failures; instead, they highlight a fundamental incompleteness in the information provided by one modality compared to the other.

 doi:10.20944/preprints202512.0450.v1

Quantitative analysis shows that the proportion of cases with significant discrepancies is less than 4% of the paired data, indicating the method's overall stability. However, the identified issues highlight several critical categories of clinical situations that require additional expert verification. Further information regarding this research step can be found in the Supplementary Materials, specifically in SM Section 4.9, Table SM-11, and Figure SM-1.

3.8. Calibration Assessment, Risk Interpretation in Clinical Scenarios, and Computational Efficiency

The calibration evaluation of the HMS system shows it has high reliability for probabilistic predictions in clinical use. The Expected Calibration Error is significantly below the 5% threshold commonly accepted in medical literature, which distinguishes between well-calibrated and poorly calibrated models. Importantly, the calibration is consistent across a wide range of predicted probabilities, ensuring that the results are interpretable for both highly sensitive cases and borderline situations.

Experiments to assess computational efficiency were conducted on a hardware configuration comprising an NVIDIA RTX 3060 GPU and a Ryzen 7 3700X CPU. The modular organization allows tasks to be separated by specialization and executed independently or in a pipeline, thereby improving portability, interpretability, and scalability across different clinical scenarios. The full pipeline processes an OCT volume of 128 B-scans in 3.15 s on the GPU, with an average latency of 24.6 ms per image, meeting real-time requirements for visualization and not limiting the clinical workflow throughput [113]. A description of the tests is provided in the Supplementary Materials, SM Section 4.10, Figure SM-2, Table SM-12.

3.9. Comparative Analysis of Model Effectiveness in Diagnosis and Staging

To present the different metrics in Figure 2, they have been scaled to 0–100, with higher values indicating better performance.

Models from the ResNet, ConvNeXt, and EfficientNet families were compared to HMS in a two-stage diagnostic and staging task. The finding that EfficientNet-B0 achieves the highest peak accuracy but does not perform as well as HMS on macro-AUROC and ECE highlights the well-documented issue of overconfidence in modern CNNs. This distinction emphasizes the difference between optimizing for accuracy and achieving high-quality probabilistic calibration [111].

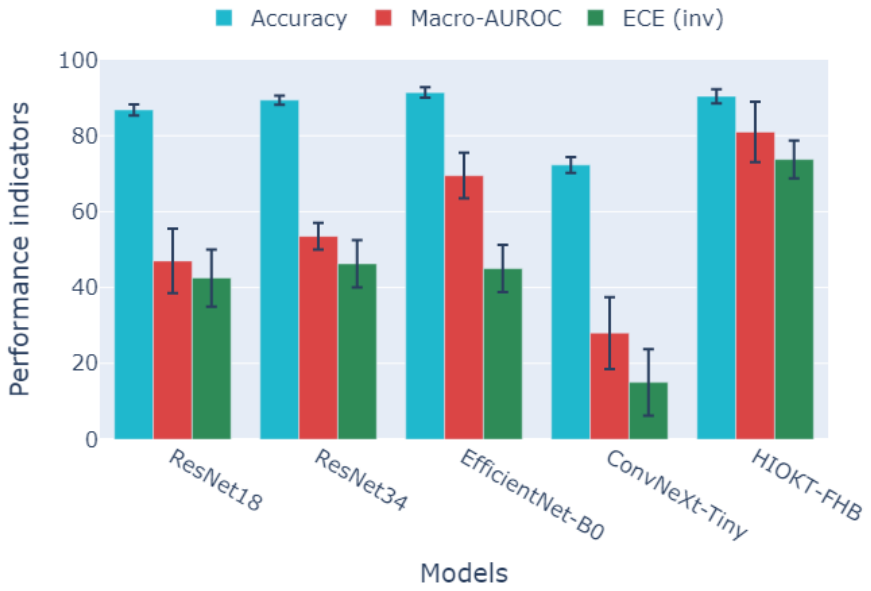

Figure 2. A comparison of the performance of the HMS system with contemporary baseline methods.

The HMS model demonstrates superior performance compared to other variants in macro-AUROC and calibration, while maintaining similar accuracy. It indicates a trade-off between probabilistic reliability and peak accuracy, favoring the hierarchical approach. Macro-AUROC averages

 doi:10.20944/preprints202512.0450.v1

across all classes and is more resilient to class imbalance. The HMS's advantage in this metric suggests that it offers better overall class separability in a multi-class setting. Additionally, a low ECE indicates a closer alignment between predicted confidence and the actual rates of correct predictions, which is essential for clinical decision-making under uncertainty. The application of the HMS system surpasses the direct use of standard CNNs for the multi-class classification and staging of AMD, DR, and DME. In contrast, CNNs exhibit an ECE of 4.3–5.3%, a finding corroborated by studies on the calibration of deep neural networks in medical imaging [100,112].

