BenchReAD: A systematic benchmark for retinal anomaly detection

Chenyu Lian[1], Hong-Yu Zhou[2(✉)], Zhanli Hu[3], and Jing Qin[1(✉)]

[1] The Center for Smart Health, School of Nursing, the Hong Kong Polytechnic University, Hong Kong, China
chenyu.lian@connect.polyu.hk, harry.qin@polyu.edu.hk
[2] School of Biomedical Engineering, Tsinghua University, Beijing, China
whuzhouhongyu@gmail.com
[3] Research Center for Medical AI, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
zl.hu@siat.ac.cn

Abstract. Retinal anomaly detection plays a pivotal role in screening ocular and systemic diseases. Despite its significance, progress in the field has been hindered by the absence of a comprehensive and publicly available benchmark, which is essential for the fair evaluation and advancement of methodologies. Due to this limitation, previous anomaly detection work related to retinal images has been constrained by (1) a limited and overly simplistic set of anomaly types, (2) test sets that are nearly saturated, and (3) a lack of generalization evaluation, resulting in less convincing experimental setups. Furthermore, existing benchmarks in medical anomaly detection predominantly focus on one-class supervised approaches (training only with negative samples), overlooking the vast amounts of labeled abnormal data and unlabeled data that are commonly available in clinical practice. To bridge these gaps, we introduce a benchmark for retinal anomaly detection, which is comprehensive and systematic in terms of data and algorithm. Through categorizing and benchmarking previous methods, we find that a fully supervised approach leveraging disentangled representations of abnormalities (DRA) achieves the best performance but suffers from significant drops in performance when encountering certain unseen anomalies. Inspired by the memory bank mechanisms in one-class supervised learning, we propose NFM-DRA, which integrates DRA with a Normal Feature Memory to mitigate the performance degradation, resulting in a more powerful and stable approach. The benchmark is publicly available at https://github.com/DopamineLcy/BenchReAD.

Keywords: Benchmark · anomaly detection · retinal imaging.

1 Introduction

Retinal imaging is critical in screening a variety of ocular and systemic conditions, such as diabetic retinopathy, glaucoma, and myopia [12,20,23]. Despite the

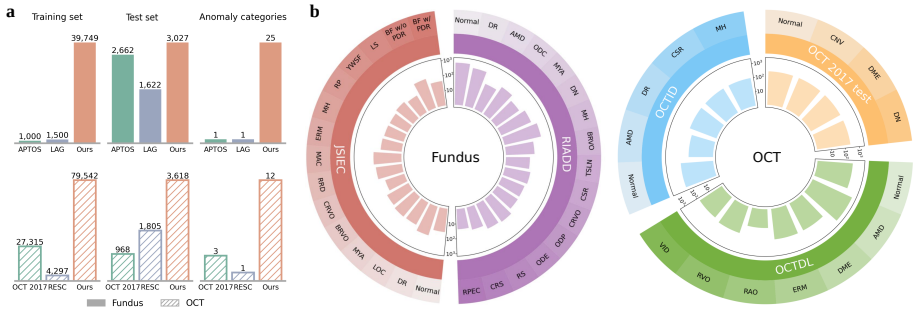

Fig. 1. (a) Comparison among widely used datasets for retinal anomaly detection (APTOS [15], LAG [19], OCT 2017 [16], and RESC [13]) and our BenchReAD. (b) Overview of the test datasets included in the proposed benchmark.

widespread application of anomaly detection in medical imaging, the field still lacks a **comprehensive and systematic benchmark** to standardize evaluation protocols and facilitate the development of novel methods. Existing benchmarks of medical anomaly detection related to retinal images often rely on **single datasets** that primarily focus on **one or a few diseases** [15,19,16,13] for both model development and evaluation, which fails to provide a robust and generalizable assessment [1,3]. We also notice that the widely used OCT 2017 dataset [16,30,11,9] has become unsuitable as a standalone benchmark because of **performance saturation**. Furthermore, most existing benchmarks focus on one-class supervised methods that only adopt negative samples for training, which originate from industrial anomaly detection where abnormal instances are rare [21,26,18,5]. In contrast, retinal imaging datasets often contain many disease samples, necessitating a more nuanced evaluation framework.

To address these limitations, we introduce BenchReAD, a systematic benchmark for retinal anomaly detection that encompasses two of the most widely used imaging modalities: fundus photography and optical coherence tomography (OCT). Our benchmark is designed to be systematic and comprehensive in two key dimensions: dataset and methodologies. From the **dataset** perspective, as illustrated in Figure 1, BenchReAD incorporates notably larger datasets with a more diverse range of anomaly categories compared to existing benchmarks, ensuring a rigorous assessment of **generalizability**. From the **methodology** perspective, we systematically categorize anomaly detection approaches into four groups based on their supervision levels: unsupervised, one-class supervised, semi-supervised, and fully supervised methods. Through extensive benchmarking, we find that a fully supervised approach leveraging disentangled representations of abnormalities (DRA) achieves the highest performance. However, it suffers from a significant drop in effectiveness when encountering certain unseen anomalies. Inspired by the memory banks in one-class supervised learning, we propose **NFM-DRA** by integrating DRA with a Normal Feature Memory to mitigate the performance drops, leading to a more powerful approach.

