Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliable Retinal Screening with Foundation Models under Distribution Shift

Amreen Batool

amreen@stu.jejunu.ac.kr

Jeju National University

Faheem Ahmed

Jeju National University

Fakhri Alam Khan

King Fahd University of Petroleum and Minerals

Article

Keywords: retinal disease classification, retinal foundation models, out-of-distribution detec tion, model calibration, open-world retinal screening, uncertainty estimation

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11262974/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reliable Retinal Screening with Foundation Models under Distribution Shift

Amreen Batool[1], Faheem Ahmed[2], Fakhri Alam Khan[3,*]

[1]Jeju National University, Jeju, Republic of Korea

[2]Biologics4U Research Institute, Seoul, Republic of Korea

[3]King Fahd University of Petroleum and Minerals, Saudi Arabia

*Corresponding author: amreen@stu.jejunu.ac.kr

Abstract

Reliable deployment of deep learning classifiers for retinal disease screening requires accurate in-distribution classification, effective out-of-distribution (OOD) detection, and well-calibrated predictions under distribution shift. This study investigates the open-world reliability of the EyeCLIP foundation model and evaluates whether calibration-oriented fine-tuning improves performance under external-domain shift. A three-class fundus classification task comprising diabetic retinopathy, glaucoma, and normal images was constructed using APTOS2019, G1020, and ORIGA, with ORIGA held out entirely for external evaluation. Standard balanced fine-tuning was compared with calibration-oriented fine-tuning and a frozen-feature linear probe. Five OOD scoring methods were evaluated alongside post-hoc temperature scaling and accept–reject analysis. Performance was assessed using classification, OOD detection, and calibration metrics. Standard balanced fine-tuning outperformed calibration-oriented fine-tuning in in-distribution accuracy (0.928 vs. 0.781), k-nearest-neighbor OOD detection (AUROC 0.938 vs. 0.878), and post-hoc OOD calibration (ECE 0.081 vs. 0.450). The calibration-oriented model misclassified 69% of normal ORIGA images as glaucoma. At 25% overall rejection, standard balanced fine-tuning rejected 50.9% of OOD samples while falsely rejecting 1.6% of in-distribution samples. Standard balanced fine-tuning combined with post-hoc calibration and feature-space k-nearest-neighbor abstention provides a simple and effective strategy for improving the reliability of open-world retinal screening.

Keywords: retinal disease classification; retinal foundation models; out-of-distribution detection; model calibration; open-world retinal screening; uncertainty estimation.

1 Introduction

Diabetic retinopathy (DR) and glaucoma are among the leading causes of irreversible vision loss worldwide, making early diagnosis essential for preventing avoidable blindness. The global prevalence of diabetic retinopathy alone is estimated to exceed 100 million individuals and is projected to increase substantially over the coming decades [1]. Color fundus photography has therefore become the primary imaging modality for large-scale retinal screening because it provides a rapid, non-invasive assessment of retinal abnormalities. Recent advances in deep learning have enabled automated retinal screening systems that achieve expert-level diagnostic performance under controlled clinical conditions, with systems such as IDx-DR demonstrating the clinical potential of artificial intelligence (AI) for ophthalmic screening [2]. Despite these advances, reliable deployment in routine clinical practice remains a significant challenge.

The primary obstacle to clinical deployment is distribution shift. Fundus images acquired in different hospitals frequently differ because of variations in imaging devices, acquisition protocols, illumination conditions, patient demographics, disease prevalence, and image quality [3]. Consequently, images encountered during deployment often differ from those observed during model training. Although deep learning models may maintain high confidence under these conditions, such confidence does not necessarily indicate reliable predictions. In safety-critical applications such as retinal screening, confidently misclassifying unfamiliar images may delay specialist referral and adversely affect patient care. Therefore, trustworthy deployment requires models that not only achieve high classification accuracy but also recognize previously unseen data and abstain from unreliable predictions. This capability, commonly referred to as out-of-distribution (OOD) detection or open-world recognition, is increasingly regarded as an essential requirement for safe clinical AI systems [4].

Foundation models have recently emerged as powerful pretrained visual encoders for retinal image analysis. Models such as EyeCLIP [5], which adapts the CLIP vision-language framework [6] to ophthalmology, and RETFound [7], which learns retinal representations through self-supervised pretraining, have demonstrated strong transferability across a wide range of downstream ophthalmic tasks. Existing studies have primarily focused on improving in-distribution classification performance through fine-tuning of these pretrained models. However, considerably less attention has been devoted to understanding their reliability under realistic deployment scenarios involving distribution shift.

To improve prediction confidence and mitigate class imbalance, many fine-tuning pipelines incorporate calibration-oriented training strategies such as label smoothing, cosine-similarity classifier heads, class-reweighted losses, and weighted sampling [8–10]. These techniques are widely adopted because they often improve calibration and minority-class recognition under conventional supervised learning settings. Nevertheless, whether these strategies also improve the open-world reliability of retinal foundation models remains largely unknown. Moreover, their interaction with post-hoc calibration and modern OOD detection methods has not been

systematically investigated under clinically relevant deployment conditions. As a result, it remains unclear whether improved in-distribution calibration necessarily translates into safer and more reliable deployment under distribution shift.

To address this question, we conduct a controlled open-world retinal disease classification study using the EyeCLIP foundation model. A three-class classification task comprising diabetic retinopathy, glaucoma, and normal retinal images is constructed using the APTOS2019, G1020, and ORIGA datasets [11–13]. To simulate realistic deployment, the ORIGA dataset is excluded entirely during training and reserved exclusively for open-world evaluation. We compare a conventional balanced fine-tuning strategy with a calibration-oriented alternative and evaluate their performance using multiple OOD detection methods, calibration metrics, and deployment-oriented accept–reject analysis. Contrary to our initial expectation that calibration-oriented training improves deployment robustness, our results show that the simpler balanced fine-tuning strategy consistently provides superior classification performance, more reliable OOD detection, and improved probability calibration under distribution shift.

The main contributions of this work are summarized as follows:

- Systematically investigate the influence of calibration-oriented fine-tuning on the open-world reliability of retinal foundation models under realistic distribution shifts.
- Demonstrate that a conventional balanced fine-tuning strategy consistently outperforms a calibration-oriented alternative across classification performance, probability calibration, OOD detection, and deployment-oriented accept–reject analysis.
- We compare five representative OOD detection methods and show that feature-space k-nearest-neighbor scoring consistently provides the most reliable identification of previously unseen retinal images.
- Provide practical evidence supporting a deployment strategy that combines balanced fine-tuning, post-hoc temperature scaling, and feature-space OOD detection for reliable open-world retinal screening.

2 Related Work

This section reviews four lines of work relevant to open-world retinal screening: ophthalmic foundation models, out-of-distribution detection, calibration methods, and class-imbalanced classification. Table 1 summarizes how each line of work has been evaluated and what gaps remain for safe clinical deployment.

2.1 Retinal and Ophthalmic Foundation Models

Foundation models pretrained on large-scale retinal data have become the dominant paradigm for fundus image analysis. RETFound [7] introduced self-supervised masked image modeling on over 1.6 million retinal images and demonstrated strong transfer across diabetic retinopathy, glaucoma, and systemic disease prediction tasks. EyeCLIP [5] adapted the CLIP vision-language framework [6] to ophthalmology by aligning paired fundus images and clinical text, yielding a visual encoder that supports both linear probing and full fine-tuning on downstream tasks. More recent ophthalmic foundation models, including FLAIR [14] and VisionFM [15], extend this paradigm with broader pretraining corpora and multi-task heads. Despite their strong in-distribution performance, evaluations of these models have focused predominantly on closed-set classification accuracy. Their behavior under dataset shift, the interaction between fine-tuning choices and downstream uncertainty estimation, and the integration of abstention mechanisms into deployment pipelines remain largely unexplored. This study addresses these gaps by evaluating EyeCLIP fine-tuning strategies through the combined lens of classification accuracy, OOD detection, calibration, and accept-reject deployment behavior.

2.2 Out-of-Distribution Detection

OOD detection methods are commonly grouped by the representation on which they operate. Output-space methods score uncertainty from the classifier's softmax or logit outputs. The Maximum Softmax Probability (MSP) [16] is the canonical baseline, flagging low-confidence predictions as potentially out-of-distribution. ODIN [17] improves MSP through input perturbation and temperature scaling, while energy-based scoring [18] exploits logit magnitudes through the negative log-sum-exp, providing a theoretically motivated alternative to MSP. Subsequent refinements such as GradNorm [19], ReAct [20], ASH [21], and ViM [22] manipulate logits or feature activations to sharpen the separation between ID and OOD inputs. A common assumption underlying output-space scores is that the logit dynamic range carries discriminative information; this assumption interacts strongly with classifier head design. Feature-space methods instead score test samples based on their geometric position relative to the training feature distribution. The Mahalanobis distance with tied covariance [23] models each class as a Gaussian in penultimate-layer feature space and scores OOD inputs by their distance to the nearest class mean. The k-nearest-neighbor (k-NN) approach [24] avoids parametric assumptions by computing the distance to the k-th nearest training feature in the embedding space. Feature-space methods are often reported to outperform output-space scores on near-OOD problems, where ID and OOD samples share the same broad visual domain [24]. However, the geometric assumptions underlying these methods, particularly the Gaussianity assumption of Mahalanobis and the choice of distance metric for L2-normalized features, are not always satisfied by foundation-model representations. OOD detection in medical imaging has received growing attention, with applications spanning chest radiography, histopathology, and derma-

tology [25–27]. However, controlled comparisons of OOD scoring methods within retinal screening pipelines, particularly under the architectural and loss configurations introduced by foundation model fine-tuning, remain scarce. To our knowledge, no prior work has systematically compared output-space and feature-space OOD scores across multiple fine-tuning recipes on a retinal foundation model.

2.3 Probability Calibration and Label Smoothing

Modern neural networks are known to produce overconfident predictions, a phenomenon first documented systematically by Guo et al. [28]. Remedies fall into two broad categories. Post-hoc calibration methods adjust predictions without retraining, with temperature scaling [28] remaining the strongest single-parameter baseline, and more expressive variants such as Dirichlet calibration [29] offering finer control at the cost of additional fit parameters. Training-time calibration methods modify the loss or training procedure to encourage well-calibrated outputs directly. Label smoothing [30] replaces challenging one-hot targets with a softened distribution, discouraging the model from producing arbitrarily large logits. Focal loss [31] downweights easy examples and has been argued to improve calibration as a side effect. Mixup and related augmentation strategies [32, 33] have similarly been reported to improve calibration through implicit regularization. A growing body of work documents that calibration behavior under distribution shift differs qualitatively from in-distribution calibration. Ovadia et al. [34] showed that post-hoc methods calibrated on validation data often fail to generalize to shifted test distributions. Subsequent studies have shown that training-time interventions, while improving ID calibration, can induce structurally biased predictions on OOD inputs that resist post-hoc correction [35, 36]. The implications of these findings for medical deployment pipelines, where calibration-aware techniques are routinely adopted without OOD evaluation, have not been systematically examined. Our study evaluates this interaction directly by comparing a calibration-aware fine-tuning recipe against a minimal baseline under a controlled domain shift.

2.4 Class-Imbalanced Classification and Cosine Similarity Heads

Class imbalance is a defining characteristic of medical imaging datasets, where pathological classes are typically underrepresented relative to healthy controls. Standard remedies include inverse-frequency class reweighting, square-root or effective-number reweighting [37], weighted sampling [38], and synthetic minority oversampling. Cosine similarity classifier heads [9, 10] were originally developed for face recognition and metric learning, where they enforce a unit-norm constraint on both features and class prototypes. By decoupling class assignment from weight-vector magnitude, cosine heads have been proposed as an architectural remedy for class imbalance: majority classes cannot dominate through magnitude growth alone. They are also widely adopted in few-shot classification [39] and have appeared in medical imaging contexts

Table 1: Summary of related work and remaining gaps addressed in this study.

