Knowing When Not to Predict: Self Supervised Learning and Abstention for Safer DR Screening

Muskaan Chopra[1,4] , **Lorenz Sparrenberg**[1,4] , **Jan H. Terheyden**[2] , **Rafet Sifa**[1,3,4]
[1]Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
[2]University Hospital Bonn - Department of Ophthalmology, Germany
[3]Fraunhofer IAIS, Sankt Augustin, Germany
[4]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

Abstract

Self-supervised learning (SSL) is now a standard way to pretrain medical image models, but performance is still mostly judged by downstream accuracy. For safety-critical screening tasks such as diabetic retinopathy grading, this is not enough: a model must also know when its predictions are unreliable and defer uncertain cases for clinical review. In this work, we examine how the length of SSL pretraining influences calibrated confidence and confidence-based abstention. We evaluate multiple SSL checkpoints under a fixed fine-tuning protocol and assess calibrated confidence, coverage, selective accuracy, and selective macro-F1. Across datasets and data regimes, SSL pretraining improves selective prediction compared to training from scratch. Unlike prior SSL studies that primarily evaluate downstream accuracy or AUROC, we analyze how SSL pretraining duration influences confidence behavior under calibrated confidence-based abstention. However, once accuracy saturates, selective performance can still change markedly across checkpoints, and longer pretraining does not consistently improve reliability. These results underscore the importance of abstention-aware evaluation and suggest that pretraining length should be treated as an important reliability-related design choice rather than only a computational detail. Code is available at https://github.com/muskaan712/ijcai-knowing-when-not-to-predict.

1 Introduction

Diabetic retinopathy (DR) screening is a safety-critical clinical task. In a screening workflow, a model that makes a confident but incorrect prediction can be more harmful than a model that defers uncertain cases to expert review. In real-world settings, screening systems must handle non-gradable images, borderline disease stages, and variability in labeling and grading protocols [Gulshan *et al.*, 2016; Krause *et al.*, 2017; Tsiknakis *et al.*, 2021]. These factors make reliability, not only average accuracy, essential for clinical deployment.

Self-supervised learning (SSL) has become a common approach for representation learning in medical image analysis, particularly when labeled data are limited [Albelwi, 2022; Chen *et al.*, 2020; He *et al.*, 2020; Grill *et al.*, 2020; Zbontar *et al.*, 2021; Bardes *et al.*, 2022]. In ophthalmology and fundus-based DR grading, SSL and transfer learning are widely used to improve performance under limited annotation budgets [Li *et al.*, 2017; Tsiknakis *et al.*, 2021]. Despite this adoption, evaluation in medical imaging studies still largely centers on downstream accuracy or AUROC [Wang *et al.*, 2021]. Such metrics provide an incomplete picture for screening applications, where uncertainty and referral decisions play a central role.

Selective prediction and calibration provide a more deployment-aligned evaluation paradigm. Under selective prediction, the model predicts only for high-confidence cases and abstains on uncertain inputs [Geifman and El-Yaniv, 2017], while calibration aligns predicted confidence with empirical correctness [Guo *et al.*, 2017]. These components are often applied post hoc to a trained model, and representation learning choices are rarely examined through the lens of selective behavior, despite their direct influence on confidence distributions.

A frequently overlooked design aspect is SSL pretraining length. In practice, the number of pretraining epochs is often set by compute budgets or marginal gains in downstream accuracy after performance saturates. Whether this choice also affects reliability properties such as calibration and selective prediction remains unclear, even though they are central to screening workflows.

In this work, we study SSL pretraining length as a reliability factor and analyze how the duration of pretraining affects downstream confidence behavior and selective prediction in DR screening. We evaluate checkpoints spanning early, intermediate, and late pretraining stages, fine-tune each under an identical supervised protocol, and assess both standard performance and reliability-aware metrics under confidence-based abstention [Geifman and El-Yaniv, 2017; Guo *et al.*, 2017].

We use APTOS as the primary benchmark [Tsiknakis *et al.*, 2021] and additionally evaluate transfer to Messidor and a 7-class fundus dataset using the same pipeline. To isolate representation-level effects on confidence behavior, CAM-based refinement [Zhou *et al.*, 2015] is excluded from selec-

tive prediction experiments.

Our main finding is that longer SSL pretraining does not monotonically improve reliability. While downstream accuracy often saturates early, selective macro-F1 and confidence behavior can vary substantially across checkpoints, and in some cases selective performance decreases at clinically relevant operating points. These effects are not apparent from accuracy alone and emerge only under selective prediction, which better reflects real screening use.

