

Evaluating Fundus-Specific Foundation Models for Diabetic Macular Edema Detection

Franco Javier Arellano , José Ignacio Orlando

Yatiris Group, PLADEMA Institute, UNICEN, Tandil, Argentina

CONICET, Tandil, Argentina

francoare@pladema.exa.unicen.edu.ar, jiorlando@pladema.exa.unicen.edu.ar

Abstract—Diabetic Macular Edema (DME) is a leading cause of vision loss among patients with Diabetic Retinopathy (DR). While deep learning has shown promising results for automatically detecting this condition from fundus images, its application remains challenging due to the limited availability of annotated data. Foundation Models (FM) have emerged as an alternative solution. However, it is unclear if they can cope with DME detection in particular. In this paper, we systematically compare different FM and standard transfer learning approaches for this task. Specifically, we compare the two most popular FM for retinal images—RETFound and FLAIR—and an EfficientNet-B0 backbone, across different training regimes and evaluation settings in IDRiD, MESSIDOR-2 and OCT-and-Eye-Fundus-Images (OEFI). Results show that despite their scale, FM do not consistently outperform fine-tuned CNNs in this task. In particular, an EfficientNet-B0 ranked first or second in terms of area under the ROC and precision/recall curves in most evaluation settings, with RETFound only showing promising results in OEFI. FLAIR, on the other hand, demonstrated competitive zero-shot performance, achieving notable AUC-PR scores when prompted appropriately. These findings reveal that FM might not be a good tool for fine-grained ophthalmic tasks such as DME detection even after fine-tuning, suggesting that lightweight CNNs remain strong baselines in data-scarce environments.

I. INTRODUCTION

Diabetic macular edema (DME) is one of the most serious vision-threatening conditions linked to diabetic retinopathy (DR) [1]. It occurs when fluid accumulates in the macula, leading to swelling that distorts central sight [2]. Early detection is crucial to initiate treatment promptly and prevent irreversible vision loss [3]. But, as the disease progresses without noticeable symptoms, it may go undetected until patient experiences vision impairment [4], so regular checkups are recommended i.e. through fundus imaging, a time-consuming process that depends on ophthalmologists skills [5].

In recent years, deep learning has shown strong potential for automating medical image analysis [6], including the detection of retinal diseases [7]. However, developing high-performing models for DME detection remains particularly challenging due to the limited availability of annotated data [8]. Foundation models (FMs) have emerged as a promising approach to address this limitation [9]. These models are pre-trained on large general-purpose datasets using pretext tasks—either via self-supervised learning from unlabeled samples [10], [11] or by learning to match images and their associated clinical reports. Such strategies help the models learn meaningful

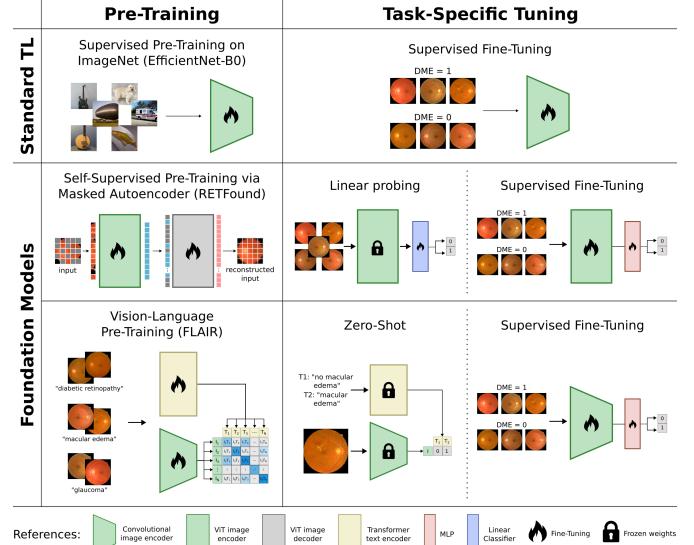


Fig. 1: Schematic representation of our study. We compared the standard transfer learning (TL) approach of supervised task-specific fine-tuning of a CNN pre-trained on ImageNet against linear probing, supervised fine-tuning and zero-shot prediction using Foundation Models (FMs).

representations of the data, serving as a robust base for developing specialized models for downstream tasks through linear probing or fine-tuning on smaller, task-specific datasets [12].

