CHILD: Human-in-the-Loop OOD Detection for Safe Clinical Deployment

Jinlun Ye[1,2,4], Kaiyue Lu[1,2,4], Runhe Lai[1,2,4], Xinhua Lu[1,2,4], Jia-Xin Zhuang[3] (✉), and Ruixuan Wang[1,2,4] (✉)

[1] School of Computer Science and Engineering, Sun Yat-sen Univerisity, Guangzhou,China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Hong Kong University of Science and Technology, Hong Kong, China
[4] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
jzhuangad@connect.ust.hk, wangruix5@mail.sysu.edu.cn

Abstract. Out-of-distribution (OOD) detection is critical for safe deployment of medical AI systems. Recently, test-time adaptation (TTA) has emerged as a new paradigm for OOD detection, automatically adjusting detector behavior during deployment. However, such automatic adaptation mechanisms may raise safety concerns in safety-critical clinical environments. While physician oversight can mitigate these risks, it is resource-intensive and must be judiciously allocated. To reconcile safety with efficiency, we propose CHILD, a training-free framework designed to enhance streaming OOD detection via sparse human feedback. Operating under strict budget constraints, CHILD employs an adaptive risk-aware sample selection mechanism to pinpoint only the most decision-uncertain samples for review. Crucially, it maximizes the utility of this sparse feedback through a retrieval-based score calibration module, which refines model predictions using a compact feature cache without any parameter updates. Extensive experiments on four medical benchmarks demonstrate that CHILD turns limited supervision into significant reliability gains: with a sparse feedback budget of only 5%, it reduces the average FPR95 from 72.63% to 60.26% and improves AUROC from 75.53% to 81.85%, consistently outperforming state-of-the-art baselines. Our code is publicly available at https://github.com/figec/CHILD.

Keywords: OOD Detection · Human-in-the-Loop · Clinical Deployment.

1 Introduction

Deep learning models demonstrate strong performance in medical diagnosis when test samples belong to the known disease categories seen during training (in-distribution, ID). However, real-world clinical deployment inevitably exposes such systems to unseen disease categories, namely out-of-distribution (OOD) samples, where models may produce overconfident yet incorrect predictions [8,

17], posing significant clinical risks. Consequently, OOD detection is critical for the safe deployment of medical artificial intelligence.

Existing research on OOD detection can be broadly categorized into two conventional paradigms. One line of work focuses on post-hoc scoring [5, 7, 8, 12–14, 22], where OOD scores are derived from model outputs or intermediate feature representations without modifying the training procedure. Another line explores training-time OOD-aware learning [3,4,15,16,24,26], aiming to improve model reliability by regularization strategies or synthetic OOD exposure during training. More recently, TTA [2,6,25,28,29,31] has emerged as a new paradigm. By leveraging unlabeled test data observed during deployment, TTA methods adjust model behavior at inference time and have been shown to achieve better OOD detection performance compared to static methods.

However, directly applying TTA methods to clinical deployment raises safety concerns. TTA methods adapt detection behavior automatically based solely on unlabeled test data, leading to unpredictable decision adjustments and progressive error accumulation [30]. At the same time, physician review of uncertain or high-risk cases is not an additional burden but an integral component of routine workflows. Such naturally available and reliable clinical supervision motivates a safer strategy: instead of relying entirely on automatic adaptation, selectively incorporating limited physician feedback can guide decision adjustment at inference time in a controlled and clinically aligned manner. Recent work [1] has explored the use of human feedback for OOD detection through additional training, but assumes offline access to the entire test set for feedback allocation and decision adjustment. In real clinical deployment, data arrive sequentially in a streaming manner, and decisions must be made online using only past observations, rendering offline methods impractical.