Study / Method	Main focus	Evaluation setting	Remaining gap addressed here
EyeCLIP [5]	Retinal vision-language foundation model	Downstream retinal classification and transfer learning	Not evaluated for OOD detection, calibration under shift, or accept-reject deployment
RETFound [7]	Self-supervised retinal foundation model	Multiple retinal downstream tasks	Focused on task accuracy rather than abstention or OOD behavior
MSP / ODIN / Energy [16–18]	Output-space OOD detection	General image classification	Sensitive to logit scale and classifier-head design; interaction with foundation model fine-tuning not established
Mahalanobis / k-NN [23, 24]	Feature-space OOD detection	General image classification	Geometric assumptions under L2-normalized foundation features not validated for retinal screening
Temperature scaling [28]	Post-hoc calibration	Mainly ID calibration	OOD calibration under retinal dataset shift remains underexplored
Label smoothing / focal loss [30, 31]	Training-time calibration	ID calibration on general datasets	Interaction with OOD calibration and post-hoc correction not documented
Cosine classifier heads [9, 10]	Feature normalization and class separation	Face recognition, metric learning, few-shot classification	Logit-magnitude OOD scores (energy, max-logit) may be incompatible; not tested in medical context
Present study	Joint evaluation of calibration-aware fine-tuning	Open-world retinal screening	Integrates classification, calibration, OOD detection, and deployment-oriented evaluation

for similar reasons. However, cosine similarity heads impose strong constraints on the logit dynamic range, since the learnable scale parameter bounds logits. This constraint has been studied in the context of representation learning. Still, its downstream consequences for OOD scoring methods that rely on logit magnitudes (energy, max-logit) have not been systematically documented. Similarly, the interaction between label smoothing, cosine normalization, and feature-space OOD scoring is not directly established in the medical imaging literature. Our study contributes a quantitative characterization of these interactions on a retinal screening task and identifies a concrete failure mode of cosine heads for logit-based OOD detection.

2.5 Summary and Position of This Study

Across these four lines of work, three recurring patterns emerge. First, foundation model evaluations have focused on closed-set accuracy rather than open-world behavior. Second, OOD detection methods have been benchmarked predominantly on general-domain datasets, with limited examination of how architectural and loss choices specific to medical imaging affect their reliability. Third, calibration-aware training and class-imbalance handling are routinely adopted as defaults in medical imaging without testing whether they transfer to deployment under distribution shift. Table 1 summarizes these observations and positions the present study with respect to each line of work. Our contribution is a controlled evaluation that connects fine-tuning design choices to four deployment-relevant axes simultaneously: classification accuracy, OOD detection, calibration under shift, and accept-reject behavior.

3 Results

We evaluate the three models from four complementary perspectives: (i) in-distribution (ID) classification performance, (ii) out-of-distribution (OOD) detection on the held-out ORIGA dataset, (iii) probability calibration under distribution shift, and (iv) deployment-oriented accept–reject performance. Finally, qualitative case studies illustrate representative model behavior in clinically relevant scenarios. Unless otherwise stated, all results are reported on the predefined test set (n = 1,353), comprising 703 ID samples from APTOS2019 and G1020 and 650 OOD samples from ORIGA, using the best-validation checkpoint of each model.

3.1 In-distribution classification

We first evaluated closed-set classification performance on the in-distribution (ID) portion of the test set (Table 2, Figure 1a).

Both fine-tuned models substantially outperformed the frozen-feature linear probe, demonstrating the benefit of task-specific adaptation. Among the evaluated models, the standard balanced fine-tuning model achieved the highest overall performance, recording an accuracy of 0.928, a balanced accuracy of 0.923, and a macro-F1 score of 0.920. In comparison, the calibration-oriented fine-tuning model achieved an accuracy of 0.781, a balanced accuracy of 0.842, and a macro-F1 score of 0.708, while the frozen-feature linear probe obtained 0.511, 0.605, and 0.462, respectively.

The per-dataset analysis (Table 3) provides additional insight into model generalization across different data sources. The calibration-oriented fine-tuning model achieved the highest macro-F1 score on APTOS2019 (0.921 versus 0.891). In contrast, the standard balanced fine-tuning model performed better on G1020 (0.380 versus 0.230) and showed a substantially larger advantage on the held-out ORIGA dataset (0.850 versus 0.237). These findings indicate that the

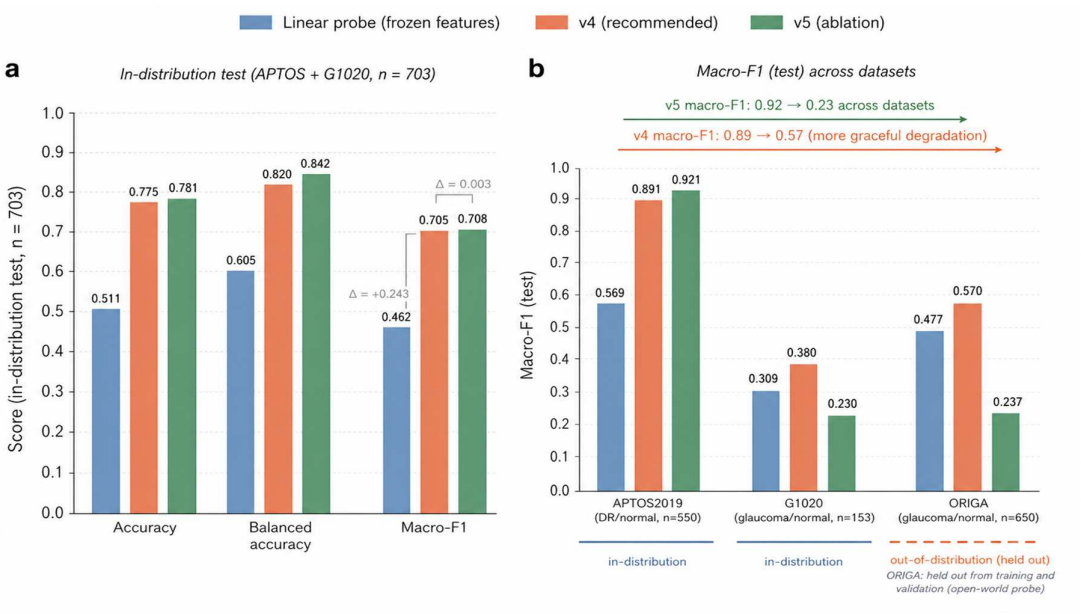


Figure 1: Classification performance across the evaluated models. (a) Overall performance on the in-distribution test set (APTOS2019 + G1020). (b) Per-dataset macro-F1 scores on APTOS2019, G1020, and the held-out ORIGA dataset.

Table 2: In-distribution classification performance on the test set.

Model	Accuracy	Balanced accuracy	Macro F1
Linear probe	0.511	0.605	0.462
Standard balanced fine-tuning (v4)	0.928	0.923	0.920
Calibration-oriented fine-tuning (v5)	0.781	0.842	0.708

performance gain of the calibration-oriented configuration was largely confined to APTOS2019, whereas the standard balanced fine-tuning model generalized more consistently across datasets.

3.2 Out-of-Distribution Detection

We next evaluated each model's ability to distinguish in-distribution (ID) samples from the held-out ORIGA dataset using the five OOD scoring methods. Table 4 summarizes the AUROC obtained for every model–score combination, while Figure 2 provides a visual comparison of the corresponding results.

Across all three models, feature-space k-nearest-neighbor (k-NN) consistently achieved the highest AUROC. The k-NN score reached 0.876 for the frozen-feature linear probe, 0.938 for the standard balanced fine-tuning model (v4), and 0.878 for the calibration-oriented fine-tuning model (v5). For the standard balanced fine-tuning model, k-NN also outperformed Mahalanobis (0.916), max-logit (0.898), energy (0.890), and maximum softmax probability (MSP, 0.814), representing the highest AUROC among all evaluated model–score combinations.

Performance differences were more pronounced for logit-based OOD scores. The calibration-

Table 3: Per-dataset macro-F1 scores on the test split.

Dataset	n	Linear probe	Standard balanced fine-tuning (v4)	Calibration-oriented fine-tuning (v5)
APTOS2019 (DR vs. normal)	550	0.569	0.891	0.921
G1020 (glaucoma vs. normal)	153	0.309	0.380	0.230
ORIGA (glaucoma vs. normal)	650	0.477	0.850	0.237

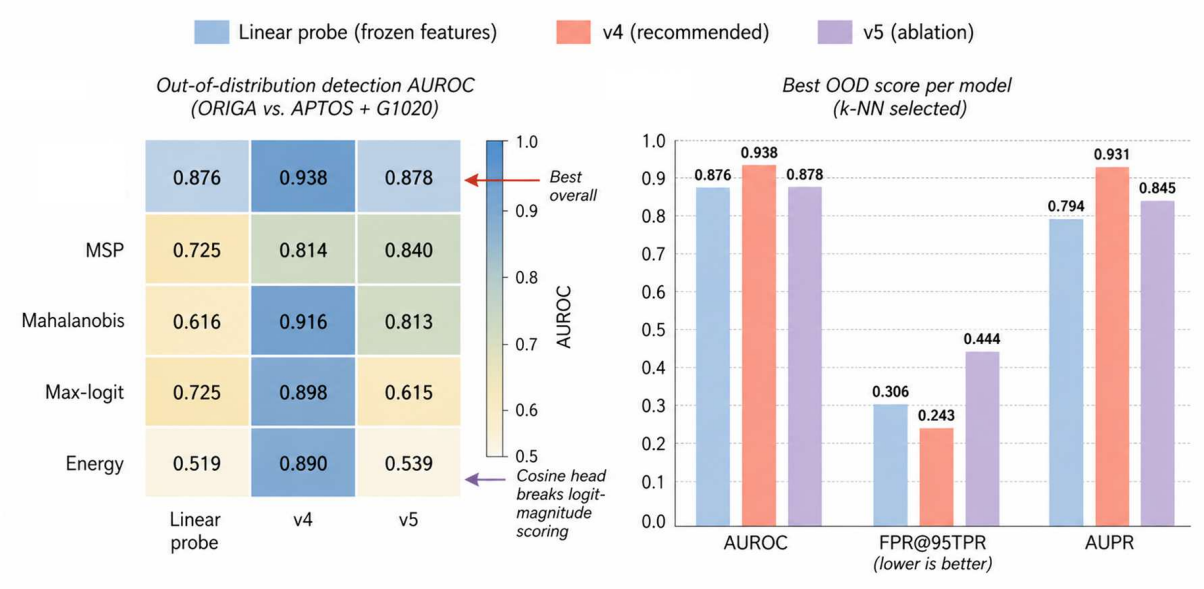


Figure 2: Comparison of OOD detection performance across models and scoring methods. The heatmap reports AUROC values for distinguishing the held-out ORIGA dataset from the in-distribution datasets (APTOS2019 and G1020). The accompanying panel summarizes the best-performing OOD score for each model.

oriented fine-tuning model achieved AUROC values of 0.615 for max-logit scoring and 0.539 for energy scoring, whereas the corresponding scores for the standard balanced fine-tuning model were 0.898 and 0.890, respectively. In contrast, feature-based methods remained competitive for both fine-tuned models, with Mahalanobis scoring achieving AUROC values of 0.916 and 0.813 for the standard balanced and calibration-oriented fine-tuning models, respectively.