3.10. Cross-Scan Validation and Robustness to Domain Shifts

To evaluate the generalizability of the HMS system across various hardware configurations, cross-scanner validation was performed using a clinical dataset from the Optimed Laser Vision Restoration Center in Ufa, Russia. This dataset consisted of 2,185 OCT images. The cross-domain validation studies, summarized in the Supplementary Materials, specifically in SM Section 4.11, Table SM-13.

The results reveal an asymmetric pattern of cross-scan generalization. Transferring from Avanti XR to REVO NX achieves significantly higher accuracy (86.1%) and discriminatory power (AUROC 0.896). In contrast, the reverse transfer shows a noticeable decline in performance, with accuracy at 74.7% and AUROC at 0.769. The systematic advantage of HMS over the best-performing baseline method, EfficientNet-B0, by 7.2 to 7.3 percentage points in both directions highlights the hierarchical architecture's improved robustness to domain variations. This finding aligns with existing evidence supporting the effectiveness of multi-stage and adaptive approaches for cross-domain generalization. While performance degradation during transfers between scanners is inevitable due to differences in optical characteristics, resolution, noise patterns, and image-acquisition parameters, the hierarchical approach significantly reduces this degradation. It achieves this by allowing for modular specialization and contrast alignment.

Essential for clinical application, the system's performance on external source data remains clinically acceptable without retraining or fine-tuning, demonstrating sufficient generalization for deployment in multi-center scenarios.

4. Discussion

The performance of the HMS hierarchical modular architecture demonstrates its superiority over traditional monolithic approaches. This architecture combines a multi-target parent model, specialized staging modules, and a cross-modality bridge, effectively modeling comorbidity through independent sigmoid outputs and focal-weighted BCE with class-balanced α_c coefficients. Compared with a strong single-stage baseline model, HMS achieves comparable peak accuracy, improved global separability (macro-AUROC), and enhanced probability calibration. These factors are crucial for ensuring the clinical reliability of decisions made under uncertainty (for detailed metrics and tables, refer to SM Sections 4.4–4.6). A significant distinction between HMS and studies such as VisionTrack is its ability to perform DR staging using OCT rather than relying on fundus imaging. It is made possible through a two-stage cross-modality alignment and a latent bridge, thus removing the impractical need for strictly paired multimodal data in clinical settings. It supports the validity of the first hypothesis outlined in the Methods chapter.

Using class-balanced weights and focal BCE enhances sensitivity to rare pathologies in multi-label settings. This improvement is demonstrated by higher overall metrics and increased robustness to class imbalance (as detailed in SM Sections 4.5–4.6). When combined with modular decomposition, these methods produce significantly better performance across various classes. This approach accounts for epidemiological factors and the expression of visual features, reflecting real clinical frequencies and helping reduce the omission of rare conditions. Additionally, the representation geometry generated in child modules, along with prototypal regularization, further enhances interclass separability and robustness against the "long tail" of distributions. This outcome confirms the validity of the second hypothesis presented in the Methods chapter.

 doi:10.20944/preprints202512.0450.v1

The third hypothesis suggests that a contrastive loss function can facilitate the training of a cross-modality bridge between OCT and fundus images without requiring strictly paired data. It is supported by evidence demonstrating the staging of DR via projection into a latent fundus space and subsequent classification according to the ICDR standards (for additional information, refer to SM Section 4.9). The two-stage approach employed—comprising NT-Xent with a momentum encoder and a queue of negative samples, alongside a regression projector with a multi-component loss function—ensures robust spatial alignment and effective transfer of diagnostically significant features across modalities. For clinical safety, a cosine similarity threshold of 0.8 was established. This results in 3.9% of cases being flagged for manual verification due to artifacts and borderline manifestations, thereby minimizing the risk of false positives while allowing for automation in 96.1% of cases (illustrations of inconsistencies can be found in SM Section 4.9).

Class-specific F1 threshold calibration significantly reduces calibration error and enhances clinical applicability compared to a fixed threshold of 0.5. This improvement is evidenced by a higher macro-F1 score of 0.989, compared to 0.923 at the 0.5 threshold, and a lower ECE of $2.1 \pm 0.4\%$, in contrast to the typical 4–6% seen with uncalibrated CNNs. Calibration curves and summaries are available in Section 4.10 of the Supplementary Materials. Furthermore, the improved calibration remains effective across a wide range of probabilities, which enhances the interpretability of risk in both high-sensitivity and borderline cases. It is crucial for making clinical decisions under uncertainty. Overall, these findings confirm the validity of the cost-sensitive postprocessing approach for tasks characterized by pronounced class imbalance and unequal error costs, thereby supporting the fourth hypothesis outlined in the Methods chapter.