To ensure alignment between the data used during pre-training and that used for fine-tuning, several domain-specific FM trained solely on retinal images have been introduced, such as RETFound [11], FLAIR [13], and others [14]–[16]. RETFound, for example, employs self-supervised learning with masked autoencoders [17] to learn an encoder that can be adapted to other fundus image-specific tasks. FLAIR [13], on the other hand, is trained to align image-text pairs using the CLIP framework [18]. It combines an image encoder based on ResNet-50 [19] with a text encoder based on BioClinicalBERT [20], enabling the development of a robust image encoder for fine-tuning, as well as a zero-shot model [21] capable of predicting outcomes by jointly processing image and text prompts and computing their cosine similarity [13].

Recent studies have explored the use of these models for detecting conditions such as hypertensive retinopathy [22], DR [14], [23], glaucoma [14], [23], and others, demonstrating

promising results. These include either superior performance compared to standard baselines [11], [14], [23] or improved data efficiency for downstream tasks [11]. However, the effectiveness of these models as a starting point for DME detection remains unexplored. In this study, we address this gap by evaluating the performance of FMs for DME detection under limited-data conditions [9] (Fig. 1). Specifically, we evaluate two of the most popular FMs—RETFound and FLAIR—under linear probing, fine-tuning, and/or zero-shot prediction scenarios, and compare them against an EfficientNet-B0 [24] baseline pre-trained on ImageNet [25] and fine-tuned for this task. Our results show that standard transfer learning remains a strong baseline for DME detection, often outperforming FM-derived models across multiple evaluation settings.

II. METHODS

Our study is schematically presented in Fig. 1. Specifically, we evaluate three methods: standard fine-tuning (Section II-A), linear probing using FMs as fixed feature extractors (Section II-B), and zero-shot prediction with vision-language FMs (Section II-C). Each approach is described in the sequel.

A. Standard Fine-Tuning

In the standard fine-tuning approach, we evaluate three models: EfficientNet-B0, the ResNet-50 encoder from FLAIR, and RETFound, which uses a Vision Transformer (ViT-S) [26]. We chose EfficientNet-B0 pre-trained on ImageNet, as it is an established baseline for transfer learning. We also included the ResNet-50 from FLAIR to study the benefits of prior training on retinal images and text, which should enable more domain-specific representations. Finally, we included RETFound’s ViT to assess the effect of using self-supervised learning.

Each network is adapted for DME detection by supervised fine-tuning (SFT) on retinal fundus images labeled as DME or non-DME. At this phase, all model layers are updated using the training data from each dataset, allowing them to specialize their representations for the DME detection task.

B. Linear Probing

To explore the representational power of FMs, we evaluate linear probing as an alternative to full fine-tuning. In this setup, the weights of the backbone model remain frozen, and only a lightweight linear classifier is trained on top of the extracted features. We use RETFound as fixed feature extractor and train two classifiers—standard Ridge Regression and LASSO [27]—on its embeddings, to distinguish between DME and non-DME cases. To mitigate the effect of the so-called curse of dimensionality, we also trained these classifiers using dimensionality reduction via Principal Component Analysis (PCA).

C. Zero-shot prediction

We also used FLAIR as a zero-shot classifier, leveraging its abilities to align visual and textual representations [13], [18]. In this setup, no fine-tuning are performed on networks’ parameters. Instead, the model receives a fundus image along with predefined text prompts describing both presence/absence

of DME-related findings. Each image is encoded by the ResNet-50 image encoder, and each prompt is processed by the BioClinicalBERT text encoder. Classification is then performed by computing the cosine similarity between image and text embeddings, and the predicted class is assigned based on the prompt with the highest similarity. No systematic prompt engineering was applied for evaluation. Instead, we evaluated the positive prompts describing DME-related findings that were originally used for training FLAIR [13] (Table I), and constructed corresponding negative prompts by negating them.

III. EXPERIMENTAL SETUP

A. Materials

Empirical evaluation was performed using three public datasets: the popular MESSIDOR-2 [28], [29] and IDRiD [30] sets, widely applied for this task [31]–[33], and the recently introduced OCT-and-Eye-Fundus-Images (OEFI) set [34].