Therefore, we propose **C**linical **H**uman-**I**n-the-**L**oop **D**ecision (**CHILD**), a training-free human-in-the-loop OOD detection framework motivated by clinical deployment. CHILD consists of two components: (i) *Risk-Aware Sample Selection*, which adaptively identifies decision-uncertain regions in the score space and allocates limited physician feedback to samples with the highest potential clinical risk; and (ii) *Score Calibration*, which maintains compact feature caches of physician-reviewed ID/OOD samples and adjusts OOD scores via similarity-driven calibration, without any parameter updates. Our contributions are summarized as follows:

1. We formulate a streaming human-in-the-loop OOD detection setting with strict feedback budget constraints, capturing the sequential decision-making requirements relevant to clinical deployment.
2. We propose CHILD, a training-free and plug-and-play framework that selectively queries physician feedback for high-risk samples and incorporates it through lightweight score-level calibration.
3. Extensive experiments on four medical benchmarks demonstrate that CHILD consistently improves performance with minimal feedback, reducing FPR95 by over 12% on average when combined with a state-of-the-art detector.

2 Method

2.1 Preliminaries and Problem Setup

OOD Detection. Given a test sample $\mathbf{x}_t$, an OOD detector $f(\cdot)$ produces a scalar score $s_t = f(\mathbf{x}_t)$. OOD detection is typically performed by comparing the score s_t with a decision threshold λ, i.e.,

$$G(\mathbf{x}_t) = \begin{cases} \text{ID}, & s_t \geq \lambda, \\ \text{OOD}, & s_t < \lambda. \end{cases} \tag{1}$$

Human-in-the-Loop Streaming OOD Detection Setting. In this work, we consider a deployment setting where test samples arrive sequentially as a stream $\{\mathbf{x}_t\}_{t=1}^{T}$, with ID and OOD samples mixed and unlabeled. During deployment, the system is allowed to request a limited amount of human feedback under a strict budget constraint. Specifically, at most $\rho\%$ of test samples can be reviewed by physicians. Rather than requiring physicians to directly determine the model-defined ID/OOD classes, they provide a diagnostic category or an intended-scope judgment, which is then mapped to ID/OOD according to the predefined set of ID classes associated with the model's intended use. In our experiments, benchmark annotations serve as a controlled proxy for such physician-confirmed feedback. We use $q_t \in \{0, 1\}$ to indicate whether physician feedback is requested for sample x_t, and enforce the budget constraint

$$\sum_{t=1}^{T} q_t \leq B, \tag{2}$$

where $B = \lfloor \rho T \rfloor$ denotes the total feedback budget. Human feedback is acquired incrementally as samples arrive, and the system has no access to the complete test set in advance.

2.2 Overview

Under the streaming OOD detection setting, we propose **C**linical **H**uman-**I**n-the-**L**oop **D**ecision (**CHILD**), a training-free human-in-the-loop framework designed for clinical deployment. As illustrated in Figure 1, CHILD operates on top of an arbitrary base OOD detector and consists of two interleaved components that operate jointly as each sample arrives. The first component, *Risk-Aware Sample Selection*, allocates limited physician feedback to decision-uncertain samples that are associated with higher potential clinical risk. The second component, *Score Calibration*, incorporates the obtained ID/OOD feedback as decision-time priors to calibrate the outputs of the base OOD detector during inference.

2.3 Risk-Aware Sample Selection

Under a limited feedback budget, physician review should be allocated to samples that are most likely to reduce high-risk errors in a streaming environment. To this

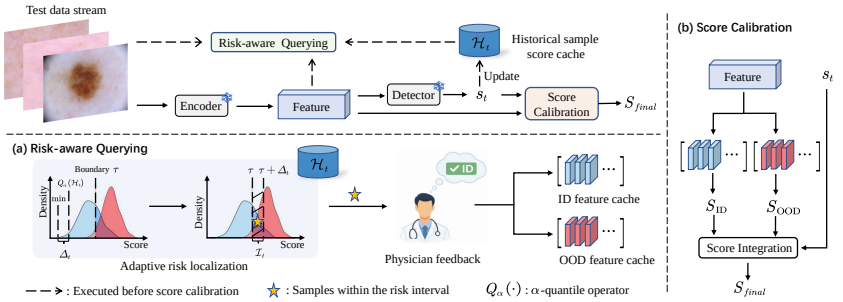

Fig. 1: Overview of the proposed framework. During deployment, samples falling within the adaptive risk interval are queried for feedback. The collected feedback is stored and subsequently used to calibrate the OOD score of incoming samples.

end, the proposed module maintains an online estimate of the score distribution and localizes high-risk regions for targeted feedback allocation (detailed below).