The frozen-feature linear probe achieved a k-NN AUROC of 0.876, which was nearly identical to that of the calibration-oriented fine-tuning model (0.878). By comparison, the standard balanced fine-tuning model improved the k-NN AUROC to 0.938, corresponding to an increase of approximately six percentage points over the frozen-feature baseline. Based on these results, k-NN was used as the OOD scoring method in the subsequent calibration and deployment analyses.

Table 4: OOD detection performance (AUROC) across models and scoring methods.

Score	Linear probe	Standard balanced fine-tuning (v4)	Calibration-oriented fine-tuning (v5)
k-NN ($k = 10$)	0.876	0.938	0.878
MSP	0.725	0.814	0.840
Mahalanobis (tied covariance)	0.616	0.916	0.813
Max-logit	0.725	0.898	0.615
Energy ($T = 1$)	0.519	0.890	0.539

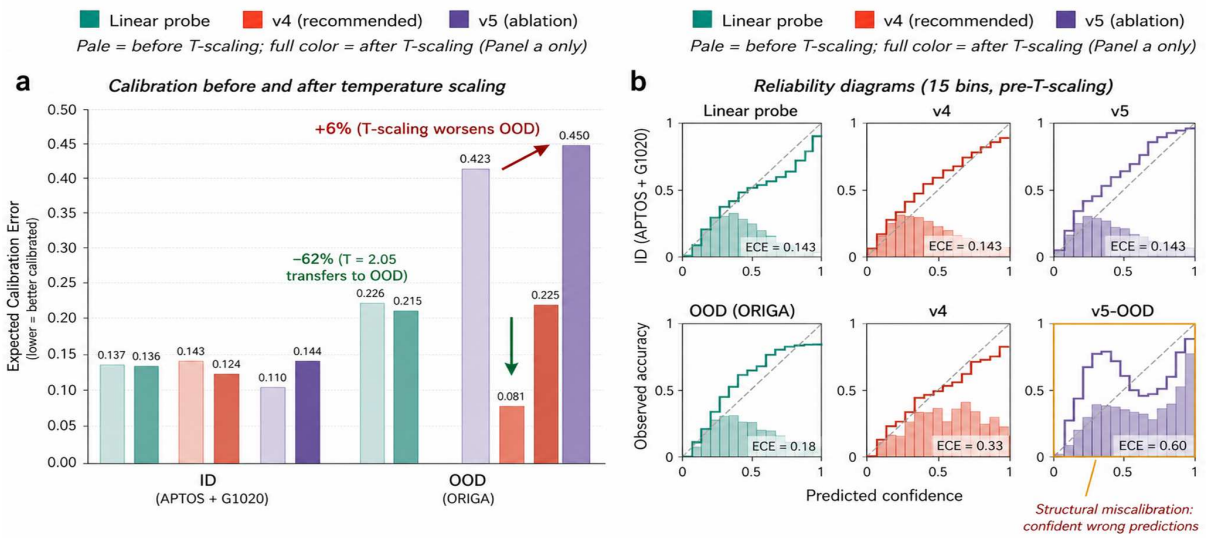


Figure 3: Calibration performance before and after post-hoc temperature scaling. Reliability diagrams are shown separately for the ID and OOD subsets, illustrating the effect of temperature scaling on the evaluated models.

3.3 Calibration Under Distribution Shift

We evaluated probability calibration separately for the in-distribution (ID) and out-of-distribution (OOD) subsets of the test set. The results before and after post-hoc temperature scaling are summarized in Table 5 and illustrated in Figure 3.

Separately evaluating ID and OOD calibration provides a more informative assessment than a single aggregate metric, as calibration quality may differ substantially between the two evaluation regimes.

Before temperature scaling, all three models exhibited similar calibration on the ID subset, with expected calibration error (ECE) values ranging from 0.137 to 0.144. In contrast, larger differences were observed on the OOD subset. The standard balanced fine-tuning model and the linear probe achieved comparable OOD ECE values (0.215 and 0.226, respectively), whereas the calibration-oriented fine-tuning model produced a substantially higher OOD ECE of 0.423.

Post-hoc temperature scaling affected the models differently. The fitted temperature was

Table 5: Calibration performance before and after post-hoc temperature scaling.

Model	Fitted T	ECE (ID)	ECE (OOD)	Brier (ID)	Brier (OOD)
Linear probe	1.01	0.137 → 0.136	0.226 → 0.225	0.597	0.647
Standard balanced fine-tuning (v4)	2.05	0.143 → 0.124	0.215 → 0.081	0.374	0.648
Calibration-oriented fine-tuning (v5)	0.89	0.144 → 0.110	0.423 → 0.450	0.307	0.877

1.01 for the linear probe, 2.05 for the standard balanced fine-tuning model, and 0.89 for the calibration-oriented fine-tuning model. Applying temperature scaling reduced the OOD ECE of the standard balanced fine-tuning model from 0.215 to 0.081, while the calibration-oriented fine-tuning model showed a slight increase from 0.423 to 0.450. The linear probe exhibited negligible changes following calibration.

The Brier scores showed a similar trend. The calibration-oriented fine-tuning model achieved the lowest ID Brier score (0.307), indicating the best probability estimates on ID samples, but produced the highest OOD Brier score (0.877). In comparison, the standard balanced fine-tuning model and the linear probe yielded substantially lower OOD Brier scores of 0.648 and 0.647, respectively. These results demonstrate that calibration performance differed considerably between the ID and OOD evaluation regimes.

3.4 Accept–Reject Tradeoff

We evaluated deployment performance using the threshold-based accept–reject strategy. The k-NN OOD score, identified as the best-performing OOD detector, was used as the abstention signal. Samples with scores exceeding a threshold τ were rejected and referred for clinical review. Figure 4 presents the corresponding operating curves.

At the 25% rejection operating point, the standard balanced fine-tuning model (v4) achieved the best overall trade-off. It correctly rejected 50.9% of OOD samples while falsely rejecting only 1.6% of ID samples. Under the same operating point, the calibration-oriented fine-tuning model (v5) rejected fewer OOD samples (47.7%) and exhibited a higher ID false-rejection rate (4.6%).

Across all evaluated rejection thresholds, the standard balanced fine-tuning model consistently achieved higher OOD catch rates while maintaining lower ID false-rejection rates than the calibration-oriented model. The accepted-ID accuracy of the standard balanced model remained stable across operating points (0.775–0.780), whereas the calibration-oriented model showed only a modest increase (0.781–0.798) as more samples were rejected.

Figure 5 further illustrates the classification performance across the full, ID, and OOD test sets. The standard balanced fine-tuning model exhibits more concentrated diagonal entries

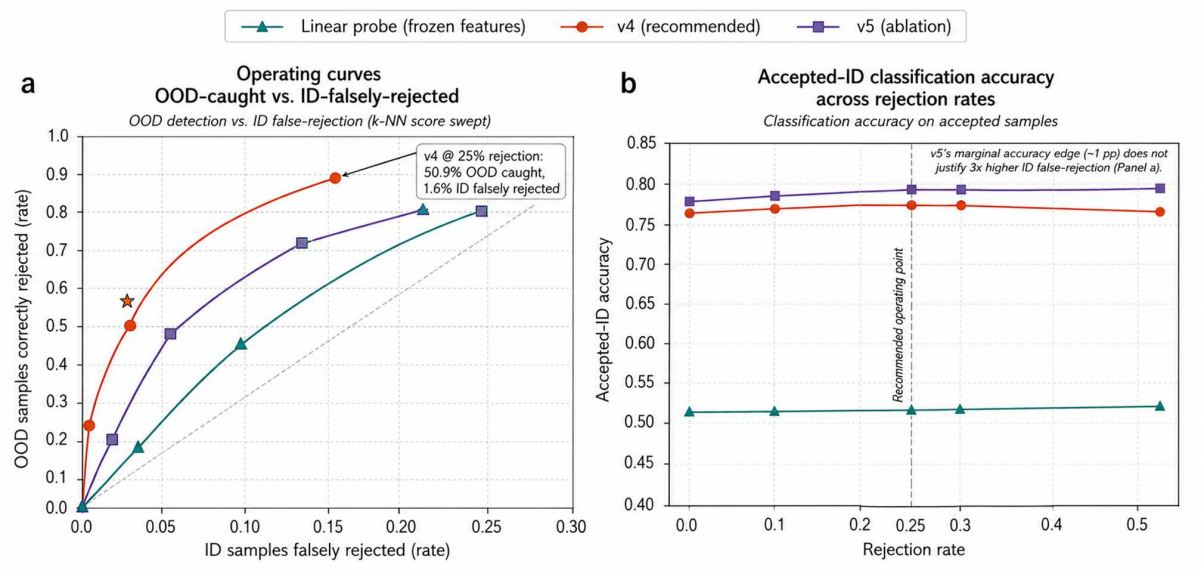


Figure 4: Accept–reject performance using the k-NN OOD score. Curves illustrate the relationship between rejection rate, accepted ID accuracy, OOD catch rate, and ID false-rejection rate for the evaluated models.

than the other models, indicating more reliable classification across both evaluation regimes. In contrast, the calibration-oriented model shows substantial confusion between normal and glaucoma cases on the OOD subset, misclassifying 334 of 482 normal ORIGA images (69%) as glaucoma.

3.5 Qualitative Per-Sample Behavior

To complement the quantitative evaluation, Figure 6 presents twelve representative test images together with the predicted class, post-temperature-scaling confidence, k-NN OOD score, and accept–reject decision for each model at the 25% rejection threshold. The examples illustrate three representative scenarios: (i) in-distribution samples correctly classified by all models, (ii) OOD normal cases where the calibration-oriented model fails while the standard balanced model succeeds, and (iii) challenging OOD glaucoma cases.

Row 1 contains representative ID samples from APTOS2019 and G1020. All three models correctly classify these images, although the fine-tuned models generally assign higher confidence than the frozen-feature linear probe.

Row 2 presents representative normal ORIGA images. The standard balanced fine-tuning model correctly classifies all four samples, whereas the calibration-oriented model consistently predicts glaucoma with high confidence. Two of these samples are rejected by the k-NN abstention mechanism, while the remaining two are accepted despite incorrect predictions. This behavior is consistent with the reduced OOD robustness of the calibration-oriented model observed in the quantitative analyses.

Row 3 illustrates representative glaucoma cases from the ORIGA dataset. The standard

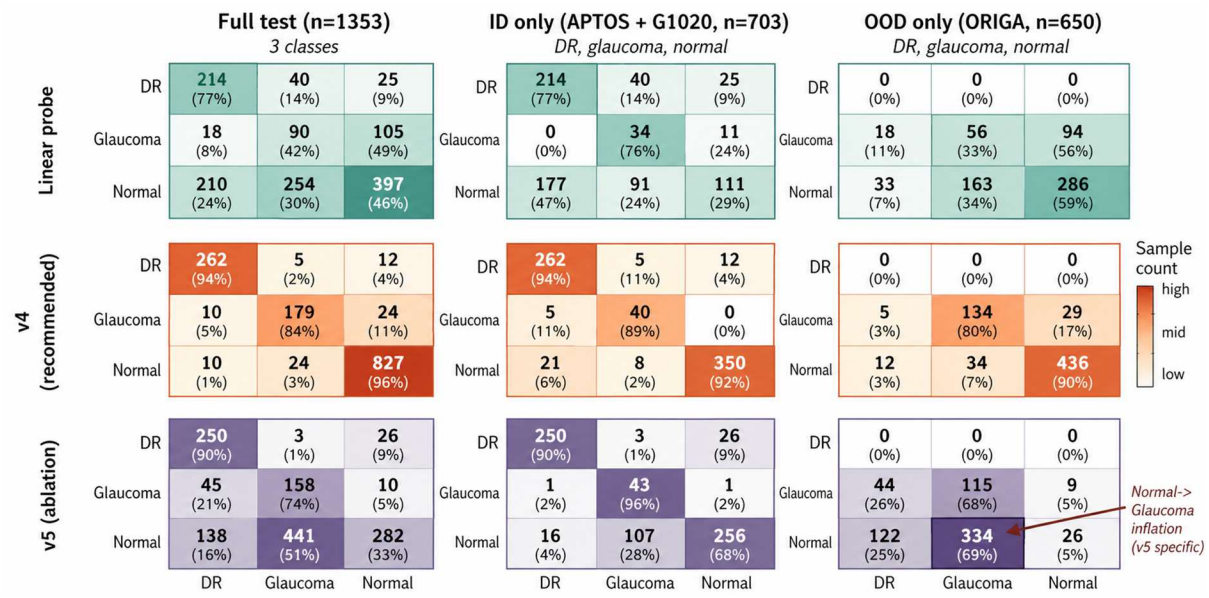


Figure 5: Confusion matrices for the linear probe, standard balanced fine-tuning model (v4), and calibration-oriented fine-tuning model (v5) on the full test set, the ID subset, and the OOD subset.

balanced fine-tuning model correctly identifies three of the four samples, whereas both fine-tuned models fail on the most challenging example. This case demonstrates that, although the proposed deployment strategy substantially improves robustness under external-domain shift, difficult OOD samples remain.