MESSIDOR-2 contains 1740 images labeled for DME as either negative (0) or positive (1) (1589 vs. 151, respectively). As no fixed training, validation and test partitions are provided, we randomly divided them using 70% and 30% for training and test, respectively, pulling off 10% of the training set for validation. Stratified sampling at a patient level was used to ensure similar distribution between classes.

IDRiD, on the other hand, consists of 516 images, annotated with three DME grades (0—no DME—, 1—non-clinically significant DME—, or 2—DME—; 222, 51 and 243 images, respectively). As our goal is to identify any DME presence, we merged labels 1 and 2 into a single positive class, resulting in a binary classification task. We followed the same partitions into training and test as provided in the set, while extracted 10% of the training samples for validation.

Finally, OEFI includes 1548 eye fundus and 1113 OCT images acquired from multiple ophthalmological institutions in Mexico, all with binary DME labels. We only used the fundus images (1053 and 495 with and without DME, respectively), as an external set to study models’ generalization performance.

B. Evaluation metrics

All models were evaluated using the Area Under the ROC (AUC-ROC) and Precision-Recall (AUC-PR) curves, two standard metrics in the literature for DME detection [11], [14], [35]. Given the noticeable class imbalance in some of the evaluation datasets—such as in MESSIDOR-2 (8.7% positive cases) and OEFI (32.0%), while IDRiD is more balanced (57.0% positive)—, we used AUC-PR on the validation set to choose the best configuration of each model.

C. Training configuration

Supervised fine-tuning was performed using Adam optimization, with learning rates empirically chosen per model and dataset. A custom version of RandAugment [36] was used for data augmentation. A grid search procedure was applied to fix the number of transformations (from 1 to 7) and augmentation strength (from 0.2 to 1.0, with increments of 0.2), choosing

TABLE I: AUC-PR/ROC results for different models trained and tested for DME detection. **First**, **second** and **third** ranked values are formatted accordingly. Values in brackets are 95% confidence intervals obtained with bootstrapping ($n = 1000$).