The fifth hypothesis regarding the robustness of transfers between OCT scanners is partially confirmed: the system maintains a clinically acceptable level of accuracy during transfers, although asymmetric degradation is observed (86.1% when transferring from Avanti XR to REVO NX, and 74.7% from REVO NX to Avanti XR). HMS consistently outperforms EfficientNet-B0 by 7.2–7.3 percentage points in both directions, highlighting the benefits of hierarchy and prototypical regularization for cross-domain generalization (see SM Section 4.11 for full tables). The remaining decrease in within-domain performance can be attributed to differences in optical parameters, noise-reduction algorithms, and resolution across manufacturers (Optovue vs. Optopol), necessitating further domain adaptation.

Compared to systems focused on single-modality scenarios and metadata integration (e.g., VisionTrack), HMS addresses more general cross-modality transfer without paired data and explicitly evaluates calibration and robustness to domain shifts, thereby improving clinical validity and the transferability of results. The presented integral indices (macro-F1=0.989 ± 0.006; micro-F1=0.994 ± 0.003; Jaccard index=0.996 ± 0.001) should be interpreted with caution due to differences in datasets and protocols, but they confirm the competitiveness of the proposed architecture for multi-target OCT classification. Links to detailed protocols and partitions are provided in the Supplementary Materials for reproducibility and independent verification.

The system combines high accuracy, calibrated probabilities, and modular explainability to support a range of scenarios, including screening for primary AMD, DR, and DME. It also aids in AMD staging according to the AREDS guidelines, enables cross-modality DR staging without fundus images, and facilitates DME monitoring through interpretable risk scores. The entire processing pipeline can analyze 128 B-scans in approximately 3.15 seconds on a GPU, yielding about 24.6 milliseconds per image. Each module can operate independently or in tandem within the pipeline. The system delivers a combined performance of 7.72 GFLOPs and uses 37.0 million parameters, making it suitable for real-time applications. However, CPU performance may limit functionality in scenarios without specialized hardware (refer to SM Section 4.10 for exact profiles). These performance characteristics make the system practical for integration into a clinical workflow, with the flexibility to adapt to different resource availability.

Key limitations of the study include the limited availability of paired OCT-Fundus data, with only 128 pairs for bridge training. There is also a severe class imbalance and the need for an $\mathcal{R}_{co}$

 doi:10.20944/preprints202512.0450.v1

penalty due to the lack of clear examples of comorbidity between AMD and DR. Additionally, the limited field of view in OCT scans limits the detection of peripheral lesions, and the labeling was conducted within a single center without multicenter consensus. These factors contribute to the observed inconsistencies and uneven transfer between scanners, underscoring the need for further validation and retraining across larger, more diverse patient cohorts. The accompanying materials include an analysis of common failures and criteria for flagging cases that require expert verification (see SM Section 4.9).

Future steps will involve multi-domain contrastive learning with adversarial alignment, the accumulation of paired data addressing real-world AMD and DR comorbidity to eliminate the $\mathcal{R}_{co}$ penalty, the implementation of uncertainty assessment methods (such as ensembles, Monte Carlo dropout, and Bayesian approaches), and the development of attention mechanisms to enable visually explainable decisions. This program aims to enhance the system's transferability, trustworthiness, and usability in multi-center and resource-constrained environments.

5. Conclusions

The HMS system demonstrates that a hierarchical modular architecture with a cross-modality latent bridge achieves high accuracy (macro-F1=0.989 $\pm$ 0.006, micro-F1=0.994 $\pm$ 0.003) and calibration (ECE=2.1 $\pm$ 0.4%) for multi-target classification and staging of AMD, DR, and DME, even without paired OCT and fundus data. A significant contribution of this system is its ability to perform OCT-based DR staging using a contrast-aligned latent bridge, thereby eliminating reliance on fundus images in 96.1% of cases. Additionally, class-specific F1-threshold optimization addresses class imbalance, outperforming monolithic CNNs in macro-AUROC and calibration, while maintaining comparable accuracy. Cross-scan validation revealed moderate robustness (86.1% accuracy for Avanti XR→REVO NX, 74.7% for the reverse direction), with a systematic 7.2–7.3 p.p. outperformance over EfficientNet-B0. However, an absolute performance drop of 15.7 p.p. indicates the need for further adaptation to domain shifts.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure SM-1: Scatterplot of cosine similarity versus Fundus consistency for the cross-modal bridge for determining DR stages; Figure SM-2: Calibration curve and contributions to Expected Calibration Error (ECE); Table SM-1: Dataset Parameters; Table SM-2: Characteristics of the experimental dataset with class distribution and statistical parameters; Table SM-3: Comparison of backbone architectures for the parent model (multi-label classification on OCT); Table SM-4: Comparison of threshold calibration methods; Table SM-5: Multi-objective performance metrics of the parent model on the test sample; Table SM-6: Multi-objective metrics of the AMD model on the test sample; Table SM-7: Multi-objective metrics of the Fundus model on the test sample; Table SM-8: Progress of Contrastive Learning; Table SM-9: Cross-modal Bridge Learning Progress; Table SM-10: Ablative Study of Loss Function Components; Table SM-11: Cross-modal Discrepancy Analysis; Table SM-12: Computational efficiency of system components; Table SM-13: Cross-scan validation