Online Score Accumulation. During deployment, the system continuously records OOD scores produced by the base detector as test samples arrive. Let $\mathcal{H}_t = \{s_i\}_{i=1}^{t}$ denote the set of accumulated historical scores up to time step t. These scores provide an evolving empirical estimate of the score distribution under the current deployment environment and form the basis for identifying decision-uncertain and potentially high-risk regions.

Adaptive Risk Sample Localization. In medical OOD detection, the most critical failure mode arises when OOD samples are misclassified as ID, a situation that typically occurs near regions of decision uncertainty where the base detector struggles to separate ID from OOD samples. Compared with confirming high-confidence samples far from the decision boundary, allocating limited human feedback to such decision-uncertain samples is more likely to influence system decisions and achieve greater risk reduction.

Accordingly, we aim to localize decision-uncertain regions along the one-dimensional OOD score axis by estimating a time-evolving reference boundary τ_t. The accumulated OOD scores form an empirical one-dimensional distribution, from which a boundary separating lower- and higher-score regions can be estimated directly in score space. To this end, the Otsu [19] thresholding method is adopted. For a candidate threshold τ, the score distribution is divided into two groups, and the between-class variance $\sigma_b^2(\tau)$ quantifies the separation between their mean scores. The threshold that maximizes $\sigma_b^2(\tau)$ is selected as the reference boundary, which typically lies between distinct score regimes, making it suitable for approximating the current ID/OOD decision boundary:

$$\tau_t = \arg\max_{\tau} \ \sigma_b^2(\tau)\,. \tag{3}$$

The boundary τ_t is continuously updated as streaming data arrive, allowing the estimate to adapt to distributional changes during deployment. To operationalize

feedback querying, τ_t is used as a reference point and define a risk query interval around it. To ensure that the width of the risk interval adapts to the evolving score distribution, the width term Δ_t is defined as a function of $\mathcal{H}_t$, i.e.,

$$\Delta_t = Q_\alpha(\mathcal{H}_t) - \min(\mathcal{H}_t)\,, \tag{4}$$

where $Q_\alpha(\cdot)$ denotes the α-quantile operator, and $\alpha \in (0, 1)$ is typically chosen to be close to 0. This formulation defines the interval width with respect to the lower extreme of the score distribution while using a quantile-based statistic to provide a robust and adaptive characterization of the distributional spread over time. In medical OOD deployment, risk is asymmetric in score-based decisions, as the most critical failure occurs when OOD samples are incorrectly accepted as ID. We then construct an asymmetric risk interval $\mathcal{I}_t$, i.e.,

$$\mathcal{I}_t = [\,\tau_t,\ \tau_t + \Delta_t\,]\,, \tag{5}$$

which expands coverage toward the more ID-like side of the boundary compared with a symmetric interval. This design increases the likelihood of capturing samples whose misclassification would lead to more severe clinical consequences under the same feedback budget.

Physician Feedback and Sample Caching. For samples within the risk interval, the system requests physician feedback when the budget permits. The provided diagnostic category or intended-scope judgment is mapped to an ID/OOD label according to the predefined ID class set. Each feedback instance is treated as high-confidence prior information. The feature $\phi(\mathbf{x})$ of a sample mapped to ID or OOD is stored in $\mathcal{C}_{\text{ID}}$ or $\mathcal{C}_{\text{OOD}}$, respectively, for subsequent score calibration. Their total size satisfies $|\mathcal{C}_{\text{ID}}| + |\mathcal{C}_{\text{OOD}}| \leq B$. Throughout this process, no model parameters are updated.