Contributions. We make three contributions. First, we analyze how SSL pretraining duration affects reliability-aware evaluation in diabetic retinopathy screening. Second, we evaluate checkpoints using selective prediction metrics, including coverage, selective accuracy, and selective macro-F1. Third, using fixed downstream splits, calibration, and training settings, we isolate representation-level effects on confidence behavior.

2 Related Work

Our work bridges two areas that are often studied separately: self-supervised representation learning for medical imaging, and reliability-aware evaluation via calibration and selective prediction. We summarize relevant work and position our contribution in diabetic retinopathy (DR) screening.

2.1 Self-Supervised Learning in Medical Imaging

Self-supervised learning (SSL) learns visual representations from unlabeled data. Contrastive methods such as SimCLR and MoCo use positive pairs from augmented views and negatives from other images [Chen *et al.*, 2020; He *et al.*, 2020], while non-contrastive methods avoid explicit negatives through architectural asymmetry or redundancy reduction, for example BYOL, Barlow Twins, and VICReg [Grill *et al.*, 2020; Zbontar *et al.*, 2021; Bardes *et al.*, 2022]. Surveys review these families and their medical imaging adaptations [Albelwi, 2022; Wang *et al.*, 2021].

SSL has been applied across medical imaging modalities, including histopathology, X-ray imaging, and retinal fundus photography [Ciga *et al.*, 2022; Shen *et al.*, 2019; Li *et al.*, 2021]. In ophthalmology, SSL is widely used for DR grading under limited labels and domain shift [Li *et al.*, 2017; Tsiknakis *et al.*, 2021]. However, evaluation still largely focuses on downstream accuracy or AUROC rather than reliability under selective prediction.

2.2 Calibration and Selective Prediction

Confidence calibration aims to align predicted probabilities with empirical correctness [Guo *et al.*, 2017]. Selective prediction allows models to abstain on low-confidence inputs, trading coverage for reduced risk [Geifman and El-Yaniv, 2017]. Learned abstention methods such as SelectiveNet optimize coverage jointly with prediction [Geifman and El-Yaniv, 2019], but add modeling and optimization complexity. A widely used baseline is confidence-threshold abstention, often applied after calibration, following the classical reject-option rule [Chow, 2003; Franc *et al.*, 2023].

In medical imaging, selective prediction is closely related to human-AI collaboration and referral workflows, including DR screening systems that defer uncertain or ungradable cases [De Fauw *et al.*, 2018]. Still, many studies treat abstention as a post hoc decision layer, and the interaction between representation learning choices and resulting confidence distributions is not analyzed in depth [Angelopoulos *et al.*, 2024; Cattelan and Silva, 2023].

2.3 Positioning of This Work

We treat selective prediction performance as an emergent downstream property influenced by representations learned during SSL pretraining. Abstention is not part of pretraining in our setup. We sample checkpoints at different pretraining lengths, fine-tune each checkpoint under an identical downstream protocol, and evaluate confidence-threshold abstention using calibrated confidence and selective metrics. Rather than focusing primarily on algorithmic novelty in SSL or abstention mechanisms, we study how SSL pretraining duration affects calibration and risk-coverage behavior in DR screening, addressing an evaluation gap that is directly relevant to deployment.

3 Methodology

We study how the length of self-supervised pretraining influences downstream confidence behavior and selective prediction in diabetic retinopathy screening. To isolate the effect of representation learning, we adopt a controlled protocol in which only the SSL checkpoint varies, while downstream training, calibration, and evaluation settings are kept fixed.

We consider two self-supervised objectives under a common pipeline: SiCoVa (Self-Informed Cross-Correlation for Variance and Invariance), a hybrid non-contrastive objective combining variance, invariance, and cross-correlation regularization, and a contrastive triplet-loss baseline.

3.1 Pretraining Setup

Self-supervised pretraining is performed on unlabeled EyePACS fundus images. From each input image x, we generate two stochastic views $v_1 = t_1(x)$ and $v_2 = t_2(x)$ using standard augmentations, including random cropping, horizontal flipping, color jittering, and Gaussian blur. A lightweight Jigsaw perturbation [Park and Ryu, 2024] is applied at the input to encourage patch-level discrimination and fine-grained spatial sensitivity.