Author Contributions: Conceptualization, E.A.L.; methodology, E.A.L.; software, E.A.L.; validation, G.M.I. and T.R.M.; formal analysis, G.M.I.; investigation, E.A.L. and G.M.I.; resources, T.R.M. and G.M.I.; data curation, G.M.I.; writing—original draft preparation, E.A.L.; writing—review and editing, E.A.L., G.S.V., E.P.T. and R.V.K.; visualization, E.A.L.; supervision, E.P.T. and G.M.I.; project administration, T.R.M. and R.V.K.; funding acquisition, E.P.T. and R.V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Science and Higher Education of the Russian Federation within the state assignment for UUST (agreement № 075-03-2024-123/1 dated 15 February 2024) and conducted in the research laboratory "Sensor systems based on integrated photonics devices" of the Eurasian Scientific and Educational Center.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202512.0450.v1

References

1. A. D. Bhatwadekar, A. Shughoury, A. Belamkar, and T. A. Ciulla, "Genetics of Diabetic Retinopathy, a Leading Cause of Irreversible Blindness in the Industrialized World," *Genes*, vol. 12, no. 8, p. 1200, July 2021, doi: 10.3390/genes12081200.
2. M. Benhamza, M. Dahlui, and M. A. Said, "Determining direct, indirect healthcare and social costs for diabetic retinopathy management: a systematic review," *BMC Ophthalmol*, vol. 24, no. 1, p. 424, Sept. 2024, doi: 10.1186/s12886-024-03665-6.
3. H. Cao, Y. Zhang, N. Zhang, and X. Ma, "Clinical trial landscape of diabetic retinopathy: global advancements and future directions," *International Journal of Surgery*, Sept. 2025, doi: 10.1097/JS9.0000000000003475.
4. A. S. A. Sakini et al., "Diabetic macular edema (DME): dissecting pathogenesis, prognostication, diagnostic modalities along with current and futuristic therapeutic insights," *Int J Retina Vitreous*, vol. 10, p. 83, Oct. 2024, doi: 10.1186/s40942-024-00603-y.
5. A. Stahl, "The Diagnosis and Treatment of Age-Related Macular Degeneration," *Dtsch Arztebl Int*, vol. 117, no. 29–30, pp. 513–520, July 2020, doi: 10.3238/arztebl.2020.0513.
6. Z. L. Teo et al., "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, Nov. 2021, doi: 10.1016/j.ophtha.2021.04.027.
7. N. Cheung, C. M. G. Cheung, S. J. Talks, and T. Y. Wong, "Management of diabetic macular oedema: new insights and global implications of DRCR protocol V," *Eye*, vol. 34, no. 6, pp. 999–1002, June 2020, doi: 10.1038/s41433-019-0738-y.
8. R. Ratnapriya and E. Y. Chew, "Age-related macular degeneration – clinical review and genetics update," *Clin Genet*, vol. 84, no. 2, pp. 160–166, Aug. 2013, doi: 10.1111/cge.12206.
9. W. L. Wong et al., "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, Feb. 2014, doi: 10.1016/S2214-109X(13)70145-1.
10. H.-T. Lin et al., "The Association between Diabetic Retinopathy and Macular Degeneration: A Nationwide Population-Based Study," *Biomedicines*, vol. 12, no. 4, p. 727, Mar. 2024, doi: 10.3390/biomedicines12040727.
11. A. J. Patel, K. Downes, A. Davis, and A. Das, "Are Proliferative Diabetic Retinopathy and Diabetic Macular Edema two different disease processes? A Retrospective Cross-sectional Study," *Invest. Ophthalmol. Vis. Sci.*, vol. 53, no. 14, p. 377, Mar. 2012.
12. W. Wang, G. Sun, A. Xu, and C. Chen, "Proliferative diabetic retinopathy and diabetic macular edema are two factors that increase macrophage-like cell density characterized by en face optical coherence tomography," *BMC Ophthalmology*, vol. 23, p. 46, Feb. 2023, doi: 10.1186/s12886-023-02794-8.
13. C. J. Flaxel et al., "Age-Related Macular Degeneration Preferred Practice Pattern®," *Ophthalmology*, vol. 127, no. 1, pp. P1–P65, Jan. 2020, doi: 10.1016/j.ophtha.2019.09.024.
14. B. E. Bouma et al., "Optical coherence tomography," *Nat Rev Methods Primers*, vol. 2, p. 79, 2022, doi: 10.1038/s43586-022-00162-2.
15. C. Metrangolo et al., "OCT Biomarkers in Neovascular Age-Related Macular Degeneration: A Narrative Review," *J Ophthalmol*, vol. 2021, p. 9994098, July 2021, doi: 10.1155/2021/9994098.
16. G. Virgili et al., "Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy," *Cochrane Database Syst Rev*, vol. 2015, no. 1, p. CD008081, Jan. 2015, doi: 10.1002/14651858.CD008081.pub3.
17. Y. Attiku et al., "Comparison of diabetic retinopathy severity grading on ETDRS 7-field versus ultrawide-field assessment," *Eye (Lond)*, vol. 37, no. 14, pp. 2946–2949, Oct. 2023, doi: 10.1038/s41433-023-02445-8.
18. Y. Xiao et al., "Assessment of early diabetic retinopathy severity using ultra-widefield Clarus versus conventional five-field and ultra-widefield Optos fundus imaging," *Sci Rep*, vol. 13, no. 1, p. 17131, Oct. 2023, doi: 10.1038/s41598-023-43947-5.
19. C. P. Wilkinson et al., "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, Sept. 2003, doi: 10.1016/S0161-6420(03)00475-5.
20. S. Kumari, P. Venkatesh, N. Tandon, R. Chawla, B. Takkar, and A. Kumar, "Selfie fundus imaging for diabetic retinopathy screening," *Eye*, vol. 36, no. 10, pp. 1988–1993, Oct. 2022, doi: 10.1038/s41433-021-01804-7.
21. B. J. Fenner, R. L. M. Wong, W.-C. Lam, G. S. W. Tan, and G. C. M. Cheung, "Advances in Retinal Imaging and Applications in Diabetic Retinopathy Screening: A Review," *Ophthalmol Ther*, vol. 7, no. 2, pp. 333–346, Dec. 2018, doi: 10.1007/s40123-018-0153-7.