2.4 Score Calibration

Given the feedback caches $\mathcal{C}_{\text{ID}}$ and $\mathcal{C}_{\text{OOD}}$, our goal is to refine OOD decisions at inference time without any test-time training. To achieve this, we perform lightweight score-level calibration that directly leverages similarity to physician-confirmed samples. For a new test sample $\mathbf{x}_t$, the base detector first produces the original OOD score s_t. The maximum cosine similarity between $\mathbf{x}_t$ and each feedback cache is then computed as follows:

$$S_{\text{ID}}(\mathbf{x}_t) = \max_{\phi(\mathbf{x})\in\mathcal{C}_{\text{ID}}} \cos(\phi(\mathbf{x}_t), \phi(\mathbf{x})), \quad S_{\text{OOD}}(\mathbf{x}_t) = \max_{\phi(\mathbf{x})\in\mathcal{C}_{\text{OOD}}} \cos(\phi(\mathbf{x}_t), \phi(\mathbf{x})), \tag{6}$$

Then the similarity difference $\delta_t = S_{\text{ID}}(\mathbf{x}_t) - S_{\text{OOD}}(\mathbf{x}_t)$ can indicate whether the sample is closer to confirmed instances and, therefore, is used as a soft calibration term to adjust the base score. However, in certain cases, a test sample may be highly similar to a previously confirmed sample, indicating strong semantic consistency; in such situations, relying solely on the soft calibration may lead to

insufficient calibration strength. To provide a more decisive adjustment under clear semantics, we introduce a hard-assignment mechanism:

$$y_t = \text{sign}(S_{\text{ID}}(\mathbf{x}_t) - S_{\text{OOD}}(\mathbf{x}_t)), \tag{7}$$

which reflects whether the sample is more similar to confirmed ID or OOD instances. To ensure that hard assignment is applied exclusively under reliable evidence, we further introduce a hard-assignment gate:

$$h_t = \mathbb{1}[\max\{S_{\text{ID}}(\mathbf{x}_t), S_{\text{OOD}}(\mathbf{x}_t)\} \geq \eta], \tag{8}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function that takes value 1 if the condition is satisfied and 0 otherwise, and η is the semantic confidence threshold. The resulting hard-assignment term is thus denoted as $H_t = h_t y_t$. Based on these quantities, the final calibrated score is defined as

$$S_{final} = s_t + \beta\Big(\delta_t + H_t\Big), \tag{9}$$

where β serves as a scaling factor that aligns the magnitude of the similarity-based calibration terms with the base detector score, while also controlling the overall calibration strength. This formulation unifies soft calibration and hard assignment within a single score-level adjustment. Higher similarity to physician-confirmed OOD samples reduces the calibrated score, pushing the decision toward OOD, whereas higher similarity to physician-confirmed ID samples increases the calibrated score, reinforcing the ID decision. In this way, limited human feedback is translated into reliable decision refinement.

3 Experiment

3.1 Experiment Setup

Datasets. Following prior evaluation protocols [10, 18, 27], we evaluate CHILD on SD-198 [21], ISIC 2019 [23], NCT-CRC [9], and BreakHis [20], covering diverse clinical imaging scenarios. For SD-198, following standard practice [11], 40 categories are designated as ID categories (referred to as Skin-40), while the remaining 158 categories are treated as OOD data. For ISIC 2019, NV, MEL, DF, and VASC are selected as ID categories, forming a long-tailed ID setting. This setting is referred to as ISIC-4, while the remaining four categories serve as OOD data. For NCT-CRC and BreakHis, we follow commonly used OOD detection setups with multiple ID/OOD splits [18], and the final performance is reported by averaging results across all splits.

Implementation Details. Unless otherwise specified, all experiments are conducted using a standard image classifier trained on the ID training set with cross-entropy loss. During deployment, the human feedback budget is set to $\rho = 5\%$. For risk interval construction, the quantile level is fixed to $\alpha = 1\%$. In the score calibration module, the calibration coefficient is set to $\beta = 0.08$,

Table 1: Performance comparison on the four benchmarks. All values are percentages. Lower FPR95 and higher AUROC are better. Best results are in **bold** and the second-best results are underlined.