Each view is processed by a shared encoder f_θ, implemented as either a ResNet50 [He *et al.*, 2016] or a ViT-B/16 [Dosovitskiy, 2020], followed by a three-layer MLP projector g_ϕ. For any view v, this produces an embedding

$$z = g_\phi(f_\theta(v)) .$$

For a mini-batch of size N, we write the paired embeddings as $Z, Z' \in \mathbb{R}^{N \times D}$, where the i-th row corresponds to the i-th image in the batch, for $i \in \{1, \dots, N\}$. Concretely, we denote the embeddings of the two views of image x_i by

$$z_i = g_\phi(f_\theta(v_{1,i})) , \qquad z'_i = g_\phi(f_\theta(v_{2,i})) .$$

During SSL training, checkpoints are saved at regular intervals spanning early, intermediate, and late stages of convergence. Since all checkpoints share the same architecture

and pretraining configuration, observed downstream differences are primarily associated with pretraining length, although some variability due to stochastic optimization and fine-tuning randomness is expected.

3.2 Non-Contrastive SSL with SiCoVa

SiCoVa is a non-contrastive objective that encourages informative, non-collapsing, and decorrelated representations without using negative samples. It combines intra-view variance and covariance regularization with inter-view invariance and cross-correlation alignment. The objective integrates principles used in VICReg [Bardes *et al.*, 2022] and Barlow Twins [Zbontar *et al.*, 2021], within a unified formulation that jointly optimizes intra-view and inter-view criteria. We weight the different regularization terms using non-negative scalars λ_{intra}, λ_{inv}, and λ_{corr}, which control their relative contribution to the total loss.

Intra-view regularization. We apply variance and covariance regularization independently to each view. Let $Z^{(j)} \in \mathbb{R}^N$ denote the j-th embedding dimension across the batch, and let $\mathrm{Std}(Z^{(j)})$ denote its empirical standard deviation. The variance term enforces a minimum standard deviation $\gamma > 0$:

$$\mathcal{L}_{\text{var}}(Z) = \frac{1}{D}\sum_{j=1}^{D} \max\Big(0,\ \gamma - \mathrm{Std}(Z^{(j)})\Big), \tag{1}$$

where the standard deviation is computed with a small constant added inside the square root for numerical stability.

The covariance regularization penalizes correlations between different embedding dimensions. Let $C_Z \in \mathbb{R}^{D\times D}$ denote the sample covariance matrix of Z (computed after centering each dimension). The covariance loss is:

$$\mathcal{L}_{\text{cov}}(Z) = \frac{1}{D}\sum_{i\neq j} (C_Z)_{ij}^2. \tag{2}$$

We apply these terms to both views and sum:

$$\mathcal{L}_{\text{intra}}(Z, Z') = \mathcal{L}_{\text{var}}(Z) + \mathcal{L}_{\text{var}}(Z') + \mathcal{L}_{\text{cov}}(Z) + \mathcal{L}_{\text{cov}}(Z'). \tag{3}$$

Inter-view regularization. To enforce invariance across views, we minimize the mean squared distance between paired embeddings:

$$\mathcal{L}_{\text{inv}}(Z, Z') = \frac{1}{N}\sum_{i=1}^{N} \|z_i - z_i'\|_2^2. \tag{4}$$

To promote alignment of corresponding embedding dimensions while discouraging redundancy across dimensions, we use a cross-correlation criterion. Let $\tilde{Z}$ and $\tilde{Z}'$ denote column-wise centered and l_2-normalized embeddings, and define the cross-correlation matrix:

$$\mathcal{R} = \frac{1}{N}\tilde{Z}^\top \tilde{Z}' \in \mathbb{R}^{D\times D}. \tag{5}$$

The correlation loss is:

$$\mathcal{L}_{\text{corr}}(Z, Z') = \sum_i \big(1 - \mathcal{R}_{ii}\big)^2 + \sum_{i\neq j} \mathcal{R}_{ij}^2. \tag{6}$$

Overall objective. The SiCoVa objective is a weighted sum of intra-view and inter-view terms:

$$\begin{aligned}\mathcal{L}_{\text{SiCoVa}} = \lambda_{\text{intra}}\mathcal{L}_{\text{intra}}(Z, Z') + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}(Z, Z') \\ + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}}(Z, Z').\end{aligned} \tag{7}$$

All weights are fixed across experiments.

3.3 Contrastive Baseline: Triplet Loss

As a contrastive baseline, we use a batch-all triplet loss [Hermans *et al.*, 2017]. For each anchor embedding z_i from Z, the positive sample is the corresponding embedding z_i' from Z', and all other embeddings z_k' with $k \neq i$ act as negatives. With margin $m > 0$ and hinge function $[\cdot]_+ = \max(\cdot, 0)$, the loss is:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{k\neq i}\Big[\|z_i - z_i'\|_2 - \|z_i - z_k'\|_2 + m\Big]_+. \tag{8}$$

This objective emphasizes inter-instance discrimination and serves as a contrastive reference to SiCoVa.