Overall, the qualitative examples closely reflect the quantitative results. The standard balanced fine-tuning model provides more reliable predictions under external-domain shift while maintaining effective OOD rejection, although no model completely eliminates challenging OOD failures.

4 Discussion

This study investigated whether a calibration-oriented fine-tuning strategy improves the robustness of a retinal foundation model under external-domain shift. Contrary to our initial expectation, the standard balanced fine-tuning model consistently outperformed the calibration-oriented model across nearly all deployment-oriented evaluations, including out-of-distribution (OOD) detection, calibration under distribution shift, and accept–reject performance. The standard balanced model achieved higher OOD detection performance (AUROC 0.938 versus 0.878), substantially lower post-temperature-scaling OOD calibration error (ECE 0.081 versus 0.450), and markedly better generalization on the held-out ORIGA dataset. Although the calibration-oriented model achieved the lowest in-distribution Brier score, this advantage did not extend to the external-domain setting, where both its calibration and classification performance deteriorated considerably. These findings highlight that improvements measured

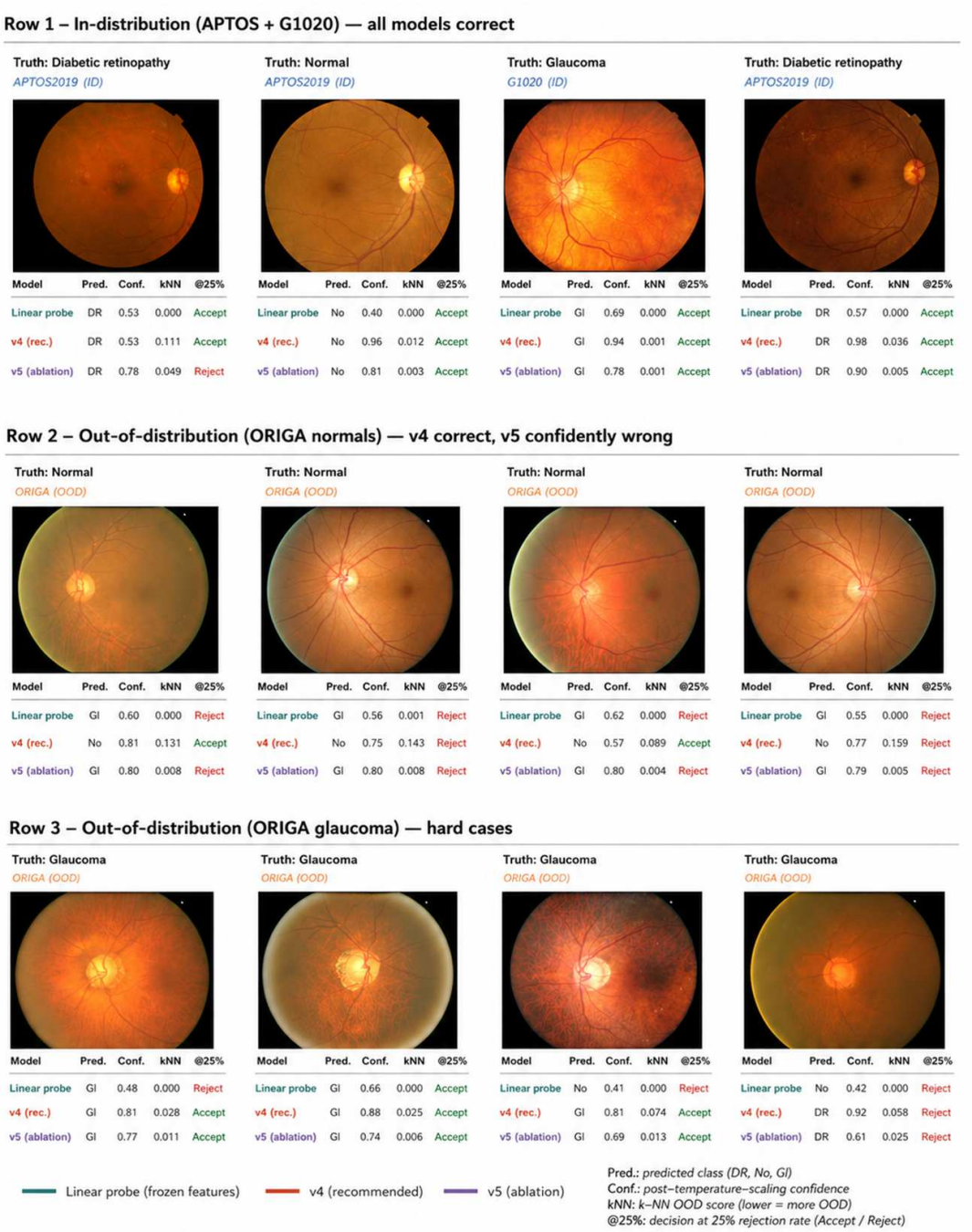


Figure 6: Representative test samples illustrating model behavior under different deployment scenarios. Row 1 shows correctly classified in-distribution cases. Row 2 presents OOD normal images that are correctly classified by the standard balanced fine-tuning model but misclassified as glaucoma by the calibration-oriented model. Row 3 shows representative OOD glaucoma cases, including a challenging example misclassified by both fine-tuned models. Predictions include the predicted class, calibrated confidence, k-NN OOD score, and accept–reject decision at the 25% rejection threshold.

on in-distribution data alone may not reliably predict deployment performance in open-world clinical environments.

One possible explanation for these observations is that the calibration-oriented training

strategy produces feature representations that are well suited for in-distribution calibration but less robust to distribution shift. The combination of cosine-similarity classification, label smoothing, and class-balancing strategies may encourage compact feature representations that improve confidence estimates within the training distribution while reducing the separation between known and unseen samples in feature space. Consistent with this interpretation, the calibration-oriented model exhibited substantially weaker OOD detection and remained poorly calibrated after post-hoc temperature scaling, whereas the standard balanced model showed a large reduction in OOD calibration error following temperature scaling. Although additional ablation studies are required to isolate the contribution of each training component, the present results suggest that calibration-oriented optimization does not necessarily improve robustness under external-domain shift.

A second important finding is the consistently strong performance of feature-space k-nearest-neighbor (k-NN) OOD detection. Across all evaluated models, k-NN achieved the highest AUROC and outperformed output-based scoring methods, including maximum softmax probability, energy, and max-logit scoring. Notably, the frozen EyeCLIP features already provided competitive OOD detection performance, indicating that pretrained foundation-model representations capture substantial information for identifying unfamiliar samples. Fine-tuning further improved this capability for the standard balanced model, whereas the calibration-oriented strategy produced little additional benefit. These observations suggest that, for retinal foundation models, feature-space information may provide a more reliable basis for OOD detection than prediction confidence alone.

Our findings are consistent with previous studies demonstrating that uncertainty estimation methods developed for in-distribution data often degrade under distribution shift. Ovadia et al. [34] showed that predictive uncertainty frequently becomes unreliable when models are evaluated on shifted data distributions, while Sun et al. [24] reported the effectiveness of feature-space k-NN for OOD detection across general computer vision benchmarks. Our results extend these observations to retinal foundation models and further suggest that classifier architecture can influence the effectiveness of different OOD scoring methods. In particular, logit-based OOD scores performed considerably worse for the calibration-oriented model than for the standard balanced model, whereas feature-space approaches remained consistently effective. These findings indicate that the choice of OOD scoring method should be considered jointly with the underlying classifier architecture.

From a deployment perspective, the proposed accept–reject framework provides an additional safety mechanism for clinical screening by allowing uncertain or unfamiliar cases to be referred for expert review. At the 25% rejection operating point, the standard balanced model correctly rejected approximately half of the OOD samples while falsely rejecting only 1.6% of in-distribution cases, representing a favorable balance between OOD detection and clinical review burden. The qualitative examples further demonstrated that the standard balanced model produced more reliable predictions on externally acquired retinal images, although challenging

OOD cases remained for all evaluated models. These observations suggest that feature-space OOD detection can improve deployment reliability but should complement, rather than replace, clinical oversight.

This study has several limitations. First, the calibration-oriented model was evaluated as a combined training strategy rather than through individual component ablations, preventing the contribution of each modification from being isolated. Second, ORIGA served as the only external-domain dataset; evaluating additional datasets acquired under different imaging conditions and patient populations would strengthen the generalizability of the conclusions. Third, we report point estimates without confidence intervals or formal statistical significance testing. Finally, the proposed accept–reject workflow has not been evaluated prospectively in clinical practice, and its operational value should be confirmed through clinician-in-the-loop studies.

Future work should systematically evaluate the individual contribution of each calibration-oriented component through controlled ablation experiments, extend the proposed framework to multiple external retinal datasets, and compare feature-space OOD detection with recent training-time approaches such as outlier exposure. Prospective clinical evaluation will also be important to determine whether improved OOD detection translates into meaningful reductions in diagnostic errors and clinical workload during real-world retinal screening.

5 Methods

5.1 Datasets and Label Harmonization

We use three publicly available color fundus photography datasets, summarized in Figure 7. All three are widely used benchmarks in retinal disease classification, and their inclusion allows the study to span two distinct disease targets (diabetic retinopathy and glaucoma) across heterogeneous acquisition conditions. APTOS2019 [11] contributes 3,662 color fundus images originally annotated for a five-class diabetic retinopathy severity task (grades 0–4). For this study, we collapse the severity grades into a binary scheme: grade 0 as normal and grades 1–4 as diabetic retinopathy. This harmonization preserves the clinically meaningful distinction between normal eyes and diabetic retinopathy while aligning the label space with the other two datasets, neither of which uses graded DR annotations, as summarized in Table 6. G1020 [12] contains 1,020 fundus images annotated for glaucoma versus normal. Labels are used as provided. G1020 was compiled to improve ethnic and demographic diversity in glaucoma benchmarks and exhibits greater within-class variability in imaging conditions than older benchmarks. ORIGA [13] contributes 650 fundus images annotated for glaucoma versus normal, drawn from the Singapore Malay Eye Study. Labels are used as provided. Compared to G1020, ORIGA originates from a single population cohort with narrower acquisition variability but differs from G1020 and APTOS in camera, illumination, and population characteristics.