ID	ISIC-4		Skin-40		NCT-CRC		BreakHis		**Average**	
Method	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP [8]	82.21	73.38	86.64	68.96	69.59	81.14	88.80	67.76	81.81	72.81
Energy [14]	73.96	74.76	84.15	70.70	62.32	81.98	78.77	68.18	74.79	73.90
ReAct [22]	69.82	78.70	85.85	71.74	62.27	85.91	80.97	66.31	74.72	75.66
LogitGap [12]	82.57	73.20	83.96	71.98	72.41	80.82	90.76	67.52	75.74	74.12
CADRef [13]	77.14	73.17	85.76	58.11	52.72	84.70	88.28	60.47	75.97	69.11
Online RTL [6]	78.70	64.68	86.71	68.81	63.78	80.43	86.96	62.01	79.03	68.98
DCAC [25]	81.36	73.27	78.75	75.06	70.31	80.53	82.09	68.52	78.12	74.34
OODD [28]	85.37	71.36	75.72	74.87	51.31	83.73	78.13	72.19	72.63	75.53
OODD+RSB	83.98	72.70	75.10	75.33	44.16	86.22	80.77	73.04	71.00	76.82
OODD+RPB	81.34	72.46	74.35	74.61	45.67	85.31	80.21	72.09	70.39	76.11
OODD+Ours	**67.74**	**82.07**	**73.73**	**76.20**	**22.36**	**93.99**	**77.24**	**75.14**	**60.26**	**81.85**

and the semantic confidence threshold is set to $\eta = 0.8$. For evaluation under the streaming setting, decision calibration is first applied at each time step using feedback accumulated from previous samples, after which the system determines whether the current sample should be queried for human feedback. Any feedback obtained from the current sample is used only for calibrating subsequent samples, ensuring a strictly causal evaluation without information leakage.

Comparison methods. All methods are evaluated using a shared ResNet-50 backbone. OOD detectors are constructed with representative post-hoc scoring functions (MSP [8], Energy [14], React [22], LogitGap [12], CADRef [13]) and TTA methods (Online RTL [6], DCAC [25], OODD [28]). Under the same feedback budget, we construct two additional baselines. RSB adopts random querying with physician feedback and applies only the soft calibration term δ_t. RPB replaces physician feedback with pseudo labels [2] while retaining both soft and hard calibration. Performance is measured by FPR95 and AUROC.

3.2 Result Analysis

Efficacy Evaluation. Table 1 presents the main results on four medical benchmarks. CHILD achieves the best AUROC and FPR95 when combined with OODD. Compared with RSB, the improvement verifies that the proposed framework can effectively exploit physician feedback. Compared with RPB, the gap indicates that pseudo labels provide less reliable guidance than controlled physician supervision in this streaming setting. Results with other OOD detectors (Fig. 2) further demonstrate the general applicability of CHILD.

Ablation Study. Table 2 evaluates the impact of Soft Calibration δ_t, Adaptive Risk Sample Localization $\mathcal{I}_t$, and Hard Assignment H_t. Each component contributes consistent performance gains, and the best results are achieved when all components are enabled, demonstrating their complementary effects. On NCT-CRC, the gain from $\mathcal{I}_t$ is less pronounced, as samples selected within the

Table 2: Ablation study of CHILD. In the $\mathcal{I}_t$ column, ✓ indicates risk-aware sample selection, while ✗ denotes random selection.

δ_t	$\mathcal{I}_t$	H_t		ISIC-4		Skin-40		NCT-CRC		BreakHis		**Average**	
		h_t	y_t	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
✗	✗	✗	✗	85.37	71.36	75.72	74.87	51.31	83.73	78.13	72.19	72.63	75.53
✓	✗	✗	✗	83.98	72.70	75.10	75.33	44.16	86.22	80.77	73.04	71.00	76.82
✓	✓	✗	✗	83.26	72.87	74.28	75.36	28.23	93.34	78.38	73.54	66.03	78.77
✓	✗	✓	✓	70.02	81.32	82.03	75.30	**17.89**	**94.91**	80.14	74.02	62.52	81.39
✓	✓	✗	✓	70.56	79.47	73.02	75.98	28.75	91.54	77.72	74.23	62.51	80.30
✓	✓	✓	✗	83.43	71.75	75.45	74.49	43.43	86.25	78.91	73.05	70.30	76.38
✓	✓	✓	✓	**67.74**	**82.07**	**73.73**	**76.20**	22.36	93.99	**77.24**	**75.14**	**60.26**	**81.85**