2 Benchmark

2.1 Datasets

The benchmark includes two of the most widely used retinal imaging modalities: fundus photography and optical coherence tomography (OCT). The test sets contain both ***seen*** (present in the training set) and ***unseen*** (absent from the training set) anomalies.

The fundus benchmark is constructed from four publicly available datasets. The training and validation sets are derived from EDDFS [27] and BRSET [22], while the test set is composed of the RIADD [24] and JSIEC [4]. We retain four common retinal anomalies of diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and myopia (MYA) in these two datasets. After preprocessing and data structuring, the training set comprises 23,361 normal images and 16,388 abnormal images, with an additional 100 normal and 100 abnormal images reserved for validation. Figure 1 (b) illustrates distribution of test sets categories, ensuring statistical robustness by excluding anomaly categories with fewer than 20 images. The RIADD dataset includes 669 normal images and 1,634 abnormal images, covering both ***seen*** anomalies (DR, AMD, optic disc cupping (ODC, highly associated with glaucoma), MYA, and drusens (DN)) and ***unseen*** anomalies (macular hole (MH), branch retinal vein occlusion (BRVO), tractional retinal detachment (TSLN), central serous retinopathy (CSR), central retinal vein occlusion (CRVO), optic disc pallor (ODP), optic disc edema (ODE), retinoschisis (RS), choroidal rupture syndrome (CRS), and retinal pigment epithelium changes (RPEC)). The JSIEC dataset comprises 38 normal images and 686 abnormal images, featuring ***seen*** categories of DR, large optic cup (LOC, highly associated with glaucoma), MYA as well as ***unseen*** categories including BRVO, CRVO, rhegmatogenous RD (RD), maculopathy (MAX), epiretinal membrane (ERM), MH, Retinitis pigmentosa (RP), yellow-white spots-flecks (YWSF), laser Spots (LS), blur fundus without PDR (BF w/o PDR), and blur fundus with suspected PDR (BF w/ PDR).

The OCT benchmark is established using three public datasets: OCT 2017 [16], OCTDL [17], and OCTID [7]. The training and validation sets are derived from OCT 2017, a large-scale classification dataset containing normal images and three abnormal classes: choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen deposits (DN). The training set includes 25,840 normal and 53,702 abnormal images, while the validation set includes 50 images per class (200 images in total). The test sets include the OCT 2017 test set, OCTDL, and OCTID, as shown in Figure 1 (b). The OCT 2017 test set includes 250 images per category (normal, CNV, DME, and DN). OCTDL includes 332 normal and 1,732 abnormal images, with two ***seen*** anomalies (AMD and DME) and four ***unseen*** anomalies: (epiretinal membrane (ERM), RAO, retinal vein occlusion (RVO), and vitreomacular interface disorders (VID)). OCTID consists of 206 normal and 336 abnormal images, including two ***seen*** anomalies (AMD and DR), and two ***unseen*** anomalies: central serous retinopathy (CSR) and MH.

2.2 Data and algorithms across different supervision levels

Supervision level settings in the training set. To systematically evaluate the performance of anomaly detection methods under different supervision levels, we partition the training set into labeled and unlabeled subsets. Specifically, the labeled subset constitutes one-third of the training data, while the unlabeled subset accounts for the remaining two-thirds. Formally, the labeled normal image subset is denoted as $\mathcal{N}_x$, with corresponding labels $\mathcal{N}_y$; the labeled abnormal image subset is denoted as $\mathcal{A}_x$, with corresponding labels $\mathcal{A}_y$; and the unlabeled image subset $\mathcal{U}_x$ contains images without any associated labels.

Algorithms across different supervision levels. Unlike previous benchmarks for medical anomaly detection that predominantly focus on one-class supervised algorithms, our benchmark systematically categorizes approaches into four groups based on supervision levels: ***Unsupervised*** methods require no annotations and make use of all available data ($\mathcal{N}_x$, $\mathcal{A}_x$, and $\mathcal{U}_x$). We choose SoftPatch [14] (NeurIPS 2023) as the representative method. ***One-class supervised*** methods rely exclusively on normal images and implicitly use their corresponding labels ($\mathcal{N}_x$, $\mathcal{N}_y$). They are trained to model the distribution of normal data and detect deviations as anomalies. We select three variants: EDC [10] (distillation-based, TMI 2023), SimpleNet [21] (self-supervised, CVPR 2023), and PatchCore [26] (memory bank-based, CVPR 2022). ***Semi-supervised*** methods represented by DDAD-ASR [2] (MedIA 2023) extend one-class supervised methods by incorporating additional unlabeled data $\mathcal{U}_x$ alongside $\mathcal{N}_x$ and $\mathcal{N}_y$ to enhance their performance. ***Fully supervised*** methods utilize both labeled normal and abnormal images along with their labels ($\mathcal{N}_x$, $\mathcal{N}_y$, $\mathcal{A}_x$, $\mathcal{A}_y$) and aims to detect both seen and unseen anomalies [6,29,31]. DRA [6] (CVPR 2022), an open-set supervised anomaly detection method, is benchmarked here.