3.4 Downstream Fine-Tuning

Each SSL checkpoint is fine-tuned on labeled diabetic retinopathy data using an identical supervised protocol. We use fixed training, validation, and calibration splits across all experiments to ensure fair comparison. The downstream task is multi-class DR severity classification.

Class activation map (CAM) based refinement is applied during downstream fine-tuning to improve localization and raw classification performance. However, CAM refinement is *not* used when evaluating selective prediction and abstention. This separation ensures that abstention behavior and confidence-based deferral are driven by the learned representations and calibrated probabilities, rather than by additional spatial regularization introduced during fine-tuning.

3.5 Calibration and Selective Prediction

After fine-tuning, predicted class probabilities are calibrated using temperature scaling on a held-out calibration split [Guo *et al.*, 2017]. Calibration is performed independently for each SSL checkpoint using identical data and procedures. In particular, we fit a single temperature parameter on the calibration split and apply it to logits at inference time to obtain calibrated class probabilities.

Selective prediction is implemented via confidence-threshold abstention [Geifman and El-Yaniv, 2017] as a post hoc decision rule, and does not affect training. For an input x, let $p_{\max}(x) = \max_k p(y = k \mid x)$ denote the maximum calibrated predicted probability. Given a threshold τ, the model predicts if $p_{\max}(x) \geq \tau$ and abstains otherwise. Coverage is defined as the fraction of samples that are retained (not abstained on).

To support comparisons across checkpoints and methods, we report selective results at a fixed coverage of 70%. This operating point was chosen as a clinically realistic trade-off that retains the majority of predictions while still allowing substantial abstention of uncertain cases for expert review. Concretely, for each checkpoint, we sweep a fixed grid of

Dataset	Role	Classes	Images
EyePACS [Kag, 2015]	SSL pretraining	–	35,000
APTOS-19 [Karthik *et al.*, 2019]	Fine-tuning	5	3,662
Messidor [Mes, 2014]	Fine-tuning	4	1,200
7-class Fundus [C Benítez *et al.*, 2021]	Fine-tuning	7	757

Table 1: Datasets used for pretraining, fine-tuning, and evaluation.

thresholds τ and select the operating point whose coverage is closest to 70%, then report selective accuracy and selective macro-F1 at that operating point. We additionally observe qualitatively similar trends across neighboring operating points in the risk-coverage analysis.

4 Experimental Setup

This section describes the controlled experimental setup used to isolate the effect of SSL pretraining length on confidence behavior and selective prediction.

4.1 Datasets

We study diabetic retinopathy (DR) severity classification using retinal fundus images. Our primary evaluation benchmark is APTOS-19, a widely used dataset in the DR literature that provides relatively consistent annotations compared to other public benchmarks [Tsiknakis *et al.*, 2021]. APTOS-19 contains five disease severity levels, ranging from no DR to proliferative DR.

For representation learning, models are pretrained on EyePACS using unlabeled images only. EyePACS provides large-scale and diverse retinal data and is commonly used for learning general-purpose fundus representations. All downstream fine-tuning and primary evaluation are performed on APTOS-19.

To assess robustness and generalization, we additionally evaluate performance on Messidor and a 7-class fundus dataset. These datasets are not used during self-supervised pretraining. For downstream evaluation, we fine-tune and evaluate models on each dataset using the same controlled protocol and labeled-data regimes as for APTOS-19.

Table 1 summarizes the datasets used in our experiments.

4.2 Data Splits and Training Protocol

For downstream training, we use fixed training, validation, and calibration splits for all SSL checkpoints. The calibration split is used exclusively for temperature scaling, while the validation split is used for evaluation both with and without abstention. Identical splits are used across all experiments to ensure fair comparison across pretraining lengths and methods.

We evaluate multiple labeled data regimes by varying the fraction of APTOS-19 training data used for fine-tuning. This allows us to assess whether the impact of SSL pretraining length on reliability persists under limited supervision, which is common in medical imaging settings.

4.3 Evaluation Metrics

We report standard downstream performance without abstention using accuracy and macro-F1. To assess reliability under selective prediction, we additionally report selective accuracy and selective macro-F1 after applying confidence-threshold abstention. Coverage is defined as the fraction of samples for which the model produces a prediction (i.e., not abstained on). For fair comparison across checkpoints and methods, we report selective results at a fixed coverage of 70%.

Macro-averaged metrics are essential in diabetic retinopathy screening due to class imbalance and severity-dependent rejection effects. In particular, selective prediction changes the evaluated set and can skew class frequencies, making accuracy alone potentially misleading. We therefore use selective macro-F1 as the primary metric for ranking checkpoints at the target coverage.

Beyond aggregate scores, we analyze coverage–risk trends and class-wise rejection patterns to characterize how pretraining length reshapes deployment-relevant trade-offs and abstention behavior across disease stages.