Training set	Model	Architecture (parameters)	Setting	AUC-PR (95% CI)			AUC-ROC (95% CI)		
				IDRID	MESSIDOR-2	OEFI	IDRID	MESSIDOR-2	OEFI
IDRID ($N = 371$)	CNN	EfficientNet-B0 (5.3M)	SFT	0.959 [0.92 - 0.98]	0.577 [0.41 - 0.74]	0.940 [0.92 - 0.95]	0.936 [0.88 - 0.98]	0.919 [0.87 - 0.96]	0.902 [0.89 - 0.92]
	RETFound	ViT-S (307M)	SFT	0.921 [0.87 - 0.96]	0.453 [0.30 - 0.59]	0.965 [0.95 - 0.97]	0.887 [0.82 - 0.94]	0.717 [0.61 - 0.81]	0.939 [0.93 - 0.95]
	RETFound	ViT-S (307M)	LP (Ridge)	0.734 [0.62 - 0.84]	0.195 [0.12 - 0.32]	0.559 [0.53 - 0.59]	0.652 [0.54 - 0.75]	0.656 [0.55 - 0.75]	0.240 [0.22 - 0.26]
	RETFound	ViT-S (307M)	LP (PCA + Ridge)	0.714 [0.59 - 0.82]	0.170 [0.09 - 0.27]	0.681 [0.65 - 0.72]	0.624 [0.51 - 0.73]	0.629 [0.54 - 0.72]	0.511 [0.48 - 0.54]
	RETFound	ViT-S (307M)	LP (LASSO)	0.727 [0.62 - 0.83]	0.218 [0.12 - 0.34]	0.665 [0.63 - 0.70]	0.626 [0.51 - 0.73]	0.653 [0.56 - 0.75]	0.501 [0.47 - 0.53]
	RETFound	ViT-S (307M)	LP (PCA + LASSO)	0.754 [0.64 - 0.85]	0.238 [0.15 - 0.38]	0.721 [0.69 - 0.75]	0.661 [0.55 - 0.76]	0.710 [0.61 - 0.80]	0.562 [0.53 - 0.59]
	FLAIR	RN-50 (26M)	SFT	0.925 [0.87 - 0.97]	0.598 [0.47 - 0.73]	0.850 [0.83 - 0.87]	0.879 [0.81 - 0.94]	0.831 [0.74 - 0.91]	0.749 [0.72 - 0.78]
MESSIDOR-2 ($N = 1096$)	CNN	EfficientNet-B0 (5.3M)	SFT	0.949 [0.90 - 0.98]	0.792 [0.68 - 0.88]	0.933 [0.91 - 0.95]	0.916 [0.85 - 0.96]	0.959 [0.93 - 0.98]	0.918 [0.90 - 0.94]
	RETFound	ViT-S (307M)	SFT	0.872 [0.80 - 0.93]	0.623 [0.48 - 0.74]	0.917 [0.90 - 0.93]	0.805 [0.72 - 0.88]	0.886 [0.83 - 0.94]	0.856 [0.84 - 0.87]
	RETFound	ViT-S (307M)	LP (Ridge)	0.729 [0.61 - 0.83]	0.250 [0.16 - 0.38]	0.554 [0.53 - 0.58]	0.636 [0.53 - 0.74]	0.784 [0.71 - 0.85]	0.315 [0.28 - 0.35]
	RETFound	ViT-S (307M)	LP (PCA + Ridge)	0.726 [0.60 - 0.83]	0.203 [0.13 - 0.33]	0.845 [0.82 - 0.87]	0.628 [0.51 - 0.73]	0.736 [0.65 - 0.81]	0.756 [0.73 - 0.78]
	RETFound	ViT-S (307M)	LP (LASSO)	0.697 [0.57 - 0.81]	0.260 [0.16 - 0.39]	0.574 [0.55 - 0.60]	0.622 [0.51 - 0.72]	0.786 [0.72 - 0.85]	0.370 [0.34 - 0.40]
	RETFound	ViT-S (307M)	LP (PCA + LASSO)	0.725 [0.60 - 0.83]	0.204 [0.13 - 0.33]	0.843 [0.82 - 0.87]	0.627 [0.51 - 0.73]	0.736 [0.65 - 0.81]	0.755 [0.73 - 0.78]
	FLAIR	RN-50 (26M)	SFT	0.916 [0.85 - 0.96]	0.703 [0.56 - 0.81]	0.941 [0.93 - 0.95]	0.878 [0.80 - 0.94]	0.908 [0.85 - 0.95]	0.897 [0.88 - 0.91]
Zero-shot ($N = 0$)	FLAIR	RN-50 (26M) / BCB (110M)	Prompt 1	0.925 [0.86 - 0.97]	0.565 [0.42 - 0.70]	0.918 [0.89 - 0.94]	0.907 [0.85 - 0.96]	0.936 [0.91 - 0.96]	0.907 [0.89 - 0.93]
	FLAIR	RN-50 (26M) / BCB (110M)	Prompt 2	0.898 [0.80 - 0.97]	0.525 [0.38 - 0.68]	0.945 [0.93 - 0.96]	0.901 [0.84 - 0.96]	0.929 [0.90 - 0.95]	0.935 [0.92 - 0.95]
	FLAIR	RN-50 (26M) / BCB (110M)	Prompt 3	0.922 [0.86 - 0.97]	0.546 [0.40 - 0.69]	0.861 [0.83 - 0.89]	0.909 [0.85 - 0.96]	0.932 [0.90 - 0.95]	0.792 [0.77 - 0.82]
	FLAIR	RN-50 (26M) / BCB (110M)	Prompt 4	0.879 [0.78 - 0.96]	0.503 [0.37 - 0.68]	0.896 [0.87 - 0.92]	0.884 [0.81 - 0.95]	0.931 [0.90 - 0.95]	0.868 [0.85 - 0.89]

RN-50: ResNet-50. *BCB*: BioClinicalBERT. *SFT*: Supervised Fine-Tuning. *LP*: Linear Probing. *Prompt 1*: "macular edema". *Prompt 2*: "leakage of fluid within the central macula from microaneurysms". *Prompt 3*: "presence of exudates". *Prompt 4*: "presence of exudates within the radius of one disc diameter from the macula center".

operations from a pool that included adjustments to brightness, contrast, saturation, and hue, random rotation, scaling, horizontal flipping, and Gaussian blur. Parameter ranges were empirically chosen to preserve clinical plausibility.