2.3 The proposed NFM-DRA with normal feature memory

The fully supervised method DRA [6] is observed to achieve the best performance, as we will discuss in Section 3.1. However, since it is trained on both normal and abnormal samples, the predictions tend to rely on capturing previously seen abnormal features instead of similarities to normal features. This reliance on seen anomaly features reduces robustness when encountering certain unseen anomalies. To address this limitation, we draw inspiration from the memory bank mechanism used in anomaly detection [8,26] and introduce a simple yet effective improvement. Specifically, we construct a normal feature memory $\mathcal{M}$ to store features of normal samples. By comparing test image features $\mathcal{X}$ against this memory, we get representative feature $\boldsymbol{x}^*$ from $\mathcal{X}$ and its closest feature $\boldsymbol{m}^*$ from $\mathcal{M}$:

$$\boldsymbol{x}^*, \boldsymbol{m}^* = \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\max} \underset{\boldsymbol{m}\in\mathcal{M}}{\arg\min} \|\boldsymbol{x} - \boldsymbol{m}\|_2 . \tag{1}$$

Then, anomaly scores can be computed through:

$$g(\mathcal{X}; \mathcal{M}) = \left(1 - \frac{\exp\left(\|\boldsymbol{x}^* - \boldsymbol{m}^*\|_2\right)}{\sum_{\boldsymbol{m}\in\mathcal{N}_b^{\mathcal{M}}(\boldsymbol{m}^*)} \exp\|\boldsymbol{x}^* - \boldsymbol{m}\|_2}\right) \cdot \|\boldsymbol{x}^* - \boldsymbol{m}^*\|_2 , \tag{2}$$

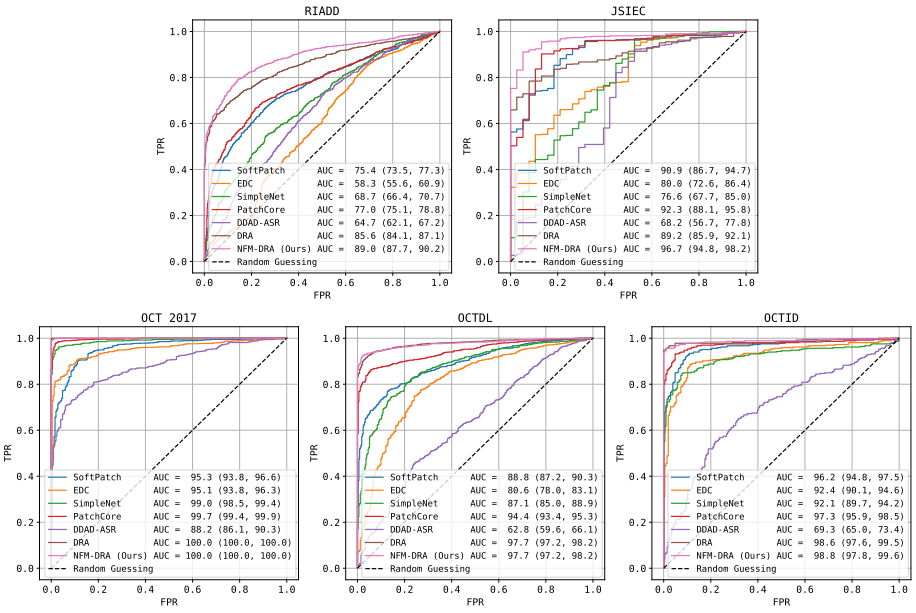

Fig. 2. ROC curves of distinguishing normal samples against all abnormal ones on test sets. Corresponding AUCs (%) are marked alongside 95% confidence intervals.

where $\mathcal{N}_b^{\mathcal{M}}(\boldsymbol{m}^*)$ denotes b features in $\mathcal{M}$ that are nearest to $\boldsymbol{m}^*$. We refine the prediction scores of DRA using anomaly values derived from Equation 2, generating the final anomaly scores:

$$f(\mathcal{X}; \mathcal{M}, \mathrm{DRA}) = \frac{1}{2}\left[g(\mathcal{X}; \mathcal{M}) + \mathrm{DRA}(\mathcal{X})\right]. \tag{3}$$

3 Experiments and analyses

3.1 Results and discussions

Overall evaluation of distinguishing normal and abnormal samples. Figure 2 presents the comparison of ROC curves and their corresponding AUCs with 95% confidence intervals (CIs). The threshold-independent metric measures the ability of the model to distinguish between normal and abnormal retinal images across all possible decision thresholds. The fully supervised method, DRA, outperforms other benchmarked approaches across four out of the five test sets, achieving notably high AUC values with narrow confidence intervals. PatchCore, a one-class supervised method, ranks second in overall performance and attains the highest AUC score on the JSIEC dataset. While other methods exhibit competitive performance on specific datasets, they generally lag behind DRA and