4.4 Implementation Details

All experiments use the same architecture and optimization settings across SSL checkpoints. Training and evaluation are performed on a single NVIDIA A100 GPU. Self-supervised pretraining requires approximately 10 hours per model, while downstream fine-tuning takes approximately 1.25 hours per model. At inference time, the abstention mechanism introduces negligible computational overhead, as it relies solely on confidence-based post-processing.

5 Results

We now present experimental results analyzing the effect of self-supervised pretraining length on downstream performance and reliability in diabetic retinopathy screening. We first examine standard classification performance without abstention, and then analyze selective prediction behavior under a fixed operating point.

5.1 Downstream Performance

We begin by analyzing downstream classification performance as a function of SSL pretraining length. Figure 1 shows downstream macro-F1 across SSL checkpoints for SiCoVa and Triplet, evaluated at downstream fine-tuning epochs 50 and 200 on APTOS-19 dataset.

Across methods, macro-F1 improves rapidly during early stages of pretraining and then saturates. Extending pretraining beyond this point yields marginal or inconsistent gains, and in some cases mild degradation. This trend is consistent across both fine-tuning durations, indicating that longer downstream training does not compensate for saturation in representation learning.

These observations align with prior work showing that SSL representations converge early with respect to downstream

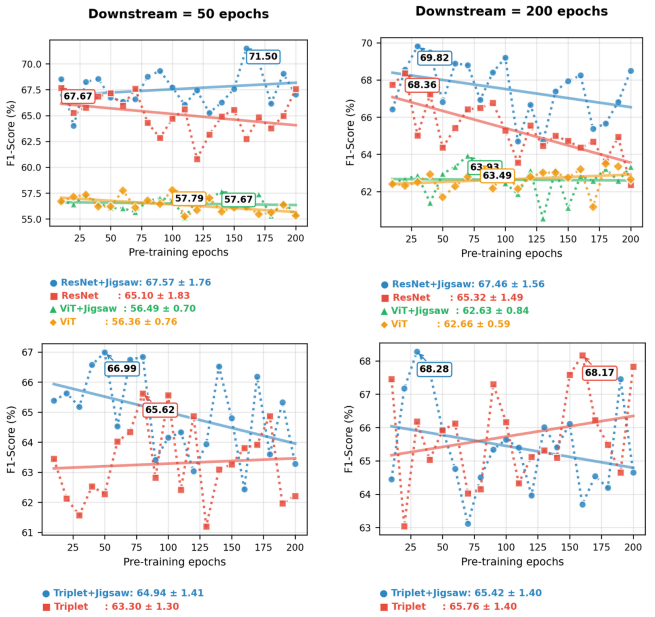

Figure 1: Downstream macro-F1 across SSL pretraining checkpoints for SiCoVa and Triplet, shown for two fine-tuning durations (epoch 50 and epoch 200).

accuracy [Truong *et al.*, 2021]. While some checkpoint-level variability is expected due to stochastic fine-tuning, we focus next on whether pretraining length systematically affects reliability.

Table 2 reports downstream performance of SiCoVa across datasets and labeled data regimes. While accuracy and macro-F1 improve with increased supervision, gains diminish at higher data fractions, particularly on APTOS.

Dataset	Regime	Acc (%)	Macro-F1 (%)	QWK
APTOS	10%	79.0	53.0	0.86
	50%	81.0	64.0	0.88
	100%	**85.27**	**71.5**	**0.91**
Messidor	10%	54.0	54.0	0.24
	50%	69.0	69.0	0.77
	100%	**71.0**	**71.0**	**0.77**
7-class Fundus	10%	67.0	67.0	0.82
	50%	86.0	86.0	0.95
	100%	**88.0**	**88.0**	**0.97**

Table 2: SiCoVa results across datasets and labeled data regimes. Metrics are computed without abstention.

To contextualize these results, Table 3 compares SiCoVa against Triplet, supervised baselines, and prior approaches on APTOS without abstention. SiCoVa achieves the strongest overall performance, particularly when combined with CAM refinement. We next examine whether these accuracy differences translate to improved reliability under selective prediction.