 doi:10.20944/preprints202512.0450.v1

22. R. Prawira, A. Bustamam, and P. Anki, "Multi Label Classification Of Retinal Disease On Fundus Images Using AlexNet And VGG16 Architectures," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2021, pp. 464–468, doi: 10.1109/ISRITI54043.2021.9702817.
23. L. Ju et al., "Synergic Adversarial Label Learning for Grading Retinal Diseases via Knowledge Distillation and Multi-Task Learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3709–3720, Oct. 2021, doi: 10.1109/JBHI.2021.3052916.
24. A. T. Nair, A. M. L, and A. K. M. N, "Disease Grading of Diabetic Retinopathy using Deep Learning Techniques," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2022, pp. 1019–1024, doi: 10.1109/ICCMC53470.2022.9754113.
25. P. Zang et al., "DcardNet: Diabetic Retinopathy Classification at Multiple Levels Based on Structural and Angiographic Optical Coherence Tomography," *IEEE Trans Biomed Eng*, vol. 68, no. 6, pp. 1859–1870, June 2021, doi: 10.1109/TBME.2020.3027231.
26. L. F. Nakayama et al., "BRSET: A Brazilian Multilabel Ophthalmological Dataset of Retina Fundus Photos," *PLOS Digital Health*, vol. 3, no. 7, p. e0000454, 2024, doi: 10.1371/journal.pdig.0000454.
27. R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automated detection of mild and multi-class diabetic eye diseases using deep learning," *Health Inf Sci Syst*, vol. 8, no. 1, p. 32, Oct. 2020, doi: 10.1007/s13755-020-00125-5.
28. E. Sükei et al., "Multi-modal representation learning in retinal imaging using self-supervised learning for enhanced clinical predictions," *Sci Rep*, vol. 14, no. 1, p. 26802, Nov. 2024, doi: 10.1038/s41598-024-78515-y.
29. Z. Xu et al., "Enhancing pathological myopia diagnosis: a bimodal artificial intelligence approach integrating fundus and optical coherence tomography imaging for precise atrophy, traction and neovascularisation grading," *British Journal of Ophthalmology*, May 2025, doi: 10.1136/bjo-2024-326252.
30. "GAMMA challenge: Glaucoma grAding from Multi-Modality imAges," *Medical Image Analysis*, vol. 90, p. 102938, Dec. 2023, doi: 10.1016/j.media.2023.102938.
31. E. Hirsch, G. Dawidowicz, and A. Tal, "MedCycle: Unpaired Medical Report Generation via Cycle-Consistency," in *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 1929–1944, doi: 10.18653/v1/2024.findings-naacl.125.
32. E. Sükei et al., "Multi-modal representation learning in retinal imaging using self-supervised learning for enhanced clinical predictions," *Sci Rep*, vol. 14, no. 1, p. 26802, Nov. 2024, doi: 10.1038/s41598-024-78515-y.
33. "DDA-Net: Unsupervised cross-modality medical image segmentation via dual domain adaptation," *Computer Methods and Programs in Biomedicine*, vol. 213, p. 106531, Jan. 2022, doi: 10.1016/j.cmpb.2021.106531.
34. Z. Zhao et al., "UOPSL: Unpaired OCT Predilection Sites Learning for Fundus Image Diagnosis Augmentation," Sept. 2025. Accessed: Sept. 16, 2025. [Online]. Available: https://www.semanticscholar.org/paper/UOPSL%3A-Unpaired-OCT-Predilection-Sites-Learning-for-Zhao-Zhao/965bc2638a8df8abd665a7106b50d4d6fd5181a6
35. W. Chen, M. I. Y. Liu, C. Wang, J. Zhu, M. I. G. Li, and S. M. I. L. Lin, "Cross-Modal Causal Intervention for Medical Report Generation," 2023. Accessed: Sept. 16, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Cross-Modal-Causal-Intervention-for-Medical-Report-Chen-Liu/f818e71f883aa8eb515331b01e24ba3530968664
36. H. Ishwaran and R. O'Brien, "Commentary: The Problem of Class Imbalance in Biomedical Data," *J Thorac Cardiovasc Surg*, vol. 161, no. 6, pp. 1940–1941, June 2021, doi: 10.1016/j.jtcvs.2020.06.052.
37. Y. Zhao, Z. S.-Y. Wong, and K. L. Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection," *J Healthc Eng*, vol. 2018, p. 6275435, May 2018, doi: 10.1155/2018/6275435.
38. V. Kumar et al., "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthcare (Basel)*, vol. 10, no. 7, p. 1293, July 2022, doi: 10.3390/healthcare10071293.
39. V. Koch et al., "Noise Transfer for Unsupervised Domain Adaptation of Retinal OCT Images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, vol. 13432, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., in Lecture Notes in Computer Science, vol. 13432, Cham: Springer Nature Switzerland, 2022, pp. 699–708, doi: 10.1007/978-3-031-16434-7_67.
40. A. Heinke et al., "Cross-instrument optical coherence tomography-angiography (OCTA)-based prediction of age-related macular degeneration (AMD) disease activity using artificial intelligence," *Sci Rep*, vol. 14, p. 27085, Nov. 2024, doi: 10.1038/s41598-024-78327-0.
41. M. R. Munk et al., "OCT-angiography: A qualitative and quantitative comparison of 4 OCT-A devices," *PLoS One*, vol. 12, no. 5, p. e0177059, May 2017, doi: 10.1371/journal.pone.0177059.