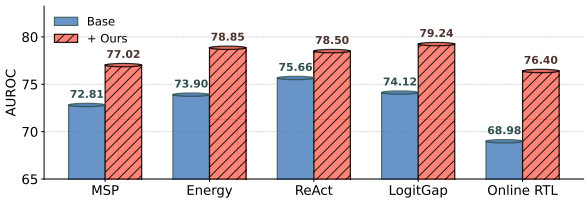

Fig. 2: Average performance with CHILD integrated into different OOD detectors.

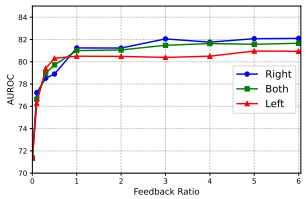

Fig. 3: Performance when varying feedback ratio.

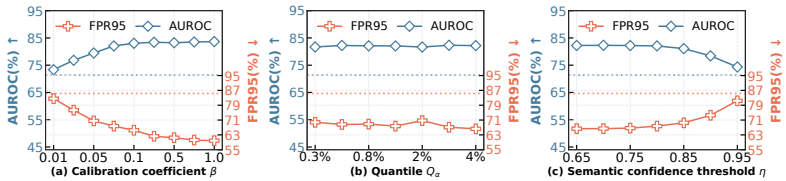

Fig. 4: Hyper-parameter sensitivity studies on ISIC-4. Dashed lines represent the performance of the best baseline (i.e., OODD).

risk interval occasionally exhibit similar characteristics, limiting the benefit of boundary-focused querying.

Budget–Performance Trade-off. Fig. 3 illustrates the effect of the feedback budget on ISIC-4 under three interval selection strategies defined relative to the reference boundary τ_t: selecting samples from the left side of τ_t, the right side of τ_t, or symmetrically from both sides. Overall, right-side selection yields stronger and more stable performance across feedback budgets, as refining the ID-side decision region more effectively corrects high-risk errors and improves ID/OOD score discrimination. Across all strategies, performance consistently improves as the feedback ratio increases, with AUROC approaching saturation at around 1% feedback. Beyond 5% feedback, further budget increases bring only marginal gains, as additional samples within the quantile-defined risk interval tend to provide increasingly redundant calibration signals. These results indicate that the proposed framework is effective and robust under limited feedback budgets.

Sensitivity Study. We analyze the robustness of the proposed framework with respect to key hyperparameters. Specifically, the calibration coefficient β is varied within $[0.01, 1.0]$, the quantile Q_α within $[0.3\%, 4\%]$, and the semantic confidence threshold η within $[0.65, 0.95]$. As shown in Fig. 4, the proposed method consistently outperforms the best baseline across broad hyperparameter ranges, indicating stable performance and low sensitivity to hyperparameter choices.

4 Conclusion

We introduced CHILD, a training-free human-in-the-loop framework for OOD detection motivated by clinical deployment. Under a strict feedback budget, CHILD selectively queries high-risk samples and incorporates sparse physician feedback through lightweight score-level calibration. Extensive experiments on multiple medical benchmarks demonstrate that CHILD consistently improves OOD detection reliability under limited feedback. These results support CHILD as a practical step toward safer medical AI systems with human oversight.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (grant No. 62571559), the Major Key Project of PCL (grant No. PCL2025AS209), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

Disclosure of Interests. Authors have no competing interests in the paper.