[1]ClementP fundus grading collection: https://huggingface.co/collections/ClementP/fundus-grading

Method	Backbone	Acc	F1	Prec.	Rec.	QWK
SiCoVa	R50 + J + C	**85.27**	**71.50**	**75.29**	**69.12**	**0.91**
	R50 + J	79.00	57.00	59.00	56.00	0.85
	R50	83.77	67.56	71.60	65.17	0.90
	ViT-B/16 + J	79.67	57.39	68.72	54.50	0.82
	ViT-B/16	79.95	57.79	69.02	54.92	0.83
Triplet	R50 + J + C	83.90	66.74	69.55	65.35	0.84
	R50	84.17	62.53	72.26	62.94	0.87
Supervised	R50	80.27	61.77	64.89	60.07	0.85
	ViT-B/16	83.00	65.00	72.00	62.00	0.89
ClementP[1]	R50	51.00	28.00	35.00	38.00	0.66
	ViT-B/16	69.00	51.00	54.00	54.00	0.88

Table 3: APTOS classification performance across SSL and supervised baselines. (R50 = ResNet50, ViT-B/16 = ViT-base (16×16), J = Jigsaw SSL, C = CAM refinement.)

5.2 Selective Prediction at Fixed Coverage

We next evaluate selective prediction using confidence-based abstention. All selective results are reported at a fixed coverage of 70%, reflecting a realistic screening setting in which uncertain cases are deferred for expert review.

Table 4 summarizes selective performance at this operating point for representative SSL checkpoints (with pretraining epochs shown). While selective accuracy remains high across checkpoints, selective macro-F1 varies substantially, even among models with similar downstream performance. Notably, longer SSL pretraining does not consistently improve selective macro-F1: intermediate checkpoints can outperform later checkpoints despite comparable accuracy, suggesting that extended pretraining reshapes confidence allocation rather than raw predictive performance.

Method	Checkpoint tier	Sel. Acc. (%)	Sel. F1 (%)	QWK
SiCoVa	Best (ep-110)	**94.69**	**78.65**	**0.95**
	Median (ep-150)	93.93	75.57	0.94
	Weak (ep-60)	92.68	68.74	0.94
Triplet	Best (ep-170)	92.31	46.56	0.92
	Median (ep-40)	93.16	37.45	0.93
	Weak (ep-20)	91.52	36.93	0.90

Table 4: Selective prediction performance at 70% coverage.

Figure 2 complements Table 4 by sweeping the confidence threshold to vary coverage and obtain a risk–coverage curve. SiCoVa consistently achieves higher selective macro-F1 across operating points, whereas Triplet is more sensitive to the choice of coverage. The highlighted point corresponds to the 70% coverage operating point used in the table.

5.3 Accuracy versus Macro-Averaged Metrics under Abstention

As confidence thresholds increase, selective accuracy rises sharply due to the removal of ambiguous samples and

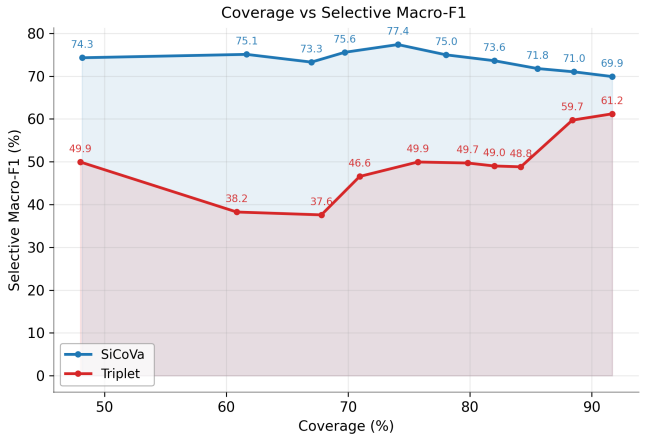

Figure 2: Selective prediction comparison between SiCoVa and Triplet under confidence-based abstention. We sweep the confidence threshold to obtain a risk-coverage curve, and highlight the operating point corresponding to 70% coverage used in Table 4.

minority-class instances. In contrast, selective macro-F1 improves more gradually and often plateaus.

This discrepancy arises because accuracy is dominated by majority classes and benefits from aggressive rejection, whereas macro-F1 penalizes uneven class-wise performance. As a result, selective macro-F1 provides a more informative measure of reliability under abstention for imbalanced screening tasks.

5.4 Representation Structure and Qualitative Analysis

Figure 3 provides qualitative Grad-CAM visualizations across APTOS, Messidor, and the 7-class fundus dataset. Accepted predictions typically exhibit more coherent activation patterns, while rejected cases more often show diffuse or unstable responses. These visualizations serve as qualitative sanity checks rather than localization claims.