 doi:10.20944/preprints202512.0450.v1

42. H. Guan and M. Liu, "Domain Adaptation for Medical Image Analysis: A Survey," *IEEE Trans Biomed Eng*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022, doi: 10.1109/TBME.2021.3117407.
43. M. Vijayan, D. K. Prasad, and V. Srinivasan, "Advancing Glaucoma Diagnosis: Employing Confidence-Calibrated Label Smoothing Loss for Model Calibration," *Ophthalmol Sci*, vol. 4, no. 6, p. 100555, June 2024, doi: 10.1016/j.xops.2024.100555.
44. T. Dawood et al., "Uncertainty aware training to improve deep learning model calibration for classification of cardiac MR images," *Med Image Anal*, vol. 88, p. 102861, Aug. 2023, doi: 10.1016/j.media.2023.102861.
45. Y. Guo, A. Liu, X. Zhu, and Y. Wang, "Calibration of Machine Learning Models for Medical Imaging: A Comprehensive Survey," *IEEE Access*, vol. 11, pp. 45789–45803, 2023, doi: 10.1109/ACCESS.2023.3275621.
46. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
47. M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.
48. Z. Liu et al., "A ConvNet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.
49. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.
50. J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
51. T. Denouden et al., "Improving Transfer Learning Through Deep Convolutional Neural Network Ensembles," in *2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852352.
52. M. Raghu et al., "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
53. A. S. Morais et al., "Optimizing Deep Learning for Image Sequence Recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 9, pp. 3228–3242, Sept. 2021, doi: 10.1109/TPAMI.2020.2990496.
54. S. Kornblith et al., "Do Better ImageNet Models Transfer Better?" in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 2656–2666, doi: 10.1109/CVPR.2019.00277.
55. C. Tan et al., "A Survey on Deep Transfer Learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, vol. 11141, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds., in Lecture Notes in Computer Science, vol. 11141, Cham: Springer International Publishing, 2018, pp. 270–279, doi: 10.1007/978-3-030-01424-7_27.
56. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science, vol. 9351, Cham: Springer International Publishing, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
57. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.
58. L. Zhang et al., "Multi-modal Deep Learning for Medical Image Analysis: A Comprehensive Survey," *IEEE Reviews in Biomedical Engineering*, vol. 15, pp. 157–179, 2022, doi: 10.1109/RBME.2021.3070036.
59. A. P. Twinanda et al., "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," *IEEE Trans Med Imaging*, vol. 36, no. 1, pp. 86–97, Jan. 2017, doi: 10.1109/TMI.2016.2593957.
60. S. M. Plis et al., "Deep learning for neuroimaging: a validation study," *Front Neurosci*, vol. 8, p. 229, Aug. 2014, doi: 10.3389/fnins.2014.00229.
61. A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.
62. K. He et al., "Momentum Contrast for Unsupervised Visual Representation Learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.
63. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 2020, pp. 1597–1607.
64. Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," in *International Conference on Learning Representations*, 2020.