For linear probing, we trained both Ridge and LASSO regression models on fixed feature vectors obtained using the RETFound backbone. Each method was evaluated both with and without dimensionality reduction using PCA, retaining 99% of the variance in the feature space when applied. The regularization parameter α was selected via grid search. For Ridge regression, we explored a broad range of α values: [0.001, 0.01, 0.1, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0, 200.0], which allowed us to assess the effect of both weak and strong regularization. In contrast, for LASSO regression, we used a narrower and lower range: [0.0001, 0.001, 0.01, 0.1, 0.5, 1.0], as LASSO tends to drive coefficients to zero more aggressively and can underfit with higher regularization strengths [27].

IV. RESULTS

All quantitative results are summarized in Table I.

For models trained on IDRID, EfficientNet-B0 achieves the highest AUC-PR and AUC-ROC on the IDRID test set. Among the FMs, FLAIR after SFT is the most competitive in terms of AUC-PR, followed closely by the fine-tuned RETFound, with the reverse ranking when using AUC-ROC as the evaluation metric. A similar trend is observed on OEFI, although in this case the fine-tuned RETFound reports the highest AUC-PR and AUC-ROC values, followed by EfficientNet-B0. Surprisingly, when these models are evaluated on MESSIDOR-2, a noticeable drop in performance is observed, with all approaches reporting AUC-PR values below 0.6. This drop is also reflected in AUC-ROC values, though less prominently, with EfficientNet-B0 remaining the most accurate model.

When models are trained on MESSIDOR-2, EfficientNet-B0 again achieves the highest AUC-PR and AUC-ROC on both MESSIDOR-2 and IDRID test sets. The second-best performing model is the fine-tuned FLAIR, although with

a larger performance gap than observed when training on IDRID. The fine-tuned RETFound consistently ranks third by a significant margin. A similar pattern is observed when evaluated on OEFI, although in this case the fine-tuned FLAIR reports slightly higher AUC-PR values than EfficientNet-B0.

All linear probing strategies using RETFound embeddings—including Ridge, LASSO, and their PCA variants—yield very low AUC-PR and AUC-ROC values compared to fully supervised fine-tuning, regardless of the training set. The use of PCA for dimensionality reduction yields mixed results; while it consistently improves performance on the OEFI test set, its effect on IDRID and MESSIDOR-2 varies: it consistently degrades performance when used with Ridge regression, yet often improves metrics when paired with LASSO.

In the zero-shot setting with FLAIR, the best AUC-PR on the IDRID test set is achieved using Prompt 1. However, when evaluated using AUC-ROC, Prompt 3 yields the highest score. Conversely, on MESSIDOR-2, Prompt 1 convincingly performed best across both metrics. On the OEFI test set, the best performance is also consistently achieved with Prompt 2, which reports the highest AUC-PR and AUC-ROC values.

When comparing zero-shot performance of FLAIR against the fine-tuned version of its ResNet-50 image encoder, the zero-shot model obtains comparable or higher AUC-ROC values across all test sets for most of the prompts. Fine-tuning only ensured better AUC-PR values in MESSIDOR-2, regardless of the training set. In any other case, however, this specialization do not improve the model.

Qualitative results are provided in Fig. 2, using explainability maps such as GradCAMs (for the CNNs) and Gradient Attention Rollout [37] (for RETFound). EfficientNet-B0 focused on the vascular arcades in non-DME eyes when trained on IDRID, whereas the version trained on MESSIDOR-2 highlighted peri-macular regions. In positive cases, the IDRID-trained network concentrated on isolated exudates—even outside the macula—and misclassified an ambiguous MESSIDOR-2 image as healthy (bottom case); the MESSIDOR-2-trained