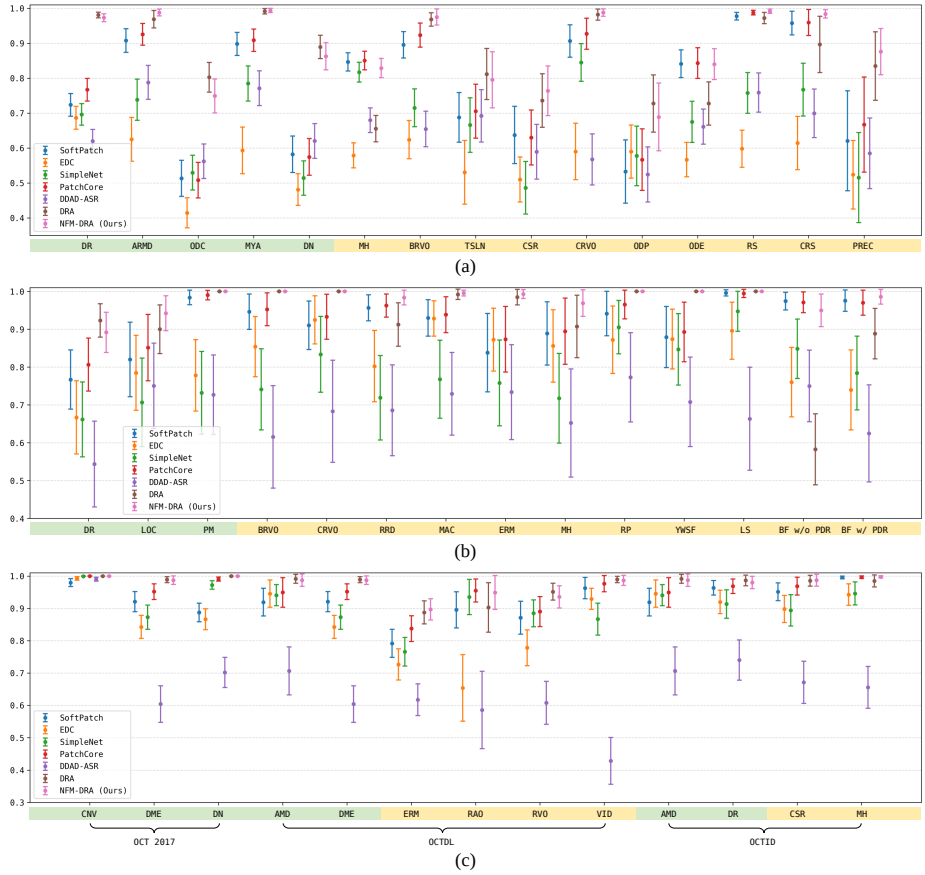

Fig. 3. AUROCs with 95% confidence intervals for normal samples compared to each abnormal category on the (a) RIADD and (b) JSIEC datasets for fundus benchmarking, as well as (c) OCT test datasets. Seen and unseen annotations highlight whether the anomaly categories appear in the training set or not, respectively.

PatchCore. The proposed NFM-DRA is the best-performing approach, showing notable improvements, particularly on the RIADD and JSIEC datasets.

Fully supervised method DRA achieves the best performance, while one-class supervised method PatchCore exhibits greater robustness. In addition to the overview comparisons (Figure 2), AUROCs with 95% CIs for normal samples compared to each abnormal category in the RIADD, JSIEC and OCT test sets are presented in Figure 3 (a), (b), and (c), respectively. DRA consistently outperforms other methods on seen anomalies and also performs well on unseen anomalies. However, for certain anomaly categories such as MH, ODE, RS, and CRS in the RIADD dataset, the performance is notably weaker, and PatchScore achieves the best performance. A similar trend is observed in

Table 1. Threshold-dependent results (%) on fundus test sets, presenting F1 scores, specificity and sensitivity. Categories of seen and unseen anomalies are highlighted to indicate whether they are included in the training set. The **best** results are shown in bold, and the second-best ones are underlined.