References

1. Bai, H., Du, X., Rainey, K., Parameswaran, S., Li, Y.: Out-of-distribution learning with human feedback. TMLR (2025)
2. Cao, C., Zhong, Z., Zhou, Z., Liu, T., Liu, Y., Zhang, K., Han, B.: Noisy test-time adaptation in vision-language models. In: ICLR (2025)
3. Chen, J., Deng, L., Gan, Z., Zheng, W., Wang, R.: Fodfom: Fake outlier data by foundation models creates stronger visual out-of-distribution detector. In: ACM MM (2024)
4. Chen, J., Zhang, T., Zheng, W., Wang, R.: Tagfog: Textual anchor guidance and fake outlier generation for visual out-of-distribution detection. In: AAAI (2024)
5. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: ICLR (2023)
6. Fan, K., Liu, T., Qiu, X., Wang, Y., Huai, L., Shangguan, Z., Gou, S., Liu, F., Fu, Y., Fu, Y., Jiang, X.: Test-time linear out-of-distribution detection. In: CVPR (2024)
7. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: ICML (2022)
8. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
9. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Medicine **16**(1), e1002730 (2019)

10. Lai, R., Lu, X., Chen, K., Chen, Q., Zheng, W.S., Wang, R.: Hierarchical vision-language learning for medical out-of-distribution detection. In: MICCAI. pp. 230–239. Springer (2025)
11. Li, Z., Zhong, C., Wang, R., Zheng, W.S.: Continual learning of new diseases with dual distillation and ensemble strategy. In: MICCAI. pp. 169–178. Springer (2020)
12. Liang, J., Hou, R., Hu, M., Chang, H., Shan, S., Chen, X.: Revisiting logit distributions for reliable out-of-distribution detection. In: NeurIPS (2025)
13. Ling, Z., Chang, Y., Zhao, H., Zhao, X., Chow, K., Deng, S.: Cadref: Robust out-of-distribution detection via class-aware decoupled relative feature leveraging. In: CVPR (2025)
14. Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS (2020)
15. Lu, X., Lai, R., Wu, Y., Chen, K., Zheng, W.S., Wang, R.: Fa: Forced prompt learning of vision-language models for out-of-distribution detection. In: ICCV (2025)
16. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. In: NeurIPS (2023)
17. Nguyen, A.M., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: CVPR (2015)
18. Oh, J.H., Falahkheirkhah, K., Bhargava, R.: Are we ready for out-of-distribution detection in digital pathology? In: MICCAI. pp. 78–89. Springer (2024)
19. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
20. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. TBME **63**(7), 1455–1462 (2015)
21. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: ECCV. pp. 206–222. Springer (2016)
22. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: NeurIPS (2021)
23. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
24. Wu, W., Ye, J., Chen, J., Chen, D., Wang, R.: Multi-os: Multimodal ood synthesis enhances out-of-distribution detection for vision-language models. In: ICASSP (2026)
25. Wu, Y., Chen, Q., Lai, R., Lu, X., Zhuang, J.X., Zhao, Z., Zheng, W.S., Wang, R.: Dcac: Dynamic class-aware cache creates stronger out-of-distribution detectors. In: AAAI (2026)
26. Xu, Z., Xiang, X., Liang, Y.: Overcoming shortcut problem in VLM for robust out-of-distribution detection. In: CVPR (2025)
27. Yan, J., Guan, X., Zheng, W., Chen, H., Wang, R.: Global and local vision-language alignment for few-shot learning and few-shot OOD detection. In: MICCAI. vol. 15964, pp. 208–218. Springer (2025)
28. Yang, Y., Zhu, L., Sun, Z., Liu, H., Gu, Q., Ye, N.: OODD: test-time out-of-distribution detection with dynamic dictionary. In: CVPR (2025)
29. Ye, J., Liao, J., Lai, R., Lu, X., Zhuang, J., Gan, Z., Wang, R.: TTL: test-time textual learning for OOD detection with pretrained vision-language models. In: CVPR (2026)
30. Zhang, J., Huang, J., Zhang, X., Shao, L., Lu, S.: Historical test-time prompt tuning for vision foundation models. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) NeurIPS (2024)

31. Zhang, Y., Wang, X., Zhou, T., Yuan, K., Zhang, Z., Wang, L., Jin, R.: Model-free test time adaptation for out-of-distribution detection. TPAMI (2025)