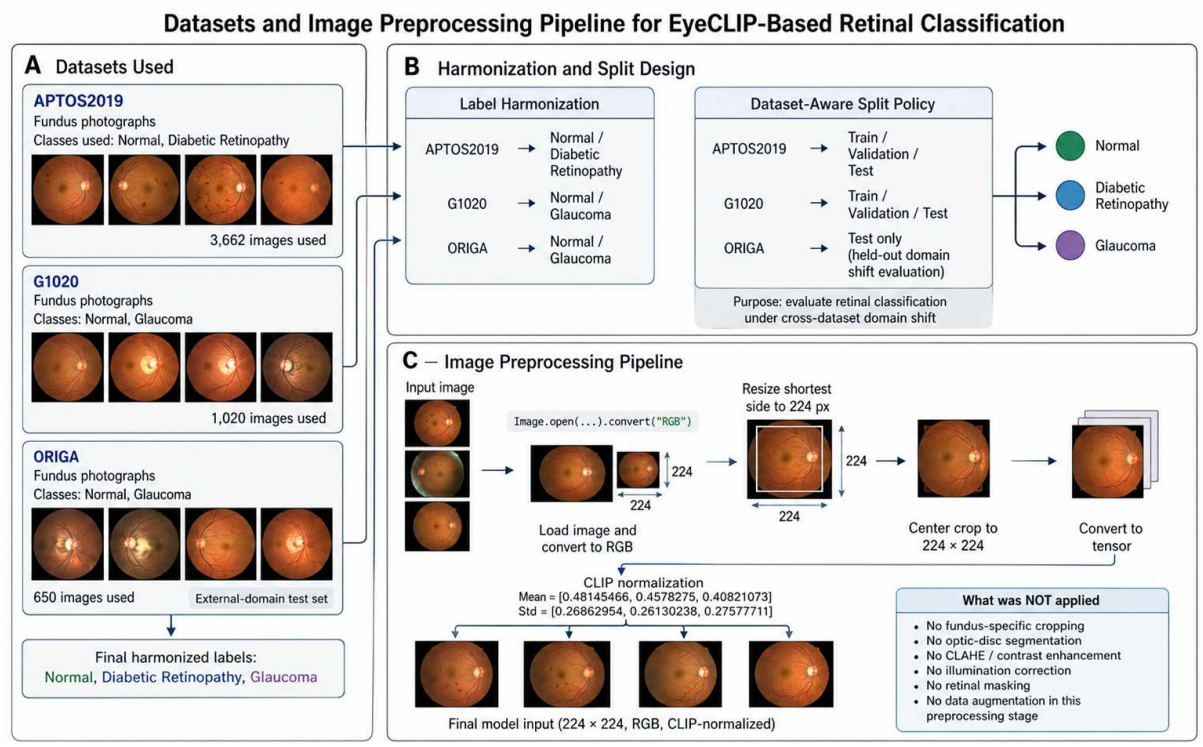


Figure 7: Three eye image datasets (APTOS2019, G1020, ORIGA) were combined into one 3-class task: normal, diabetic retinopathy, and glaucoma. APTOS and G1020 were used for training, validation, and testing, whereas ORIGA was used only as an external test set to check how the model handled new data. All images were processed in the same way using the CLIP pipeline: converted to RGB, resized so the shorter side is 224 pixels, center-cropped to 224 X 224, converted to tensors, and normalized with CLIP mean and standard deviation.

Label harmonization across the three sources, Figure 7(B), yields a unified three-class target space: normal, diabetic_retinopathy, glaucoma. Image preprocessing, shown in Figure 7(C) follows the standard CLIP visual pipeline to remain consistent with the pretrained EyeCLIP backbone. Each input fundus image is loaded from disk and converted to RGB, resized so that its shortest side matches 224 pixels, and center-cropped to 224×224, converted to a normalized tensor, and channel-wise normalized using the CLIP mean [0.4815, 0.4578, 0.4082] and standard deviation [0.2686, 0.2613, 0.2758]. The preprocessing pipeline is identical across all three models compared in this study and across all three datasets; therefore, any observed differences in performance or calibration reflect model design and training rather than input-processing differences. The three datasets differ substantially in acquisition conditions and class distribution. Class distributions are heterogeneous: APTOS2019 after binary relabeling is approximately 49.3% normal and 50.7% diabetic retinopathy, G1020 is 71.0% normal and 29.0% glaucoma, and ORIGA is 74.2% normal and 25.8% glaucoma. When aggregated across all three datasets, the combined three-class task is imbalanced, with approximately 35% diabetic retinopathy, 9% glaucoma, and 56% normal samples, a roughly 6:1 normal-to-glaucoma ratio that motivates the class-weighted training strategies compared in this study.

Table 6 reports the total number of images per dataset, class-wise distribution, and split

Table 6: Dataset composition and split distribution for retinal classification.

Dataset	Total Images	Normal	Diabetic Retinopathy	Glaucoma	Train	Validation	Test
APTOS2019	3,662	1,805	1,857	0	2,563	549	550
G1020	1,020	724	0	296	714	153	153
ORIGA	650	482	0	168	0	0	650
Total (combined)	5,332	3,011	1,857	464	3,277	702	1,353

allocation, highlighting both the class imbalance and the separation between in-domain and out-of-domain evaluation.

5.2 Open-World Split Protocol

We construct an open-world evaluation protocol by partitioning the three datasets across train, validation, and test with an explicit domain-shift design. APTOS2019 and G1020 are each randomly partitioned into training, validation, and test subsets at a 70/15/15 ratio, with stratification on the harmonized class label to preserve the original class balance within each subset. Random assignments use a fixed seed (42) to ensure reproducibility. ORIGA is not partitioned: all 650 ORIGA samples are assigned to the test split exclusively and never appear in the training or validation sets. The resulting compositions are listed in Table 7. This design establishes two evaluation regimes within the test split. The in-distribution (ID) regime comprises test samples drawn from APTOS and G1020 ($n = 703$). The out-of-distribution (OOD) [24] regime comprises 650 ORIGA samples. Because no ORIGA sample is observed during training or validation, performance on ORIGA measures zero-shot transfer to an unseen dataset. Any systematic difference between ORIGA and the APTOS/G1020 sources in the camera model, acquisition protocol, patient demographics, image resolution, or labeling convention contributes to the distribution shift that the model must implicitly detect at test time. This is the sense in which ORIGA serves as an open-world probe for our evaluation. Two design choices warrant explanation. First, the split is deliberately unbalanced across datasets: ORIGA contributes to the test split only, whereas APTOS and G1020 contribute to all three splits. This is not an artifact of random sampling but a design choice that creates a controlled distribution shift between the training distribution and a subset of the test distributions. Our protocol intentionally sacrifices uniformity across sources to create a genuine open-world evaluation regime. Second, we do not report macro-averaged test metrics across the three datasets. Such averages conflate in-distribution generalization and zero-shot dataset transfer into a single number that has no clear operational interpretation. Therefore, all test-set results are reported either per dataset (APTOS, G1020, ORIGA separately) or per regime (ID vs. OOD). When a single aggregate test number is required for completeness, it is reported alongside the per-regime breakdown and its

limited interpretability is highlighted. Because ORIGA is held out entirely, any property that differs systematically between ORIGA and the other two datasets contributes to our open-world signal, including properties unrelated to disease presentation, such as the camera model and illumination. We view this as a faithful reflection of real-world deployment, where models encounter distribution shifts driven by several simultaneous factors. However, this means that we cannot isolate the contribution of any single shift factor from this protocol alone; extending the evaluation to additional held-out datasets with more controlled axes of variation is a planned extension.

Table 7: Data composition by split and dataset

Split	APTOS (DR / normal)	G1020 (glauc. / normal)	ORIGA (glauc. / normal)	Total
Train	1300 / 1263	207 / 507	0 / 0	3277
Val	278 / 271	44 / 109	0 / 0	702
Test	279 / 271	45 / 108	168 / 482	1353

5.3 Backbone

All three models in this study shared a common visual encoder based on the OpenAI CLIP ViT-B/32 architecture [6], with weights initialized from the published EyeCLIP checkpoint. The backbone structure is summarized in Figure 8. Only the visual encoder was used; EyeCLIP's text encoder was omitted because our task was purely image-to-class classification, and the text modality played no downstream role. Each input image ($224 \times 224 \times 3$ RGB, preprocessed as described in 5.1) is divided into 49 non-overlapping 32×32 patches through a stride convolutional patch embedding, producing a sequence of 49 patch tokens of dimension 768. A learned [CLS] token is prepended, giving a 50×768 input sequence to the transformer stack, with further details mentioned in Figure 8 and in the classifier head attached to the backbone output. This shared backbone design isolates the effects of fine-tuning strategy and classifier choice on downstream behavior.

5.4 Model v4: Standard Fine-Tuning

Model v4 is a simpler fine-tuning variant designed to reflect the minimum recipe that might be adopted when applying EyeCLIP to a new classification task. v4 and v5 share the EyeCLIP ViT-B/32 visual backbone but differ in fine-tuning scope, classifier head architecture, loss parameterization, augmentation, and sampling strategy. The linear probe is trained on frozen L2-normalized EyeCLIP features using multinomial logistic regression. **Classifier head:** A linear classifier (`nn.Linear(512,3)`) is appended to the L2-normalized backbone output. The head produces logits $z = Wf + b$ with $W \in \mathbb{R}^{3\times512}$ and $b \in \mathbb{R}^3$, followed by a softmax to obtain class probabilities. The head weights are initialized from $\mathcal{N}(0, 0.01^2)$ and the bias is initialized