Dataset	Normal vs	SoftPatch			EDC			SimpleNet			PatchCore			DDAD-ASR			DRA			NFM-DRA (Ours)		
		F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN
RIADD	All abnormal	82.7	5.8	97.7	**83.0**	0.0	**100.0**	**83.0**	0.9	99.8	82.4	13.6	94.9	37.4	88.6	24.1	79.3	90.3	68.3	70.0	**99.3**	54.0
	DR	54.8	5.8	97.0	54.5	0.0	**100.0**	54.6	0.9	99.8	56.1	13.6	95.3	26.5	88.6	18.2	90.0	90.3	95.0	**93.2**	**99.3**	88.3
	AMD	18.0	5.8	**100.0**	17.1	0.0	**100.0**	17.2	0.9	**100.0**	19.3	13.6	**100.0**	33.3	88.6	42.0	64.6	90.3	92.8	**87.9**	**99.3**	84.1
	ODC	30.8	5.8	92.3	31.7	0.0	**100.0**	31.7	0.9	99.4	29.7	13.6	82.6	22.3	88.6	18.7	**50.8**	90.3	48.4	28.9	**99.3**	17.4
	MYA	19.2	5.8	**100.0**	18.3	0.0	**100.0**	18.5	0.9	**100.0**	20.6	13.6	**100.0**	32.2	88.6	38.7	68.5	90.3	97.3	**90.4**	**99.3**	88.0
	DN	31.5	5.8	96.1	31.2	0.0	**100.0**	31.3	0.9	99.3	30.6	13.6	86.8	27.9	88.6	24.3	**67.7**	90.3	73.0	60.4	**99.3**	44.7
	MH	50.0	5.8	**100.0**	48.5	0.0	**100.0**	48.7	0.9	**100.0**	**51.9**	13.6	99.4	38.8	88.6	29.8	42.7	90.3	32.7	35.9	**99.3**	22.2
	BRVO	20.7	5.8	**100.0**	19.7	0.0	**100.0**	19.8	0.9	**100.0**	21.9	13.6	98.8	14.1	88.6	14.6	66.4	90.3	89.0	**85.5**	**99.3**	79.3
	TSLN	11.3	5.8	**100.0**	10.7	0.0	**100.0**	10.5	0.9	97.5	12.2	13.6	**100.0**	18.8	88.6	30.0	**33.3**	90.3	52.5	30.2	**99.3**	20.0
	CSR	13.9	5.8	94.4	13.9	0.0	**100.0**	14.0	0.9	**100.0**	13.8	13.6	87.0	12.9	88.6	16.7	**32.4**	90.3	42.6	31.4	**99.3**	20.4
	CRVO	11.8	5.8	**100.0**	11.2	0.0	**100.0**	11.2	0.9	**100.0**	12.7	13.6	**100.0**	5.0	88.6	7.1	54.4	90.3	95.2	**85.4**	**99.3**	83.3
	ODP	12.0	5.8	97.7	11.6	0.0	**100.0**	11.7	0.9	**100.0**	12.1	13.6	90.9	6.5	88.6	9.1	**31.0**	90.3	45.5	25.0	**99.3**	15.9
	ODE	20.5	5.8	**100.0**	19.5	0.0	**100.0**	19.6	0.9	**100.0**	21.9	13.6	**100.0**	24.6	88.6	27.2	36.9	90.3	40.7	**45.0**	**99.3**	30.9
	RS	17.5	5.8	**100.0**	16.7	0.0	**100.0**	16.8	0.9	**100.0**	18.8	13.6	**100.0**	32.7	88.6	41.8	60.3	90.3	85.1	**86.6**	**99.3**	82.1
	CRS	9.7	5.8	**100.0**	9.2	0.0	**100.0**	9.3	0.9	**100.0**	10.5	13.6	**100.0**	13.6	88.6	23.5	44.1	90.3	82.4	**78.1**	**99.3**	73.5
	RPEC	5.9	5.8	87.0	6.4	0.0	**100.0**	6.5	0.9	**100.0**	6.4	13.6	87.0	7.8	88.6	17.4	27.5	90.3	60.9	**44.4**	**99.3**	34.8
	Average	25.6	5.8	97.6	25.2	0.0	**100.0**	25.3	0.9	99.7	26.3	13.6	95.2	22.1	88.6	24.0	53.1	90.3	68.8	**61.1**	**99.3**	52.4
JSIEC	DR	81.9	10.5	91.5	84.8	0.0	**100.0**	86.1	13.2	99.1	84.3	21.1	93.4	70.8	47.4	65.1	**90.9**	94.7	84.9	90.7	**100.0**	83.0
	LOC	74.6	10.5	**100.0**	72.5	0.0	**100.0**	75.2	13.2	**100.0**	76.9	21.1	**100.0**	**83.3**	47.4	**100.0**	71.6	94.7	58.0	46.2	**100.0**	30.0
	PM	76.1	10.5	**100.0**	74.0	0.0	**100.0**	76.6	13.2	**100.0**	78.3	21.1	**100.0**	79.7	47.4	90.7	98.2	94.7	**100.0**	**100.0**	**100.0**	**100.0**
	BRVO	72.1	10.5	**100.0**	69.8	0.0	**100.0**	72.7	13.2	**100.0**	74.6	21.1	**100.0**	76.9	47.4	90.9	97.8	94.7	**100.0**	**100.0**	**100.0**	**100.0**
	CRVO	56.4	10.5	**100.0**	53.7	0.0	**100.0**	57.1	13.2	**100.0**	59.5	21.1	**100.0**	66.7	47.4	95.5	95.7	94.7	**100.0**	**100.0**	**100.0**	**100.0**
	RRD	77.0	10.5	**100.0**	75.0	0.0	**100.0**	77.6	13.2	**100.0**	79.2	21.1	**100.0**	**85.1**	47.4	**100.0**	68.9	94.7	54.4	77.4	**100.0**	63.2
	MAC	81.3	10.5	**100.0**	79.6	0.0	**100.0**	81.8	13.2	**100.0**	83.1	21.1	**100.0**	88.1	47.4	**100.0**	98.0	94.7	98.6	**99.3**	**100.0**	98.6
	ERM	60.5	10.5	**100.0**	57.8	0.0	**100.0**	61.2	13.2	**100.0**	63.4	21.1	**100.0**	70.4	47.4	96.2	**90.2**	94.7	88.5	87.0	**100.0**	76.9
	MH	57.5	10.5	**100.0**	54.8	0.0	**100.0**	58.2	13.2	**100.0**	60.5	21.1	**100.0**	65.6	47.4	91.3	**81.0**	94.7	73.9	75.7	**100.0**	60.9
	RP	56.4	10.5	**100.0**	53.7	0.0	**100.0**	57.1	13.2	**100.0**	59.5	21.1	**100.0**	68.8	47.4	**100.0**	95.7	94.7	**100.0**	**100.0**	**100.0**	**100.0**
	YWSF	63.0	10.5	**100.0**	60.4	0.0	**100.0**	63.7	13.2	**100.0**	65.9	21.1	**100.0**	71.1	47.4	93.1	96.7	94.7	**100.0**	**100.0**	**100.0**	**100.0**
	LS	54.1	10.5	**100.0**	51.3	0.0	**100.0**	54.8	13.2	**100.0**	57.1	21.1	**100.0**	66.7	47.4	**100.0**	95.2	94.7	**100.0**	**100.0**	**100.0**	**100.0**
	BF w/o PDR	87.0	10.5	**100.0**	85.3	0.0	99.1	87.4	13.2	**100.0**	**88.4**	21.1	**100.0**	86.0	47.4	88.6	29.4	94.7	17.5	44.9	**100.0**	28.9
	BF w/ PDR	72.6	10.5	**100.0**	70.3	0.0	**100.0**	73.2	13.2	**100.0**	75.0	21.1	**100.0**	70.0	47.4	77.8	65.7	94.7	51.1	**78.4**	**100.0**	64.4
	Average	71.2	10.5	99.3	69.3	0.0	**99.9**	72.0	13.2	**99.9**	73.5	21.1	99.5	76.1	47.4	91.9	83.9	94.7	80.0	**85.6**	**100.0**	78.6