 doi:10.20944/preprints202512.0450.v1

65. M. Caron et al., "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9912–9924.
66. X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15745–15753, doi: 10.1109/CVPR46437.2021.01549.
67. J.-B. Grill et al., "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21271–21284.
68. I. Diakou et al., "Multi-label classification of biomedical data," *Med Int (Lond)*, vol. 4, no. 6, p. 68, Sept. 2024, doi: 10.3892/mi.2024.192.
69. M. Priyadharshini, A. F. Banu, B. Sharma, S. Chowdhury, K. Rabie, and T. Shongwe, "Hybrid Multi-Label Classification Model for Medical Applications Based on Adaptive Synthetic Data and Ensemble Learning," *Sensors (Basel)*, vol. 23, no. 15, p. 6836, July 2023, doi: 10.3390/s23156836.
70. S. Yuan, Y. Chen, C. Ye, M. W. Bhatt, M. Saradeshmukh, and M. S. Hossain, "Cross-modal multi-label image classification modeling and recognition based on nonlinear," *Nonlinear Engineering*, vol. 12, no. 1, Jan. 2023, doi: 10.1515/nleng-2022-0194.
71. A. Lemay, C. Gros, E. Naga Karthik, and J. Cohen-Adad, "Label fusion and training methods for reliable representation of inter-rater uncertainty," *MELBA journal*, vol. 1, no. January 2023 issue, pp. 1–27, Jan. 2023, doi: 10.59275/j.melba.2022-db5c.
72. M. S. Neyestanak et al., "A Quantitative Comparison Between Focal Loss and Binary Cross-Entropy Loss in Brain Tumor Auto-Segmentation Using U-Net," *Journal of Biostatistics and Epidemiology*, Aug. 2025, doi: 10.18502/jbe.v11i1.19315.
73. C. Watanabe, "Interpreting Layered Neural Networks via Hierarchical Modular Representation," in *Neural Information Processing*, vol. 1143, T. Gedeon, K. W. Wong, and M. Lee, Eds., in Communications in Computer and Information Science, vol. 1143, Cham: Springer International Publishing, 2019, pp. 376–388, doi: 10.1007/978-3-030-36802-9_40.
74. B. N. Iduh and P. Anwaitu Fraser Egba, "An Enhanced Modular-Based Neural Network Framework for Effective Medical Diagnosis," *Journal of Computational Mechanics, Power System and Control*, May 2024, doi: 10.46253/J.MR.V7I2.A1.
75. N. Waugh et al., "Introduction to age-related macular degeneration," in *Treatments for dry age-related macular degeneration and Stargardt disease: a systematic review*, NIHR Journals Library, 2018.
76. M. Sasaki, R. Kawasaki, and Y. Yanagi, "Early Stages of Age-Related Macular Degeneration: Racial/Ethnic Differences and Proposal of a New Classification Incorporating Emerging Concept of Choroidal Pathology," *J Clin Med*, vol. 11, no. 21, p. 6274, Oct. 2022, doi: 10.3390/jcm11216274.
77. M. D. Davis et al., "Comparison of Time-Domain OCT and Fundus Photographic Assessments of Retinal Thickening in Eyes with Diabetic Macular Edema," *Invest Ophthalmol Vis Sci*, vol. 49, no. 5, pp. 1745–1752, May 2008, doi: 10.1167/iovs.07-1257.
78. B. L. Sikorski, G. Malukiewicz, J. Stafiej, H. Lesiewska-Junk, and D. Raczynska, "The Diagnostic Function of OCT in Diabetic Maculopathy," *Mediators Inflamm*, vol. 2013, p. 434560, 2013, doi: 10.1155/2013/434560.
79. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, in ICML'20, vol. 119, JMLR.org, July 2020, pp. 1597–1607.
80. K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, June 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.
81. M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014, doi: 10.1109/TKDE.2013.39.
82. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.
83. Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, June 2019, pp. 9260–9269, doi: 10.1109/CVPR.2019.00949.
84. O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vis*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
85. Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," *Mach Learn Knowl Discov Databases*, vol. 8725, pp. 225–239, 2014, doi: 10.1007/978-3-662-44851-9_15.