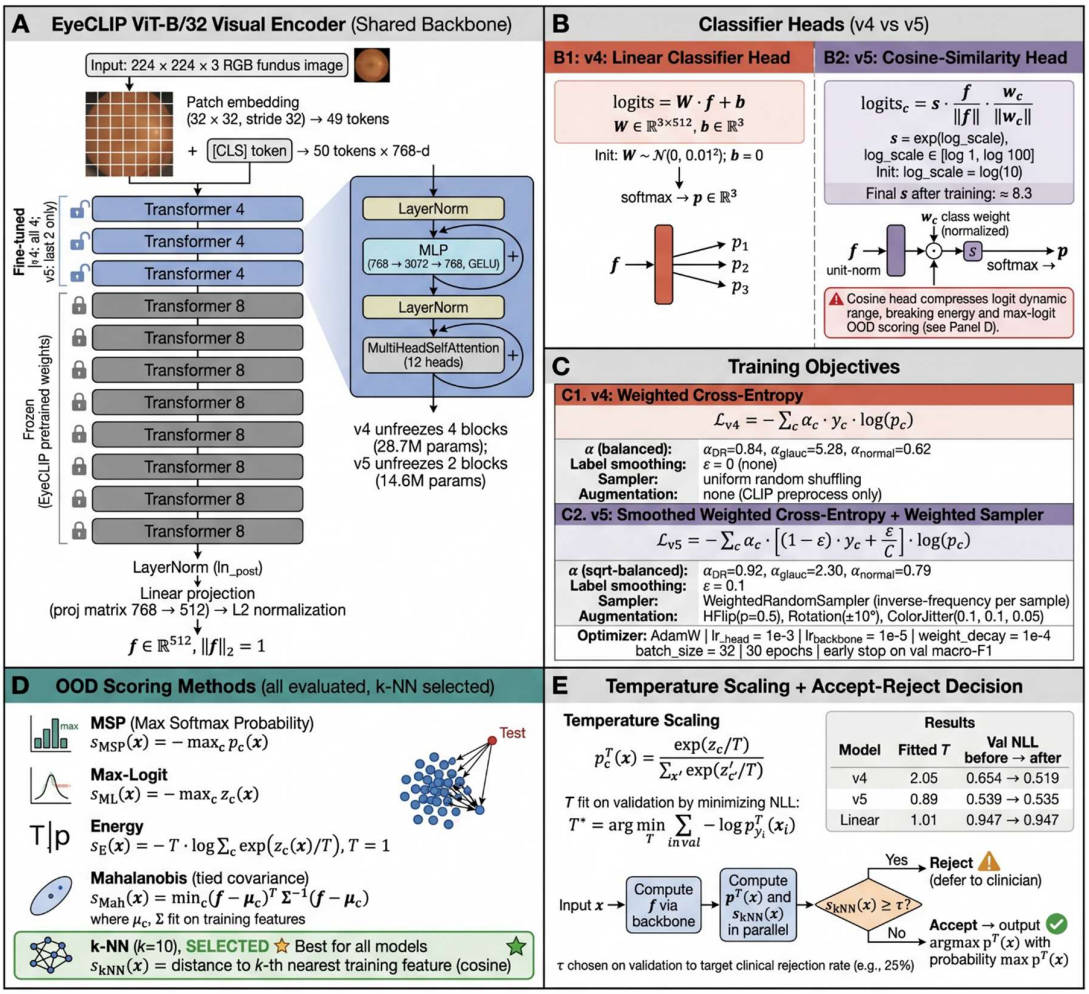


Figure 8: Architectural and procedural details of the three compared models. (A) Shared EyeCLIP ViT-B/32 visual backbone initialized from pretrained EyeCLIP weights. (B) Classifier head. v4 uses a standard linear layer with weighted cross-entropy, and v5 replaces this with a cosine similarity head with a learnable temperature. (C) Training objectives. v4 uses inverse-frequency class weights, and v5 applies square-root-scaled weights and label smoothing. (D) Out-of-distribution scoring methods evaluated in this study. (E) Post-hoc temperature scaling, with temperature fitted on the validation set by minimizing negative log-likelihood, followed by the accept–reject deployment decision.

to zero. **Fine-tuning scope:** The last four transformer blocks of the visual encoder, together with the final LayerNorm (`ln_post`) and the 768 → 512 projection matrix, are unfrozen and updated during training. Further model details are provided in Table 8, which summarizes the full configurations of the three models compared in this study. All parameter counts are exact, obtained by summing over `model.parameters()`. This yields 28.7M trainable parameters out of 151.3M total backbone-plus-head parameters. The choice to fine-tune the top four blocks follows standard practice for CLIP-family transfer learning [40], where fine-tuning deep blocks is sufficient to specialize high-level representations while preserving the low-level features learned during pretraining.

Training objective: The loss is standard cross-entropy with inverse-frequency class weights,

implemented via PyTorch's `CrossEntropyLoss(weight=α)` as shown in Eq. 1:

$$L_{v4} = -\sum_{c} \alpha_c y_c \log(p_c) \tag{1}$$

Here, α_c denotes the class weight computed from the training split class frequencies using the scikit-learn "balanced" scheme, yielding $\alpha_{\text{DR}} = 0.84$, $\alpha_{\text{glaucoma}} = 5.28$, and $\alpha_{\text{normal}} = 0.62$. No label smoothing is applied ($\varepsilon = 0$.)

Optimization: Training uses AdamW with separate learning rates for the classifier head (1×10^{-3}) and the unfrozen backbone parameters (1×10^{-5}), weight decay 1×10^{-4}, gradient clipping at max-norm 1.0, and batch size 32. Training runs for up to 30 epochs with early stopping on the validation macro-F1 (patience 7). In practice, v4 converged to its best validation checkpoint at epoch 9, and training was terminated at epoch 16.

5.5 Model v5: Calibration-Aware Fine-Tuning

Model v5 shares v4's EyeCLIP ViT-B/32 backbone and overall training recipe but introduces four modifications drawn from the calibration and class-imbalance literature, each independently associated with improved probability quality or better minority-class handling in prior work. The intent behind bundling these four modifications rather than evaluating them in isolation was to test whether the combined "calibration-aware" recipe, as it might be adopted by a practitioner following standard recommendations from the calibration and class-imbalance literature, produces better open-world behavior than the simpler v4 baseline.

The first and most architecturally consequential modification was the replacement of v4's linear classifier head with a cosine-similarity head incorporating a learnable temperature. In this head, the L2-normalized 512-dimensional feature vector f produced by the backbone is compared against $C = 3$ class-prototype vectors w_c, also L2-normalized at forward time, producing logits, as shown in Eq. 2:

$$z_c = s \cdot \left(\frac{f}{\|f\|}\right) \cdot \left(\frac{w_c}{\|w_c\|}\right) \tag{2}$$

where s is a learnable scalar parameter defined as $s = \exp(\text{log_scale})$. The log-scale parameter is initialized at $\log(10)$, a commonly used default that places the initial effective temperature in a numerically stable range and is clamped during training to the interval $[\log 1, \log 100]$ to prevent the temperature from drifting to degenerate values. Class-prototype weights are initialized from $\mathcal{N}(0, 0.01^2)$, independently of the log_scale. During v5 training, s drifted from its initial value of 10 to approximately 8.3 at the best-validation checkpoint (epoch 25), indicating that the learning dynamics converged to a mildly softened softmax compared with the initial setting. Cosine similarity heads are popular in face recognition and few-shot classification literature and have been recommended for imbalanced classification because the unit-norm class prototypes prevent majority classes from dominating through weight-vector magnitude alone.

Table 8: Model configuration summary for v4 (Standard balanced fine-tuning), v5 (Calibration-oriented fine-tuning), and linear probe.

Component	**v4 (Standard balanced fine-tuning)**	**v5 (Calibration-oriented fine-tuning)**	**Linear probe**
Backbone	EyeCLIP ViT-B/32	EyeCLIP ViT-B/32	EyeCLIP ViT-B/32
Input resolution	$224 \times 224 \times 3$	$224 \times 224 \times 3$	$224 \times 224 \times 3$
Patch size / stride	32 / 32	32 / 32	32 / 32
Tokens (incl. [CLS])	50	50	50
Embedding dim	768	768	768
Transformer blocks	12	12	12
Attention heads / block	12	12	12
MLP hidden dim	3072	3072	3072
Output feature dim	512 (L2-normalized)	512 (L2-normalized)	512 (L2-normalized)
Fine-tuning scope	Last 4 blocks + ln_post + proj	Last 2 blocks + ln_post + proj	None (frozen)
Trainable params	28,747,779	14,572,033	~1,500
Total params	151,278,852	151,278,850	151,278,340
Classifier head	Linear ($\mathbb{R}^{512} \to \mathbb{R}^3$)	Cosine similarity + learned T	Multinomial LogReg
Head init	$W \sim \mathcal{N}(0, 0.01^2)$, $b = 0$	$W \sim \mathcal{N}(0, 0.01^2)$, log_scale = $\log(10)$	sklearn default
Learnable temperature	No	Yes (final ≈ 8.3)	n/a
Class weights	Balanced (0.84 / 5.28 / 0.62)	sqrt-balanced (0.92 / 2.30 / 0.79)	Balanced
Label smoothing (ε)	0	0.1	n/a
Sampler	Uniform shuffle	WeightedRandomSampler (inverse-freq)	n/a
Augmentation	None	HFlip(0.5), Rot($\pm 10°$), ColorJitter(0.1, 0.1, 0.05)	None
Optimizer	AdamW	AdamW	LBFGS (sklearn default)
Learning rate (head)	1×10^{-3}	1×10^{-3}	n/a
Learning rate (backbone)	1×10^{-5}	1×10^{-5}	n/a
Weight decay	1×10^{-4}	1×10^{-4}	L2, $C = 1.0$
Batch size	32	32	full-batch
Max epochs	30	30	converge
Early stopping	patience 7 (val macro-F1)	patience 7 (val macro-F1)	convergence
Gradient clipping	max-norm 1.0	max-norm 1.0	n/a
Best epoch (achieved)	9	25	–
Best val macro-F1	0.690	0.704	0.479

The second modification was a softer class-weight scheme for the cross-entropy loss. Whereas v4 uses the full inverse-frequency (balanced) scheme yielding $\alpha_{\text{glaucoma}} = 5.28$, v5 applies its square-root, as shown in Eq. 3.

$$\alpha_c = \sqrt{\frac{N}{C \cdot N_c}} \tag{3}$$

where N is the training set size, $C = 3$ is the number of classes, and N_c is the count in class c. This yields a softer reweighting: v5's $\alpha_{\text{glaucoma}} = 2.30$, compared to v4's $\alpha_{\text{glaucoma}} = 5.28$ (full values: $\alpha_{\text{DR}} = 0.92$, $\alpha_{\text{glaucoma}} = 2.30$, $\alpha_{\text{normal}} = 0.79$).

The third modification is the weighted random sampler, which shapes the class composition of each training batch for the model. Whereas v4 draws training batches through uniform random shuffling (each sample has an equal probability of being drawn in each epoch), v5 uses PyTorch's `WeightedRandomSampler` with per-sample weights equal to the square-root-balanced class weight of each sample class. Weighted sampling is motivated by the practical observation that class-weighted cross-entropy alone can still produce noisy minority-class gradients when minority samples are rare within each batch: with glaucoma at $\sim 6\%$ and a batch size of 32, many batches in uniform sampling contain zero glaucoma samples, leaving the model with no glaucoma signal to fit against in those steps.

The fourth modification was label smoothing with $\varepsilon = 0.1$ applied to the cross-entropy loss. Label smoothing replaces each one-hot target y_c with a soft target $(1-\varepsilon)y_c + \varepsilon/C$, yielding the smoothed loss in Eq. 4:

$$L_{v5} = -\sum_c \alpha_c \cdot \left[(1-\varepsilon)y_c + \frac{\varepsilon}{C}\right] \cdot \log(p_c) \tag{4}$$

Label smoothing is among the most widely adopted training-time calibration techniques, as reported in the literature, to reduce the over-confidence characteristic of modern deep networks trained with standard cross-entropy [41]. In the three-class setting with $\varepsilon = 0.1$, smoothing effectively asks the model to produce probabilities of $(0.933, 0.033, 0.033)$ for the correct class and its two alternatives rather than the one-hot $(1, 0, 0)$ target.

In addition to these four head-and-loss modifications, v5 applies training-time augmentation that v4 omits: each training image after the standard CLIP preprocessing described in Section 5.1 is subject to a random horizontal flip with a probability of 0.5, a random rotation uniformly sampled from $[-10°, +10°]$, and mild ColorJitter with brightness ± 0.1, contrast ± 0.1, and saturation ± 0.05. These augmentations were chosen to be fundus-appropriate: horizontal flips are valid for fundus images because laterality is not encoded as a pattern; small rotations simulate acquisition-angle jitter; and mild color jitter simulates camera/illumination variation across clinics without introducing clinically implausible color shifts.

A final training configuration difference partially offsets v5's added capacity and regularization. Because the cosine head, weighted sampler, and label smoothing each increase the risk

of overfitting to the training distribution, v5 unfreezes only the last two transformer blocks of the EyeCLIP visual encoder rather than v4's four, together with the final LayerNorm (`ln_post`) and the 768 $\rightarrow$ 512 projection matrix. This yielded 14.6M trainable parameters, which was approximately half of v4's 28.7 M.

All remaining hyperparameters were identical to v4 to ensure that any observed performance or calibration difference between the two models was attributable to the modifications above, not to confounding differences in optimization. Specifically, v5 uses the same AdamW optimizer with the same differentiated learning rates (1×10^{-3} for the cosine classifier head, 1×10^{-5} for the unfrozen backbone blocks), same weight decay (1×10^{-4}), same gradient-norm clipping (max-norm 1.0), same batch size (32), same early stopping criterion (validation macro-F1 with patience 7), and the same maximum epoch budget (30) as v4. In practice, v5 was trained for the full 30-epoch budget without triggering early stopping, converging to its best validation checkpoint at epoch 25. v4, by contrast, reached its best checkpoint at epoch 9 and early stopped at epoch 16.