the JSIEC dataset, where DRA consistently surpasses counterparts on seen categories but underperforms on specific unseen anomalies, including CRVO, BF w/o PDR, and BF w/ PDR. Notably, DRA performs the worst among all comparative methods on the BF w/o PDR category. The observed contradiction between the strong overall performance of supervised methods and their unstable performance on unseen anomalies underscores the need to improve robustness.

The proposed NFM-DRA outperforms all benchmarked methods and shows robustness for unseen anomalies. Motivated by the findings above, we propose NFM-DRA to boost the robustness of DRA across diverse anomaly types by leveraging a Normal Feature Memory. Specifically, we augment prediction scores of DRA with anomaly scores derived from comparing the testing images with the proposed normal feature memory, thereby obtaining the final anomaly scores (refer to Section 2.3 for more details). As illustrated in Figure 2, the proposed NFM-DRA is the best method on fundus datasets and delivers performance comparable to DRA on OCT datasets, further enhancing robustness. NFM-DRA mitigates the significant performance drops in detecting unseen anomalies such as MH and ODE in the RIADD dataset, as well as BF w/o PDR in the JSIEC dataset, as shown in Figure 3 (a) and (b). Furthermore, the threshold-dependent performance is improved, as evidenced by the average F1 scores in Table 1 and 2.

Table 2. Threshold-dependent results (%) on OCT test sets, presenting F1 scores, specificity and sensitivity. Seen and unseen anomaly categories are highlighted to show whether they are included in the training set. The **best** results are in bold and the second-best ones are underlined.