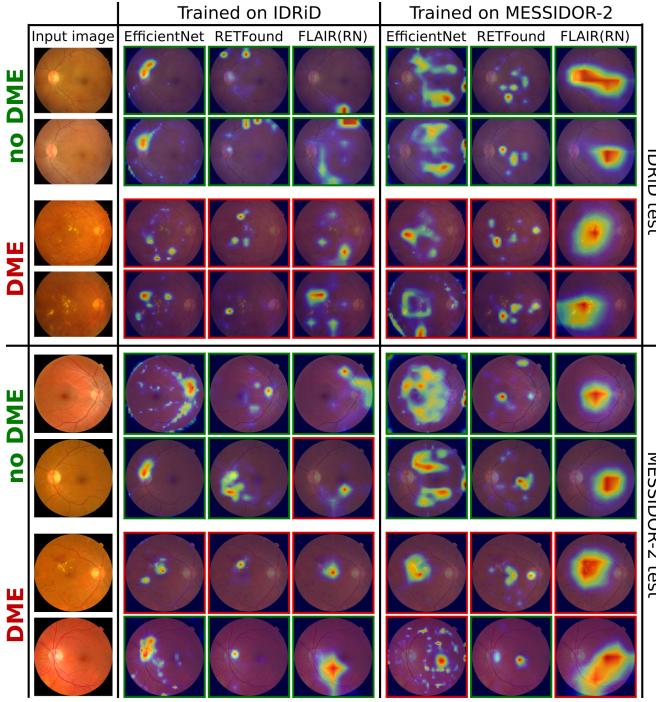


Fig. 2: Qualitative comparison of explainability maps across models and datasets, for randomly selected IDRiD and MESSIDOR test images (top and bottom blocks), with and without DME. Heatmaps for EfficientNet-B0 and FLAIR (ResNet-50) are Grad-CAMs of the predicted class, while RETFound maps are Gradient Attention Rollouts. Border color indicates the model’s predicted label (green = no DME, red = DME).

counterpart centered its attention on the macula and resolved that error. RETFound, on the other hand, displayed less consistent saliency: although it frequently activated around exudates in DME images, it also produced scattered responses in healthy eyes. FLAIR, finally, showed very sparse, pinpoint activations after IDRiD training, offering no clear distinction between healthy and diseased cases, whereas the MESSIDOR-2-trained model generated broader maps consistently centered on the macula, suggesting that it uses lesion presence or absence in that region to guide its predictions.

V. DISCUSSION

In this study we benchmarked two FMs for retinal image analysis—RETFound and FLAIR—for DME detection, comparing them with the typical approach of fine-tuning a CNN pre-trained on ImageNet. We selected these two due to their increased popularity and usage in the field [38], [39].

The prevailing view in the literature is that high-capacity FMs outperform lighter models on downstream tasks because pre-training endows them with rich, transferable representations requiring little additional refinement. Our results challenge this assumption. A lightweight CNN such as EfficientNet-B0 (with ≈ 5 M parameters) pre-trained on ImageNet and fine-tuned on fundus images consistently matched—or even surpassed—the performance of much larger FM backbones such as the ViT-S encoder of RETFound (≈ 307 M

parameters) and the ResNet-50 image encoder of FLAIR (≈ 26 M parameters). We hypothesize that DME detection depends on fine-grained, local cues (e.g., exudates near the macula or subtle vessel curvature) that broad, heterogeneous FM objectives may overlook, whereas smaller CNNs preserve inductive biases better suited to low-level retinal features. Qualitative heatmaps in Fig. 2 support this notion: EfficientNet-B0 produces clinically meaningful activations, and FLAIR benefits from similar CNN-based inductive biases, while RETFound is noticeably less interpretable from a clinical perspective.

Our experiments also show that linear probing on RETFound embeddings yields markedly lower performance than full fine-tuning, regardless of dataset or classifier, indicating that successful transfer to specialized tasks such as DME detection requires deeper adaptation—updating internal representations, not merely the classifier head. This gap likely arises because RETFound’s self-supervised pretraining captures broad retinal structures rather than the subtle, localized biomarkers of DME, making the representations insufficiently discriminative for linear probing without deeper adaptation.

FLAIR remains competitive in a zero-shot setting without task-specific training. Surprisingly, fine-tuning for DME detection does not provide consistent gains and can even degrade accuracy relative to the original zero-shot model, likely because the positive prompts we used were already present during pre-training [13]. While this shows the potential of language supervision, our results also reveal a strong dependence on prompt wording: as no single prompt guaranteed stable performance across datasets. Further work is needed to clarify how prompt phrasing (e.g. prompt 4) and dataset characteristics (e.g., prevalence of exudates near the optic disc within the positive class) interact to affect zero-shot accuracy.

Beyond these model-specific findings, we consistently observed lower performance on MESSIDOR-2 compared to IDRiD. This drop is likely driven by domain shift between the datasets: IDRiD images are high-resolution, captured with dilated pupils and systematically annotated for exudates near the macula, whereas MESSIDOR-2 comprises routine clinical acquisitions with lower resolution, variable illumination, and a much lower prevalence of DME cases. These discrepancies in acquisition protocols, image quality, and disease distribution make cross-dataset generalization particularly challenging, explaining the reduced transferability observed across all models.