5.6 Linear Probe Baseline

In addition to the two fine-tuning variants, we included a frozen-feature linear-probe baseline Figure 9 that made no updates to the EyeCLIP backbone. The linear probe is motivated by two considerations: First, it establishes a lower bound on the performance achievable from EyeCLIP's pretrained features alone, without any task-specific feature adaptation. Any benefit observed in v4 or v5 relative to the linear probe is attributable to fine-tuning rather than EyeCLIP pretraining itself. Conversely, if the linear probe is competitive with a fine-tuned variant on a particular metric, this indicates that fine-tuning contributes little beyond what pretraining already captured on that metric. Second, a linear probe on L2-normalized features is a widely used transfer-learning baseline for CLIP-family models and is often reported to be surprisingly competitive for downstream tasks, particularly when labeled training data are limited.

These features, together with the harmonized class labels, were used to fit a multinomial logistic regression using scikit-learn's LogisticRegression class [42]. The logistic regression uses class_weight='balanced' (equivalent to v4's class-weighting scheme), inverse-regularization strength C = 1.0 (standard L2 penalty), the LBFGS solver for optimization, and multi_class='multinomial' to fit a true multinomial rather than a one-versus-rest classifier. The logistic regression model was optimized using the LBFGS solver on the extracted training features, without mini-batch gradient-based fine-tuning of the EyeCLIP backbone. At inference time, the linear probe extracts features through the (still frozen) EyeCLIP encoder and passes them through the fitted logistic regression classifier to obtain class probabilities. Because logistic regression produces probabilities directly rather than raw logits, reconstructed pseudo-logits, defined as $\log(p_c)$, are used foMSP, Mahalanobis distance, and k-nearest-neighbor distance do not require raw logits and are computed consistently across all three models. Notably, the linear

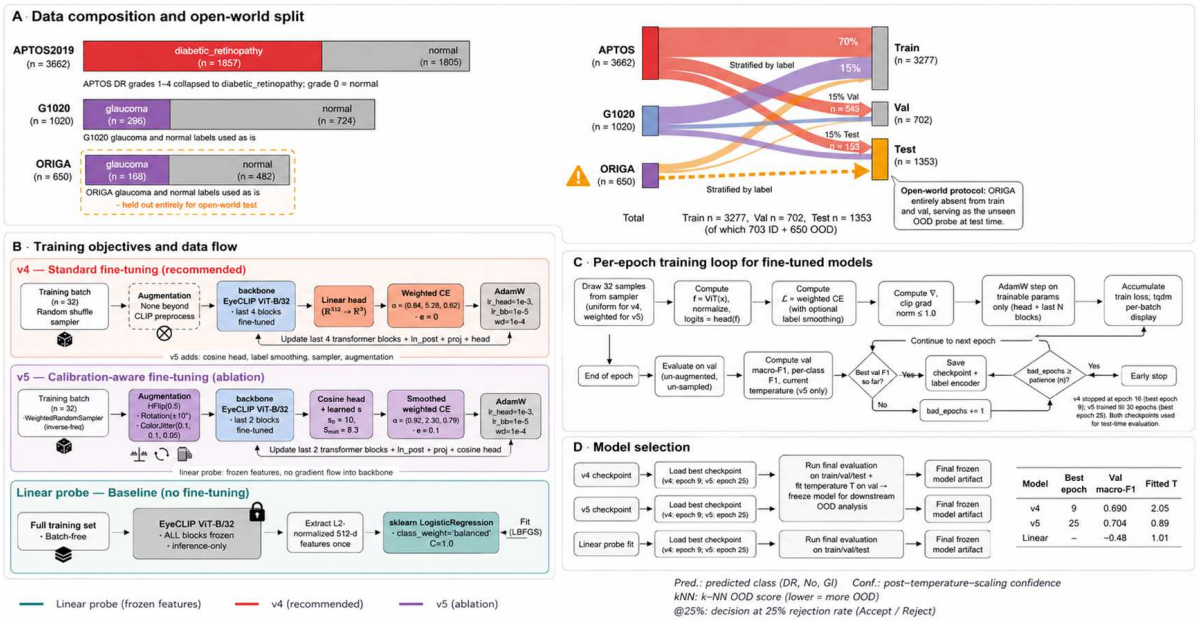


Figure 9: Overview of the open-world data split and training pipelines. APTOS2019 and G1020 are split into train, validation, and test subsets, ORIGA is reserved exclusively as an external open-world test set. Model v4 uses standard balanced cross-entropy fine-tuning with a linear classifier head. Model v5 uses calibration-oriented fine-tuning with a cosine-similarity head, weighted sampling, data augmentation, and label smoothing.

probe, despite its architectural simplicity, is not merely a weak baseline: EyeCLIP's pretrained features encode substantial retinal-image structure acquired during multimodal pretraining, and a linear probe on these features provides a meaningful comparison point for interpreting the contribution of fine-tuning.

5.7 OOD Detection Scores

To evaluate the capacity of each model for open-world detection, we computed five different out-of-distribution scoring functions for every test-set sample. The five scores span both output-space and feature-space methods (which operate on the penultimate 512-dimensional L2-normalized features), allowing us to identify the most informative score type for retinal fundus classification using EyeCLIP-derived features. All five scores follow the sign convention that a higher score indicates more out-of-distribution, with negation applied where needed so that the downstream threshold-based evaluation uses a single uniform convention.

The Maximum Softmax Probability (MSP) is the canonical output-space baseline for OOD detection. It treats the top 1 predicted probability as a proxy for model confidence and flags low-confidence predictions as out of distribution. Formally, as shown in Eq. 5:

$$s_{\text{MSP}}(x) = -\max_{c} \; p_c(x) \tag{5}$$

where $p(x)$ is the softmax output of the classifier.

Max-logit is a simple variant of MSP that operates directly on the pre-softmax logits, as shown in Eq. 6:

$$s_{\mathrm{ML}}(x) = -\max_{c} \; z_c(x) \tag{6}$$

Because max-logit depends on the absolute scale of the logits, it interacts with architectural choices that constrain logit magnitude, a property that helps explain the observed failure mode of cosine-similarity classifier heads.

Energy [18] is an output-space score derived from the negative log-sum-exp of the scaled logits, as defined in Eq. 7

$$s_E(x) = -T \cdot \log \sum_{c} \exp\left(\frac{z_c(x)}{T}\right) \tag{7}$$

with $T = 1$ was used throughout this study. Energy was introduced by [18] as a theoretically motivated alternative to MSP, with the appealing property that low energy (high log-sum-exp) corresponds to high predicted confidence across all classes simultaneously, rather than just the top 1 class. Energy is sensitive to logit-magnitude structure in the same way as max-logit and is therefore subject to the same cosine-head interaction observed in Section 3.2.

The Mahalanobis distance [23] is a classical feature space OOD score that models the in-distribution feature space as a mixture of class-conditional Gaussians. Following Lee et al.'s tied-covariance formulation, we fit a single shared covariance matrix Σ from the training-set features (pooling within-class residuals across all three classes) and per-class means μ_c. The OOD score for a test sample is the minimum Mahalanobis distance to any class of Gaussian, as shown in Eq. 8.

$$s_{\mathrm{Mah}}(x) = \min_{c} (f - \mu_c)^T \Sigma^{-1} (f - \mu_c) \tag{8}$$

where f is the 512-dimensional L2-normalized feature vector of the sample. To ensure that the empirical covariance is invertible, we add a shrinkage of 1×10^{-3} to the diagonal before inversion. The Mahalanobis distance measures how far a test point lies from the nearest class-conditional Gaussian under the learned feature geometry. It assumes that in-distribution features are approximately Gaussian-distributed per class.

Additionally, the k-nearest neighbor (kNN) distance is a non-parametric feature-space OOD score that avoids the Gaussian assumption of Mahalanobis. For a test feature f, we compute the cosine distance to every training-set feature and report the distance to the k-th nearest neighbor as follows:

$$s_{\mathrm{kNN}}(x) = \text{distance to the } k\text{-th nearest training feature (cosine distance)} \tag{9}$$

with $k = 10$ in this study. Cosine distance, rather than Euclidean distance, is the geometrically natural metric for L2-normalized features that lie on the unit hypersphere; it corresponds to angular separation and is invariant to the magnitude rescaling inherent to L2 normalization.

Moreover, the linear probe produces class probabilities directly through scikit-learn's predict$_proba$, $ratherthanrawpre-softmaxlogits.Allfivescoreswerecomputedfromasingleinferenc$ $rejectanalysesreusedthesameunderlyingpredictions.Thisensuresthatalltest-setanalysesofclosed-$ $setclassification, OODdetection, calibration, andaccept-rejecttradeoffsaremutuallyconsistentana$

Table 9: OOD detection and calibration procedure.

Procedure	**Details**
OOD scoring methods	MSP, Max-Logit, Energy ($T = 1$), Mahalanobis (tied covariance), k-NN ($k = 10$)
Score convention	Higher score = more out-of-distribution, across all methods
Feature reference (Mahalanobis, k-NN)	Training-split L2-normalized features
Mahalanobis shrinkage	1×10^{-3} (added to diagonal of tied covariance)
k-NN distance metric	Cosine distance ($1 - f_{\text{test}} \cdot f_{\text{train}}$)
ID / OOD assignment	ID: APTOS + G1020 ($n = 703$); OOD: ORIGA ($n = 650$)
Temperature scaling	Single scalar T fit on validation by minimizing NLL; grid search over $[0.05, 10.0]$
Fitted T (v4 / v5 / Linear)	2.05 / 0.89 / 1.01
Accept-reject threshold	τ on k-NN score; reported at 0%, 10%, 25%, and 50% rejection rates
Evaluation metrics (OOD)	AUROC, FPR@95TPR, AUPR
Evaluation metrics (calibration)	ECE (15 equal-width bins), adaECE (15 equal-count bins), MCE, Brier score, NLL
Software	PyTorch 2.x, OpenAI CLIP 1.0, scikit-learn 1.x, Python 3.10
Hardware	2 × NVIDIA RTX 5090 GPUs (32 GB VRAM each), Ubuntu 24.04 WSL2
Random seed	42 for Python, NumPy, PyTorch, and data splitting

5.8 Calibration Metrics

We evaluate calibration quality using five complementary metrics, all of which are standard in the modern calibration literature. The Expected Calibration Error (ECE) [28] is the weighted average discrepancy between predicted confidence and observed accuracy across confidence bins, as defined in Eq.

$$\text{ECE} = \sum_m \left(\frac{|B_m|}{N} \right) \cdot |\text{acc}(B_m) - \text{conf}(B_m)| \tag{10}$$

where samples are partitioned into $M = 15$ equal-width bins over the $[0, 1]$ confidence range. We additionally report adaptive ECE [43], which uses equal-count rather than equal-width binning and is more robust when confidence distributions concentrate sharply near 1.0, a common pattern with overconfident neural

networks. The full list of calibration metrics and binning schemes used is summarized in Table 9.

The Maximum Calibration Error (MCE) is the largest absolute gap over any bin,

$\max_m |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$, and captures the worst-case calibration behavior. The multiclass Brier score [44] is a proper scoring rule computed as the mean squared L2 distance between the predicted probability vector and the one-hot label, as shown in Eq. 11:

$$\text{Brier} = \frac{1}{N} \sum_i \|p(x_i) - y_i\|^2 \tag{11}$$

Unlike ECE, the Brier score is sensitive to the full probability vector rather than just the top 1 confidence alone and is therefore a more comprehensive measure of probability quality. Finally, we report the mean negative log-likelihood (NLL), which is the standard cross-entropy loss evaluated on the test data; like Brier, NLL is a proper scoring rule and is the natural objective for temperature scaling.