Dataset	Normal vs	SoftPatch			EDC			SimpleNet			PatchCore			DDAD-ASR			DRA			NFM-DRA (Ours)		
		F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN	F1	SPC	SEN
OCT 2017	CNV	75.0	34.0	99.6	71.9	22.0	**100.0**	89.0	75.2	**100.0**	91.6	81.6	**100.0**	68.1	6.4	**100.0**	**99.2**	**98.4**	**100.0**	98.8	97.6	**100.0**
	DME	75.2	34.0	**100.0**	71.9	22.0	**100.0**	88.8	75.2	99.6	91.6	81.6	**100.0**	67.9	6.4	99.6	**99.2**	**98.4**	**100.0**	98.8	97.6	**100.0**
	DN	74.4	34.0	98.4	70.3	22.0	96.4	87.4	75.2	96.8	91.0	81.6	98.8	67.6	6.4	98.8	**99.2**	**98.4**	**100.0**	98.6	97.6	99.6
	Average	78.6	34.0	99.3	75.5	22.0	98.8	90.1	75.2	98.8	92.7	81.6	99.6	72.5	6.4	99.5	**99.3**	**98.4**	**100.0**	98.9	97.6	99.9
OCTDL	AMD	90.4	33.7	97.2	88.1	0.0	**100.0**	84.8	**83.4**	76.8	92.2	45.2	98.1	88.1	0.0	**100.0**	89.6	15.1	99.8	**94.7**	61.4	99.2
	DME	55.8	33.7	96.6	47.0	0.0	**100.0**	**72.2**	**83.4**	77.6	60.3	45.2	96.6	47.0	0.0	**100.0**	51.0	15.1	**100.0**	69.0	61.4	98.6
	ERM	54.9	33.7	91.6	48.3	0.0	**100.0**	58.6	**83.4**	56.1	59.0	45.2	91.0	48.3	0.0	**100.0**	51.4	15.1	97.4	**65.9**	61.4	89.7
	RAO	16.7	33.7	**100.0**	11.7	0.0	**100.0**	**41.2**	**83.4**	90.9	19.5	45.2	**100.0**	11.7	0.0	**100.0**	12.9	15.1	95.5	24.6	61.4	95.5
	RVO	46.4	33.7	96.0	37.8	0.0	**100.0**	**66.7**	**83.4**	77.2	49.9	45.2	93.1	37.8	0.0	**100.0**	41.7	15.1	**100.0**	58.2	61.4	93.1
	VID	40.0	33.7	97.4	31.4	0.0	**100.0**	**60.6**	**83.4**	75.0	44.6	45.2	97.4	31.4	0.0	**100.0**	35.0	15.1	**100.0**	53.2	61.4	97.4
	Average	56.6	33.7	96.5	50.8	0.0	**100.0**	**66.9**	**83.4**	75.5	59.8	45.2	96.2	50.8	0.0	**100.0**	53.4	15.1	98.9	65.8	61.4	95.9
OCTID	AMD	63.0	72.8	92.7	34.8	0.0	**100.0**	42.6	28.2	**100.0**	82.9	95.1	83.6	34.8	0.0	**100.0**	90.8	95.1	98.2	**96.2**	**100.0**	92.7
	DR	77.4	72.8	96.3	51.0	0.0	**100.0**	56.7	28.2	94.4	89.6	95.1	88.8	51.0	0.0	**100.0**	93.6	95.1	96.3	**96.1**	**100.0**	92.5
	CSR	76.6	72.8	96.1	49.8	0.0	**100.0**	55.5	28.2	94.1	89.7	95.1	89.2	49.8	0.0	**100.0**	93.3	95.1	96.1	**96.4**	**100.0**	93.1
	MH	78.5	72.8	**100.0**	49.8	0.0	**100.0**	56.7	28.2	97.1	93.8	95.1	97.1	49.8	0.0	**100.0**	94.3	95.1	98.0	**98.0**	**100.0**	96.1
	Average	77.3	72.8	96.4	52.7	0.0	**100.0**	58.6	28.2	96.3	89.9	95.1	89.8	52.7	0.0	**100.0**	93.8	95.1	97.1	**96.7**	**100.0**	93.6

OCT 2017 is no longer sufficient as a standalone benchmark. Figure 3 (c) presents the results for each abnormal category on OCT test sets. As shown in the figure, the AUROCs on OCT 2017 are nearly saturated for DRA, particularly for the CNV category, where DRA, PatchCore, and SimpleNet all achieve near-perfect scores. This saturation indicates that the OCT 2017 dataset is no longer adequate as a standalone benchmark for evaluating retinal anomaly detection methods [1,3,10], highlighting the importance of our work.

More attention should be paid to threshold-dependent metrics. Most anomaly detection methods and benchmarks only take threshold-independent metrics for evaluation [1,3,28], which limits the evaluation of their readiness for real-world deployment. Benchmarks and analyses for threshold-dependent metrics are critical to draw more attention to the evaluation tailored for clinical practice. Thresholds were selected based on the best F1 scores on the validation sets. As shown in Table 1 and 2, the F1 scores, specificities, and sensitivities reveal that while the proposed NFM-DRA achieves the highest average F1 scores, no single method consistently outperforms others across different datasets and anomaly categories. Moreover, the performance of threshold-dependent metrics exhibits more significant fluctuations compared to threshold-agnostic metrics. We hope this benchmark will inspire a rethinking of evaluation protocols for practical anomaly detection.

3.2 Implementation

We conduct benchmarking following the original papers and official codebases based on PyTorch [25]. Grid search is employed to identify the optimal learning rate and number of epochs around the default settings. The best-performing checkpoints on the validation set are evaluated on the test set. All experiments are performed on a single NVIDIA RTX 3090 GPU.

4 Conclusion

In this paper, we establish a systematic evaluation framework based on seven public datasets. We evaluate anomaly detection methods under different levels of supervision. DRA, a fully supervised approach, achieves the best overall performance, while PatchCore, a memory bank-based one-class supervised method, demonstrates greater robustness for unseen anomalies. Based on these findings, we propose NFM-DRA to enhance DRA with a novel feature memory, which outperforms all benchmarked methods and exhibits improved robustness. We hope the proposed benchmark will inspire rigorous evaluation of retinal anomaly detection and foster impactful research in this field.