To conclude, our study shows that FMs are not a one-size-fits-all solution for DME detection. Although they offer clear advantages—particularly for zero-shot inference via prompting—lightweight CNNs remain robust and efficient baselines, probably because DME detection is an inherently fine-grained task that is benefited by the inductive bias of CNNs, particularly under limited data regimes. We encourage future work to compare these results against new FMs or even large, generalistic [40], [41] or specialized VLMs [42].

ACKNOWLEDGMENTS

This study was partially funded with a CONICET PIP 2021-2023 (11220200102472CO).

REFERENCES

- [1] A. N. Kollias and M. W. Ulbig, "Diabetic retinopathy: early diagnosis and effective treatment," *Deutsches Arzteblatt International*, vol. 107, no. 5, p. 75, 2010.
- [2] P. Kohli, K. Tripathy, and B. Patel, "Macular edema," *StatPearls*, 2024.
- [3] N. Elyasi and H. Hemmati, "Diabetic macular edema: diagnosis and management," *Am Acad Ophthalmol EyeNet Mag*, vol. 66, pp. 35–7, 2021.
- [4] O. Musat, C. Cernat, M. Labib, A. Gheorghe, O. Toma, M. Zamfir, and A. M. Boureanu, "Diabetic macular edema," *Romanian journal of ophthalmology*, vol. 59, no. 3, p. 133, 2015.
- [5] Y. Xu, Y. Wang, B. Liu, L. Tang, L. Lv, X. Ke, S. Ling, L. Lu, and H. Zou, "The diagnostic accuracy of an intelligent and automated fundus disease image assessment system with lesion quantitative function (SmartEye) in diabetic patients," *BMC Ophthalmology*, vol. 19, no. 1, p. 184, 2019.
- [6] S. K. Zhou, H. Greenspan, C. Davatzikos *et al.*, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, 2021.
- [7] M. Badar, M. Haris, and A. Fatima, "Application of deep learning for retinal image analysis: A review," *Computer Science Review*, 2020.
- [8] R. Gelman, "Evaluation of transfer learning for classification of: (1) diabetic retinopathy by digital fundus photography and (2) diabetic macular edema, choroidal neovascularization and drusen by optical coherence tomography," 2019.
- [9] V. van Veldhuizen, V. Botha, C. Lu, M. E. Cesur, K. G. Lipman, E. D. de Jong, H. Horlings, C. I. Sanchez, C. G. M. Snoek, L. Wessels, R. Mann, E. Marcus, and J. Teuwen, "Foundation models in medical imaging – a review and outlook," 2025. [Online]. Available: <https://arxiv.org/abs/2506.09095>
- [10] R. Balestrieri, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian *et al.*, "A cookbook of self-supervised learning," *arXiv preprint arXiv:2304.12210*, 2023.
- [11] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, no. 7981, pp. 156–163, 2023.
- [12] D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen *et al.*, "A real-world dataset and benchmark for foundation model adaptation in medical image classification," *Scientific Data*, vol. 10, no. 1, p. 574, 2023.
- [13] J. Silva-Rodriguez, H. Chakor, R. Kobbi, J. Dolz, and I. B. Ayed, "A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision," *Medical Image Analysis*, vol. 99, p. 103357, 2025.
- [14] J. Du, J. Guo, W. Zhang, S. Yang, H. Liu, H. Li, and N. Wang, "RET-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024.
- [15] K. Yu, Y. Zhou, Y. Bai, Z. D. Soh, X. Xu, R. S. M. Goh, C.-Y. Cheng, and Y. Liu, "UrFound: Towards universal retinal foundation models via knowledge-guided masked modeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 753–762.
- [16] S. Yang, J. Du, J. Guo, W. Zhang, H. Liu, H. Li, and N. Wang, "ViLReF: an expert knowledge enabled vision-language retinal foundation model," *arXiv preprint arXiv:2408.10894*, 2024.
- [17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.
- [18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: <https://arxiv.org/abs/1512.03385>
- [20] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019. [Online]. Available: <https://arxiv.org/abs/1904.03323>