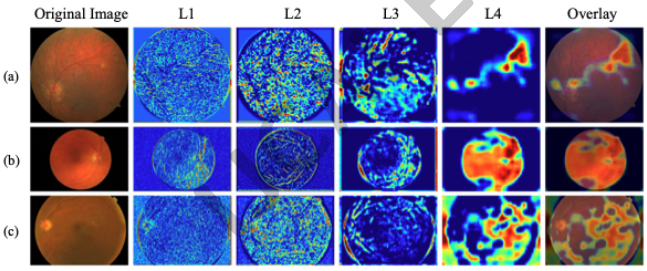

Figure 3: Grad-CAM visualizations for SiCoVa across three downstream datasets, contrasting accepted and rejected cases [Dataset: (a) 7-Class Fundus, (b) Messidor, and (c) Aptos-19].

Figure 4 visualizes representation structure and severity-dependent rejection using PaCMAP and class-wise acceptance statistics. Disease severity classes form a continuous and partially overlapping manifold, reflecting the progressive nature of diabetic retinopathy. After abstention, accepted samples occupy denser and more class-consistent regions, while rejected samples concentrate near inter-class boundaries rather than forming isolated outliers, suggesting that abstention primarily filters semantically ambiguous cases.

5.5 Cross-Dataset Transfer

Finally, we evaluate selective prediction on Messidor and the 7-class fundus dataset. While downstream accuracy remains relatively stable across SSL checkpoints, selective macro-F1 continues to vary, indicating that the impact of pretraining length on confidence behavior generalizes beyond a single benchmark.

Overall, these results demonstrate that SSL pretraining length has a limited effect on downstream accuracy but a pronounced impact on confidence calibration and selective prediction behavior, with direct implications for safe deployment in clinical screening systems.

6 Discussion

Our results demonstrate that the length of self-supervised pretraining plays a meaningful role in shaping confidence behavior and selective prediction performance in diabetic retinopathy screening. While downstream accuracy often saturates early, reliability-aware evaluation reveals substantial variation across SSL checkpoints that is not captured by standard metrics. In this section, we discuss the implications of these findings for representation learning, clinical interpretation, and deployment.

6.1 Why Longer Pretraining Does Not Guarantee Better Reliability

A common assumption in SSL-based medical imaging is that longer pretraining is either beneficial or, at worst, neutral once downstream accuracy has converged. Our results challenge this assumption. Although extended pretraining does not consistently degrade standard downstream performance, it can still reshape confidence distributions, which may affect selective macro-F1 at a fixed coverage. We examine this relationship directly by reporting selective performance across pretraining epochs in the subsequent analysis.

One plausible explanation is that later stages of SSL pretraining emphasize invariances that improve global discrimination while compressing features relevant for borderline or minority classes. In diabetic retinopathy, disease severity evolves along a continuum rather than forming well-separated categories. As a result, representations that appear strong under accuracy-based evaluation may still yield poorly calibrated confidence near class boundaries. These effects remain largely invisible unless selective prediction behavior is explicitly analyzed.

Importantly, our findings do not suggest that longer pretraining is inherently harmful. Rather, they indicate that pretraining length interacts with downstream reliability in non-monotonic ways, and that accuracy alone is insufficient to identify favorable operating points.

6.2 Clinical Interpretation of Accepted and Rejected Cases

To assess the clinical plausibility of abstention decisions, a clinician inspected a subset of accepted and rejected pre-

Figure 4: Representation structure and severity-dependent abstention using PaCMAP and class-wise acceptance statistics at 70% coverage.

dictions. Rejected cases often exhibited factors known to complicate automated grading, including decentration, shadowing, media opacities, and confounding retinal conditions, consistent with criteria for non-gradable images in screening workflows. The clinician also noted occasional inconsistencies within the accepted set and a small number of rejected images of acceptable quality. This underscores that confidence-based abstention reflects model uncertainty, not clinical ground truth or image quality, and may conservatively reject borderline but usable cases. Overall, abstention should be interpreted as decision support rather than a replacement for expert judgment [De Fauw *et al.*, 2018].

6.3 Accuracy versus Reliability in Screening Systems

Our analysis highlights limitations of accuracy as a deployment-oriented metric. Under selective prediction, accuracy can increase as uncertain samples are rejected even when performance on clinically relevant minority or borderline classes degrades. Macro-averaged metrics expose this effect by penalizing uneven class-wise performance and severity-dependent rejection. Selective macro-F1 therefore, provides a more balanced assessment of reliability under abstention and aligns better with equitable performance across severity levels.

6.4 Implications for Deployment and Evaluation

First, SSL pretraining length should be treated as a design choice with direct consequences for confidence behavior. While selective performance varies across checkpoints, we do not observe a clear monotonic trend with pretraining length, motivating further experiments. Second, benchmarking SSL methods solely on downstream accuracy can miss clinically relevant differences in uncertainty handling. Third, confidence-based abstention is a simple deferral mechanism, but its behavior should be assessed across operating points and disease severities.