Acknowledgments. This work was supported in part by a grant of Collaborative Research with World-leading Research Groups in The Hong Kong Polytechnic University (project no. G-SACF), and a Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project) under Shenzhen Municipal Science and Technology Innovation Commission (project no. SGDX20230821092359002).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: Bmad: Benchmarks for medical anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4042–4053 (2024)
2. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Medical Image Analysis **86**, 102794 (2023)
3. Cai, Y., Zhang, W., Chen, H., Cheng, K.T.: Medianomaly: A comparative study of anomaly detection in medical images. arXiv preprint arXiv:2404.04518 (2024)
4. Cen, L.P., Ji, J., Lin, J.W., Ju, S.T., Lin, H.J., Li, T.P., Wang, Y., Yang, J.F., Liu, Y.F., Tan, S., et al.: Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature communications **12**(1), 4828 (2021)
5. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9737–9746 (2022)
6. Ding, C., Pang, G., Shen, C.: Catching both gray and black swans: Open-set supervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7388–7398 (2022)
7. Gholami, P., Roy, P., Parthasarathy, M.K., Lakshminarayanan, V.: Octid: Optical coherence tomography image database. Computers & Electrical Engineering **81**, 106532 (2020)
8. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1705–1714 (2019)

9. Guo, J., Jia, L., Zhang, W., Li, H., et al.: Recontrast: Domain-specific anomaly detection via contrastive reconstruction. Advances in Neural Information Processing Systems **36** (2024)
10. Guo, J., Lu, S., Jia, L., Zhang, W., Li, H.: Encoder-decoder contrast for unsupervised anomaly detection in medical images. IEEE Transactions on Medical Imaging (2023)
11. Han, X., Chen, X., Liu, L.P.: Gan ensemble for anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 4090–4097 (2021)
12. Han, Y., Li, W., Liu, M., Wu, Z., Zhang, F., Liu, X., Tao, L., Li, X., Guo, X.: Application of an anomaly detection model to screen for ocular diseases using color retinal fundus images: design and evaluation study. Journal of medical Internet research **23**(7), e27822 (2021)
13. Hu, J., Chen, Y., Yi, Z.: Automated segmentation of macular edema in oct using deep neural networks. Medical image analysis **55**, 216–227 (2019)
14. Jiang, X., Liu, J., Wang, J., Nie, Q., Wu, K., Liu, Y., Wang, C., Zheng, F.: Softpatch: Unsupervised anomaly detection with noisy data. Advances in Neural Information Processing Systems **35**, 15433–15445 (2022)
15. Karthik, Maggie, Dane, S.: Aptos 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection (2019), kaggle
16. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
17. Kulyabin, M., Zhdanov, A., Nikiforova, A., Stepichev, A., Kuznetsova, A., Ronkin, M., Borisov, V., Bogachev, A., Korotkich, S., Constable, P.A., et al.: Octdl: Optical coherence tomography dataset for image-based deep learning methods. Scientific Data **11**(1), 365 (2024)
18. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9664–9674 (2021)
19. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10571–10580 (2019)
20. Li, Y., Mitchell, W., Elze, T., Zebardast, N.: Association between diabetes, diabetic retinopathy, and glaucoma. Current Diabetes Reports **21**, 1–16 (2021)
21. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20402–20411 (2023)
22. Nakayama, L.F., Restrepo, D., Matos, J., Ribeiro, L.Z., Malerbi, F.K., Celi, L.A., et al.: Brset: A brazilian multilabel ophthalmological dataset of retina fundus photos. PLOS Digital Health **3**(7), e0000454 (2024). https://doi.org/10.1371/journal.pdig.0000454, https://doi.org/10.1371/journal.pdig.0000454
23. Narasimha-Iyer, H., Can, A., Roysam, B., Stewart, V., Tanenbaum, H.L., Majerovics, A., Singh, H.: Robust detection and classification of longitudinal changes in color retinal fundus images for monitoring diabetic retinopathy. IEEE transactions on biomedical engineering **53**(6), 1084–1098 (2006)
24. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research. Data **6**(2), 14 (2021)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-

performance deep learning library. Advances in neural information processing systems **32** (2019)
26. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
27. Xia, X., Li, Y., Xiao, G., Zhan, K., Yan, J., Cai, C., Fang, Y., Huang, G.: Benchmarking deep models on retinal fundus disease diagnosis and a large-scale dataset. Signal Processing: Image Communication **127**, 117151 (2024). https://doi.org/https://doi.org/10.1016/j.image.2024.117151, https://www.sciencedirect.com/science/article/pii/S0923596524000523
28. Xie, G., Wang, J., Liu, J., Lyu, J., Liu, Y., Wang, C., Zheng, F., Jin, Y.: Im-iad: Industrial image anomaly detection benchmark in manufacturing. IEEE Transactions on Cybernetics (2024)
29. Yao, X., Li, R., Zhang, J., Sun, J., Zhang, C.: Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24490–24499 (2023)
30. Zhai, M., Wu, X., He, Z., Wang, C., Wang, H., Wang, P.: Dual-branch retinal oct anomaly detection based on knowledge distillation and reconstruction. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
31. Zhu, J., Ding, C., Tian, Y., Pang, G.: Anomaly heterogeneity learning for open-set supervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17616–17626 (2024)