 doi:10.20944/preprints202512.0450.v1

86. S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, Mar. 2022, doi: 10.1093/bib/bbab569.
87. "OCTID: Optical coherence tomography image database," *Computers & Electrical Engineering*, vol. 81, p. 106532, Jan. 2020, doi: 10.1016/j.compeleceng.2019.106532.
88. Traslational-Visual-Health-Laboratory, *Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET*. (Sept. 30, 2025). Accessed: Oct. 09, 2025. [Online]. Available: https://github.com/Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET
89. "ResNet and its application to medical image processing: Research progress and challenges," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107660, Oct. 2023, doi: 10.1016/j.cmpb.2023.107660.
90. M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, May 2019, pp. 6105–6114.
91. N. Agastya, L. Novamizanti, and G. Budiman, "Tuna Loin Quality Grading Using Image Processing and EfficientNetV2," in *2025 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, Bali, Indonesia: IEEE, July 2025, pp. 833–840, doi: 10.1109/IAICT65714.2025.11100818.
92. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International conference on learning representations*, 2017.
93. K. Alomar, H. I. Aysel, and X. Cai, "Data Augmentation in Classification and Segmentation: A Survey and New Strategies," *J Imaging*, vol. 9, no. 2, p. 46, Feb. 2023, doi: 10.3390/jimaging9020046.
94. S. Natarajan, A. Jain, R. Krishnan, A. Rogye, and S. Sivaprasad, "Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone," *JAMA Ophthalmol*, vol. 137, no. 10, pp. 1182–1188, Oct. 2019, doi: 10.1001/jamaophthalmol.2019.2923.
95. R. Hao, K. Namdar, L. Liu, M. A. Haider, and F. Khalvati, "A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks," *J Digit Imaging*, vol. 34, no. 4, pp. 862–876, Aug. 2021, doi: 10.1007/s10278-021-00478-7.
96. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, July 2009, doi: 10.1016/j.ipm.2009.03.002.
97. "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006, doi: 10.1016/j.patrec.2005.10.010.
98. M. Kulyabin et al., "OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods," *Sci Data*, vol. 11, p. 365, Apr. 2024, doi: 10.1038/s41597-024-03182-7.
99. M. K. Bizaki et al., "Deep Neural Networks-based Malignant Breast Lesions Detection and Segmentation from Mammography," in *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Italy: IEEE, Nov. 2022, pp. 1–3, doi: 10.1109/NSS/MIC44845.2022.10399058.
100. S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLoS ONE*, vol. 17, no. 1, p. e0262838, Jan. 2022, doi: 10.1371/journal.pone.0262838.
101. S. Kiss, H. S. Chandwani, A. L. Cole, V. D. Patel, O. E. Lunacsek, and P. U. Dugel, "Comorbidity and health care visit burden in working-age commercially insured patients with diabetic macular edema," *Clin Ophthalmol*, vol. 10, pp. 2443–2453, Dec. 2016, doi: 10.2147/OPTH.S114006.
102. C. Lobo et al., "Characterisation of progression of macular oedema in the initial stages of diabetic retinopathy: a 3-year longitudinal study," *Eye (Lond)*, vol. 37, no. 2, pp. 313–319, Feb. 2023, doi: 10.1038/s41433-022-01937-3.
103. Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," *Mach Learn Knowl Discov Databases*, vol. 8725, pp. 225–239, 2014, doi: 10.1007/978-3-662-44851-9_15.
104. C. Toma et al., "Microvascular changes in eyes with non-proliferative diabetic retinopathy with or without macular microaneurysms: an OCT-angiography study," *Acta Diabetol*, vol. 62, no. 5, pp. 753–761, May 2025, doi: 10.1007/s00592-024-02394-y.
105. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.
106. Y. Bi, J. Xie, and H. Wang, "Contrastive Learning-Based Feature Modulation Strategy for Test-Time Adaptation in Medical Image Segmentation," in *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, May 2025, pp. 916–921, doi: 10.1109/CSCWD64889.2025.11033539.
107. S. Rajaraman, G. Zamzmi, and S. K. Antani, "Novel loss functions for ensemble-based medical image classification," *PLoS One*, vol. 16, no. 12, p. e0261307, Dec. 2021, doi: 10.1371/journal.pone.0261307.
108. R. Viñals and J.-P. Thiran, "A KL Divergence-Based Loss for In Vivo Ultrafast Ultrasound Image Enhancement with Deep Learning," *J. Imaging*, vol. 9, no. 12, p. 256, Nov. 2023, doi: 10.3390/jimaging9120256.

 doi:10.20944/preprints202512.0450.v1

109. X. Lei et al., "A Cross-Modal Feature Fusion Method to Diagnose Macular Fibrosis in Neovascular Age-Related Macular Degeneration," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece: IEEE, May 2024, pp. 1–5, doi: 10.1109/ISBI56570.2024.10635126.
110. A. Zedadra, M. Y. Salah-Salah, O. Zedadra, and A. Guerrieri, "Multi-Modal AI for Multi-Label Retinal Disease Prediction Using OCT and Fundus Images: A Hybrid Approach," *Sensors*, vol. 25, no. 14, p. 4492, July 2025, doi: 10.3390/s25144492.
111. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, June 2016.
112. J. Carse, A. Alvarez Olmo, and S. McKenna, "Calibration of deep medical image classifiers: an empirical comparison using dermatology and histopathology datasets," in *International workshop on uncertainty for safe utilization of machine learning in medical imaging*, Springer, 2022, pp. 89–99.
113. W. Chen and H. Wang, "OCTSharp: an open-source and real-time OCT imaging software based on C#," *Biomed Opt Express*, vol. 14, no. 11, pp. 6060–6071, Oct. 2023, doi: 10.1364/BOE.505308.