- [21] J. Liu, T. Hu, Y. Zhang, X. Gai, Y. Feng, and Z. Liu, "A ChatGPT aided explainable framework for zero-shot medical image diagnosis," *arXiv preprint arXiv:2307.01981*, 2023.
- [22] J. Silva-Rodriguez, J. Chelbi, W. Kabir, H. Chakor, J. Dolz, I. B. Ayed, and R. Kobbi, "Exploring the transferability of a foundation model for fundus images: Application to hypertensive retinopathy," in *Computer Graphics International Conference*. Springer, 2023, pp. 427–437.
- [23] D. Shi, W. Zhang, X. Chen, Y. Liu, J. Yang, S. Huang, Y. C. Tham, Y. Zheng, and M. He, "EyeFound: A multimodal generalist foundation model for ophthalmic imaging," 2024.
- [24] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [27] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.
- [28] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, J.-R. Ordóñez-Varela, P. Massin, A. Erginay *et al.*, "Feedback on a publicly distributed image database: the MESSIDOR database," *Image Analysis & Stereology*, pp. 231–234, 2014.
- [29] M. D. Abramoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang *et al.*, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA ophthalmology*, vol. 131, no. 3, pp. 351–357, 2013.
- [30] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018.
- [31] S. Sundaram, M. Selvamani, S. K. Raju, S. Ramaswamy, S. Islam, J.-H. Cha, N. A. Almjally, and A. Elaraby, "Diabetic retinopathy and diabetic macular edema detection using ensemble based convolutional neural networks," *Diagnostics*, vol. 13, no. 5, p. 1001, 2023.
- [32] T.-Y. Wang, Y.-H. Chen, J.-T. Chen, J.-T. Liu, P.-Y. Wu, S.-Y. Chang, Y.-W. Lee, K.-C. Su, and C.-L. Chen, "Diabetic macular edema detection using end-to-end deep fusion model and anatomical landmark visualization on an edge computing device," *Frontiers in medicine*, vol. 9, p. 851644, 2022.
- [33] T. Nazir, M. Nawaz, J. Rashid, R. Mahum, M. Masood, A. Mehmood, F. Ali, J. Kim, H.-Y. Kwon, and A. Hussain, "Detection of diabetic eye disease from retinal images using a deep learning based centernet model," *Sensors*, vol. 21, no. 16, p. 5283, 2021.
- [34] J. A. Hughes Cano, U. Olivares Pinto, and S. C. Thébault, "Dataset of eye fundus and OCT images for the study of diabetic macular edema and diabetic retinopathy," <https://github.com/Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET>, 2022, accessed: 2025-07-08.
- [35] I. Bressler, R. Aviv, D. Margalit, G. Y. Cohen, T. Ianchulev, S. V. Savant, D. J. Ramsey, and Z. Dvey-Aharon, "Autonomous screening for diabetic macular edema using deep learning processing of retinal images," *medRxiv*, pp. 2022–08, 2022.
- [36] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.
- [37] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.0928*, 2020.
- [38] J. Zhang, S. Lin, T. Cheng, Y. Xu, L. Lu, J. He, T. Yu, Y. Peng, Y. Zhang, H. Zou *et al.*, "RETFound-enhanced community-based fundus disease screening: real-world evidence and decision curve analysis," *NPJ digital medicine*, vol. 7, no. 1, p. 108, 2024.
- [39] K. Sun, S. Xue, F. Sun, H. Sun, Y. Luo, L. Wang, S. Wang, N. Guo, L. Liu, T. Zhao *et al.*, "Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions," *Artificial Intelligence in Medicine*, p. 103265, 2025.
- [40] K. Tomita, T. Nishida, Y. Kitaguchi, K. Kitazawa, and M. Miyake, "Image recognition performance of GPT-4V (ision) and GPT-4o in ophthalmology: Use of images in clinical questions," *Clinical Ophthalmology*, pp. 1557–1564, 2025.
- [41] X. Liang, M. Bian, M. Chen, L. Liu, J. He, J. Xu, and L. Li, "A novel ophthalmic benchmark for evaluating multimodal large language models with fundus photographs and OCT images," *arXiv preprint arXiv:2503.07094*, 2025.
- [42] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau *et al.*, "MedGemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.