All five metrics were computed separately in the in-distribution regime (APTOS + G1020 test samples) and out-of-distribution regime (ORIGA test samples). Reporting per regime is essential to this study's contribution: a model that is well-calibrated on ID inputs but poorly calibrated on OOD inputs will appear well-calibrated under a single aggregate metric, obscuring exactly the failure mode we wish to measure.

5.9 Temperature Scaling

Temperature scaling [28] is a post-hoc calibration method that applies a single learned scalar $T > 0$ to the logits before softmax, as defined in Eq. 12.

$$p_c^T(x) = \frac{\exp(z_c/T)}{\sum_{c'} \exp(z_{c'}/T)} \tag{12}$$

Values $T > 1$ soften the softmax (reducing overconfidence), and values $0 < T < 1$ sharpen it. A crucial property of temperature scaling is that it preserves argmax: the predicted class remains unchanged at any T, so classification accuracy is invariant under scaling. Only probability magnitudes, and therefore confidence-based metrics (ECE, Brier, and NLL), change under scaling. This makes temperature scaling a low-cost, accuracy-preserving intervention, and it is widely regarded as the strongest single-parameter baseline for post-hoc calibration of neural network classifiers.

For each of the three trained models, we fitted an independent temperature T by minimizing the validation negative log-likelihood, as shown in Eq. 13:

$$T^* = \arg\min_T \sum_{i \in \text{val}} -\log p_{y_i}^T(x_i) \tag{13}$$

We perform the minimization via a two-stage grid search: a coarse pass over $T \in [0.05, 10.0]$ with 200 points, followed by a finer pass over $T \in [0.5, 2.0]$ with 151 points, with the overall best T from the two grids selected. A grid search is preferred over gradient-based optimization because the search space is one-dimensional, bounded, and well-behaved, and the grid approach is transparent and deterministic. The fitted T is then applied to the test-time logits before computing the softmax probabilities for the test metrics.

5.10 Accept-Reject Protocol

To quantify the deployment behavior under abstention, we operationalize the open-world task as a two-stage decision [45]: for each test input x, an OOD score $s(x)$ is computed using the selected scoring method and a threshold τ partitions the samples into accepted and rejected. Specifically, samples with $s(x) \geq \tau$ are rejected, the model abstains from prediction, deferring the input to a human clinician, and samples with $s(x) < \tau$ are accepted, with the class prediction obtained as the argmax of the temperature-scaled softmax probabilities. The threshold τ is a deployment-time hyperparameter that encodes the operator's preference over two competing costs: rejecting too many in-distribution samples wastes clinician time on cases the model could handle, while accepting too many out-of-distribution samples risks silent misclassification on inputs the model is not equipped to handle [46].

To characterize the resulting tradeoff, we sweep τ across the full range of observed test-set k-NN scores and report four summary quantities at each threshold. The rejection rate is the overall proportion of test samples rejected, as defined in Eq. 14:

$$\text{rejection rate} = \frac{|\{x : s(x) \geq \tau\}|}{N} \tag{14}$$

This serves as a simple operating-point descriptor. The accepted-ID accuracy is the classification accuracy of accepted samples drawn from the in-distribution test partition, which is the accuracy a clinician would observe in deployed use, conditional on the model choosing to produce a prediction [47]. Finally, the ID false-rejection rate is the proportion of true in-distribution samples incorrectly rejected and the cost paid in unnecessary clinician review to

achieve a given OOD catch rate.

Together, these five quantities describe the full operating landscape available at the time of testing. We additionally report anchor points at 0%, 10%, 25%, and 50% overall rejection rates to provide deployment-realistic comparison points across models.

5.11 Experimental Setup and Implementation Details

Training and inference are performed on a single workstation running Ubuntu 24.04 under Windows Subsystem for Linux 2 (WSL2), with an AMD Threadripper PRO 7975WX 32-core CPU, 256 GB system RAM, and two NVIDIA RTX 5090 GPUs (32 GB VRAM each). Fine-tuning and feature extraction are executed on a single GPU per run; the second GPU is used for parallel experimentation but never to parallelize the training of a single model. The software dependencies are PyTorch 2.6 with CUDA 12.8, OpenAI CLIP 1.0 (from which the ViT-B/32 backbone architecture and preprocessing pipeline are sourced), scikit-learn 1.4 (for the linear-probe logistic regression and for all classification and calibration metrics), tqdm 4.66 (for training progress display), NumPy 1.26, pandas 2.2, and Python 3.10. The EyeCLIP visual weights are loaded from the published checkpoint without architectural modification. The random seeds for PyTorch, NumPy, and Python's random module are fixed at the start of each training run to support reproducibility. The approximate wall-clock training time for each fine-tuned model is 45--50 min on a single RTX 5090 GPU (v4 approximately 45 min over 16 epochs before early stopping; v5 approximately 50 min over the full 30-epoch budget). Feature extraction for the linear probe completes in under 5 min.

5.12 Reproducibility Details

All train/validation/test splits are generated at the image level using a fixed random seed of 42 with stratification by the harmonized class label. ORIGA is excluded entirely from model training and validation and used only for external test-time evaluation. The same RGB conversion, resizing, center-croppi tensor conversion, and CLIP normalization are applied to APTOS2019, G1020, and ORIGA. Images that could not be opened or decoded are excluded before split generation. The EyeCLIP ViT-B/32 visual checkpoint used in this study is loaded without architectural modification. The split CSV files, preprocessing scripts, and training code will be released via a public repository upon acceptance to enable exact reproduction.

6 Conclusion

This study presented an open-world evaluation of retinal disease classification using the EyeCLIP foundation model, comparing a standard balanced fine-tuning strategy with a calibration-oriented fine-tuning approach. Contrary to our initial expectation, the standard balanced model consistently achieved superior performance across multiple deployment-relevant criteria, including out-of-distribut (OOD) detection, calibration under distribution shift, cross-dataset classification and accept--reject decision making. Although the calibration-oriented model improved probability estimation on in-distribution data, these benefits did not generalize to the external-domain setting.

Among the evaluated OOD detection methods, feature-space k-nearest-neighbor (k-NN) scoring consistently provided the strongest performance across all model variants. Furthermore, combining standard balanced fine-tuning with post-hoc temperature scaling substantially improved OOD calibration while preserving classification performance, highlighting the value of simple post-trainin calibration for open-world deployment.

Overall, our findings demonstrate that improved in-distribution calibration alone is insufficient to ensure reliable performance under distribution shift. Instead, the combination of robust feature representations, effective feature-space OOD detection, and appropriate post-hoc calibration provides a practical framework for more reliable open-world retinal screening. Future work should validate this framework across additional external datasets and prospective clinical settings.

Acknowledgements

This research was supported by King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia.

Ethics Statement

This study used only publicly available, de-identified retinal image datasets and involved no direct interaction with human participants. Therefore, additional institutional ethical approval was not required for the present computational study.

Author Contributions

Amreen Batool: Conceptualization, methodology, experiments, formal analysis, and writing--original draft. Faheem Ahmed: Methodology, validation, and writing, review, and editing. Fakhri Alam Khan: Supervision, conceptual guidance, validation, writing, review, and editing.

Data Availability

The datasets used in this study, including APTOS2019, G1020, and ORIGA/ORIGA-light, are publicly accessible through the following sources: APTOS2019 is available at https://www.kaggle.com/competitions/aptos2019-blindness-detection/data; G1020 is available at https://www.dfki.de/web/forschung/projekte-publikationen/publikation/10840/; and ORIGA/ORIGA-light can be accessed at https://www.kaggle.com/datasets/arnavjain1/glaucoma-datasets. The processed metadata, train/validation/test split files, and scripts used for preprocessing and evaluation will be made available upon reasonable request or through the project repository after publication.

Competing Interests

The authors declare no competing interests.

References

[1] Z. L. Teo et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128(11):1580--1591, 2021. doi: 10.1016/j.ophtha.2021.04.027.

[2] W. H. Yang et al. An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy. *Diabetes Therapy*, 10(5):1811, 2019. doi: 10.1007/s13300-019-0652-0.

[3] E. Beede et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2020. doi: 10.1145/3313831.3376718.

[4] V. Mehta et al. Evaluating transparency in ai/ml model characteristics for fda-reviewed medical devices. *npj Digital Medicine*, 8(1):673, 2025. doi: 10.1038/s41746-025-02052-9.

[5] D. Shi et al. A multimodal visual--language foundation model for computational ophthalmology. *npj Digital Medicine*, 8(1):381, 2025. doi: 10.1038/s41746-025-01772-2.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision, 2021.

[7] Y. Zhou et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156--163, 2023. doi: 10.1038/s41586-023-06555-x.

[8] Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator, 2022.

[9] H. Wang et al. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265--5274, 2018. doi: 10.1109/CVPR.2018.00552.

[10] W. Liu et al. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738--6746, 2017. doi: 10.1109/CVPR.2017.713.

[11] Maria Herrero. APTOS-2019 Dataset. Kaggle dataset, 2019. URL https://www.kaggle.com/datasets/mariaherrerot/aptos2019. Accessed: 2026-05-20.

[12] M. N. Bajwa et al. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *International Joint Conference on Neural Networks*, 2020. doi: 10.1109/IJCNN48605.2020.9207664.

[13] Z. Zhang et al. Origa-light: An online retinal fundus image database for glaucoma analysis. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3065--3068, 2010. doi: 10.1109/IEMBS.2010.5626137.

[14] Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025.

[15] Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):AIoa2300221, 2024.

[16] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples. In *International Conference on Learning Representations*, 2017.

[17] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[18] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21464--21475, 2020.

[19] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 677--689, 2021.

[20] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 144--157, 2021.

[21] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2023.

[22] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4921--4930, 2022.

[23] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[24] Y. Sun et al. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of Machine Learning Research*, volume 162, pages 20827--20840, 2022.

[25] Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-based out-of-distribution detection: A comparative study and analysis. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging Workshop, MICCAI*, 2021.

[26] Jasper Linmans, Jeroen van der Laak, and Geert Litjens. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. *Medical Imaging with Deep Learning (MIDL)*, 2020.

[27] Camila González, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. 2021.

[28] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[29] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[30] Y. Chen et al. Alternating training-based label smoothing enhances prompt generalization, 2025.

[31] J. Mukhoti et al. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 15288--15299, 2020.

[32] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[33] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[34] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[35] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C. Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[37] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268--9277, 2019.

[38] S. Y. Alaba et al. Weighted neural network for imbalanced information with undersampling. *Array*, 29:100668, 2026. doi: 10.1016/j.array.2025.100668.

[39] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019.

[40] A. Davila et al. Bio-inspired fine-tuning for selective transfer learning. *IEEE Access*, 2026. doi: 10.1109/ACCESS.2025.3587524.

[41] H. Ren et al. Learning label smoothing for text classification. *PeerJ Computer Science*, 10:e2005, 2024. doi: 10.7717/peerj-cs.2005.

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825--2830, 2011.

[43] L. Tao et al. Confidence calibration under ambiguous ground truth, 2026.

[44] K. Zhu et al. Weighted brier score. *Statistical Biosciences*, 2025. doi: 10.1007/s12561-025-09505-5.

[45] N. Madhusudhan et al. Evaluating abstention in multimodal reasoning systems, 2026.

[46] L. Doorenbos et al. Non-linear outlier synthesis for ood detection, 2024.

[47] J. Ara et al. Balanced datasets and ensemble methods for heart disease classification. *Scientific Reports*, 16:11706, 2026. doi: 10.1038/s41598-026-47691-4.