7 Conclusion

We study how self-supervised pretraining length influences reliability and selective prediction in diabetic retinopathy screening. Although downstream accuracy saturates early, confidence behavior and selective macro-F1 vary substantially across SSL checkpoints, affecting deployment decisions under abstention. These findings motivate reliability-aware evaluation and treating pretraining length as a primary design choice.

Limitations. We focus on a single SSL framework and a limited set of architectures, and the clinical inspection is qualitative and based on a small sample. We rely on confidence-based abstention rather than learned deferral. In addition, we do not perform statistical significance testing across checkpoints or repeated fine-tuning runs. Future work should extend to additional SSL objectives and architectures, analyze reliability trends across multiple random seeds, and include larger-scale human-in-the-loop evaluation.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

[Albelwi, 2022] Saleh Albelwi. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022.

[Angelopoulos *et al.*, 2024] Anastasios N Angelopoulos, Stuart Pomerantz, Synho Do, Stephen Bates, Christopher P Bridge, Daniel C Elton, Michael H Lev, R Gilberto González, Michael I Jordan, and Jitendra Malik. Conformal triage for medical imaging ai deployment. *medRxiv*, pages 2024–02, 2024.

[Bardes *et al.*, 2022] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022.

[C Benítez *et al.*, 2021] Veronica Elisa C Benítez, Ingrid Castro Matto, Julio César Mello Román, José Luis Vázquez Noguera, Miguel García-Torres, Jordan Ayala, Diego P. Pinto-Roa, Pedro E. Gardel-Sotomayor, Jacques Facon, and Sebastian Alberto Grillo. Dataset from fundus images for the study of diabetic retinopathy (0.2), 2021.

[Cattelan and Silva, 2023] Luís Felipe P Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. *arXiv preprint arXiv:2305.15508*, 2023.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[Chow, 2003] Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 2003.

[Ciga *et al.*, 2022] Olivier Ciga, Tony Xu, and Anne L. Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[De Fauw *et al.*, 2018] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George Van Den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Franc *et al.*, 2023] Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

[Geifman and El-Yaniv, 2017] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 4878–4887, 2017.

[Geifman and El-Yaniv, 2019] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 2019.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[Gulshan *et al.*, 2016] Varun Gulshan, Lily Peng, Marc Coram, Martin Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip Nelson, Jessica Mega, and Dale Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316:2402–2410, 12 2016.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Kag, 2015] Kaggle eyepacs diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection, 2015.

[Karthik *et al.*, 2019] Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. Kaggle Competition, 2019. Accessed 2026-01-14.

[Krause *et al.*, 2017] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg Corrado, Lily Peng, and Dale Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125, 10 2017.

[Li *et al.*, 2017] Xiaogang Li, Tiantian Pang, Biao Xiong, Weixiang Liu, Ping Liang, and Tianfu Wang. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, pages 1–11. IEEE, 2017.

[Li *et al.*, 2021] Zekun Li, Wei Zhao, Feng Shi, Lei Qi, Xingzhi Xie, Ying Wei, Zhongxiang Ding, Yang Gao, Shangjie Wu, Jun Liu, Yinghuan Shi, and Dinggang Shen. A novel multiple instance learning framework for covid-19 severity assessment via data augmentation and self-supervised learning. *Medical Image Analysis*, 69:101978, 2021.

[Mes, 2014] Messidor-2. http://www.adcis.net/en/third-party/messidor2/, 2014.

[Park and Ryu, 2024] Wongi Park and Jongbin Ryu. Fine-grained self-supervised learning with jigsaw puzzles for medical image classification. *Comput. Biol. Med.*, 174(C), May 2024.

[Shen *et al.*, 2019] Tianyu Shen, Chao Gou, Fei-Yue Wang, Zilong He, and Weiguo Chen. Learning from adversarial medical images for x-ray breast mass segmentation. *Computer Methods and Programs in Biomedicine*, 180:105012, 2019.

[Truong *et al.*, 2021] Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. How transferable are self-supervised features in medical image classification tasks? In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 54–74. PMLR, 2021.

[Tsiknakis *et al.*, 2021] Nikos Tsiknakis, Dimitris Theodoropoulos, Georgios Manikis, Emmanouil Ktistakis, Ourania Boutsora, Alexa Berto, Fabio Scarpa, Alberto Scarpa, Dimitrios I. Fotiadis, and Kostas Marias. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. *Computers in Biology and Medicine*, 135:104599, 2021.

[Wang *et al.*, 2021] Jian Wang, Hengde Zhu, Shuihua Wang, and Yudong Zhang. A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26:3518–3535, 2021.

[Zbontar *et al.*, 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.