Research Square

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

# Activation Fluctuation-Driven Scaling for Reliable Medical Out-of-Distribution Detection

**Feichao Wang**
School of Software, Shandong University
**Rundong He**
School of Software, Shandong University
**Yilong Yin**
ylyin@sdu.edu.cn

School of Software, Shandong University

**Research Article**

**Keywords:** Out-of-distribution detection, Activation fluctuation, Scoring funciton, Medical image diagnosis

**Posted Date:** May 31st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9608616/v1

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

**Additional Declarations:** No competing interests reported.

# Activation Fluctuation-Driven Scaling for Reliable Medical Out-of-Distribution Detection

Feichao Wang[a,b], Rundong He[a], Yilong Yin[a,*]

[a] *School of Software, Shandong University*
[b] *Network Information Center, Qilu Normal University*

**Abstract**

Reliable out-of-distribution (OOD) detection is essential for safely deploying deep learning models in open-world medical imaging scenarios, where unexpected inputs may lead to high-risk diagnostic errors. Scale, a representative post-hoc method, enlarges the distributional discrepancy between in-distribution (ID) and OOD samples by globally scaling neural activations. However, it does not explicitly account for activation fluctuations, which characterize the overall variance of the activation distribution and exhibit clear differences between ID and OOD samples in our empirical observations. Motivated by this finding, we propose Activation Fluctuation-Driven Scaling (AFDS), a sample-level scaling strategy that dynamically estimates each sample's activation fluctuation and adaptively adjusts the scaling intensity to better amplify ID–OOD separability. In addition, medical images often involve subtle inter-class differences and substantial covariate shifts, making maximum-probability-based confidence insufficient for capturing sample uncertainty. To address this issue, we design an uncertainty scoring function that combines the maximum prediction with the relative differences between it and the remaining class predictions, thereby exploiting richer inter-class information for OOD identification. Extensive experiments on the MEDMNIST and MIDOG datasets show that AFDS achieves state-of-the-art performance on Near-OOD, Far-OOD, and covariate-shift ID (CS-ID) detection tasks, demonstrating its effectiveness and generalizability in complex medical imaging scenarios.

*Keywords:* Out-of-distribution detection; Activation fluctuation; Scoring funciton;

*Corresponding author

## 1. Introduction

In real-world open-world applications, the input data processed by deep learning models often deviate from the independent and identically distributed (i.i.d.) assumption relied upon during training [1, 2, 3]. Out-of-Distribution (OOD) samples pose a significant reliability threat to models during deployment [4]. If a model fails to effectively identify OOD samples, it may lead to severe consequences in real-world dynamic environments. OOD detection methods aim to enable models to autonomously identify and reject predictions for unknown or anomalous samples, thereby preventing high-confidence erroneous decisions. This is crucial for enhancing the reliability and safety of AI systems in safety-critical domains such as autonomous driving [5], industrial quality inspection [6], and medical image diagnosis [7, 8]. In medical image diagnosis, the importance of OOD detection is particularly pronounced [9, 10, 11]. AI-driven medical image diagnostic systems are typically trained on data specific to particular diseases. But in clinical practice, models are likely to encounter rare abnormalities not seen in the training set, which can be considered OOD samples. If the model misclassifies an OOD sample as a known category with high confidence, it could lead to a misdiagnosis, posing a threat to patient health.

Among various OOD detection methods, post-hoc approaches [12, 13] have gained significant attention due to their high computational efficiency, ease of implementation, and ability to operate without retraining. These methods directly use features or outputs generated by pre-trained models for discrimination. A representative approach is the scaling mechanism based on intermediate layer activations. For example, the Scale method [14] enhances the distributional difference between In-Distribution (ID) and OOD samples by globally scaling neural activations, enabling effective separation. However, Scale overlooks the influence of activation fluctuations during the scaling process, which reflect the overall variance of the activation distribution. Additionally, medical image data presents unique challenges, such as subtle inter-class differences and significant covariate shifts, resulting in a failure to fully capture in terms of sample

uncertainty.

Our experiments reveal a significant difference in activation fluctuations between ID and OOD samples. This observation motivates the design of a novel scaling mechanism that more effectively responds to the activation feature of different sample types. Based on this, we propose an innovative activation fluctuation-driven scaling method, named Activation Fluctuation-Driven Scaling (AFDS) for medical OOD detection. This method introduces a novel sample-level fluctuation-aware mechanism, dynamically assessing each sample's activation fluctuations and adaptively adjusting the scaling intensity to effectively amplify the difference between ID and OOD samples. Additionally, to address the subtle inter-class differences and significant covariate shifts inherent in medical imaging data, we further introduce a new uncertainty scoring function. This function not only considers the maximum response in the predictions but also accounts for its differences from the remaining class responses, enabling a more comprehensive extraction of structural information from the output distribution and achieving effective separation between ID and OOD samples. Our experiments on public medical imaging datasets such as MEDMNIST and MIDOG show that the proposed approach achieves leading performance on Near-OOD, Far-OOD, and covariate shift (CS-ID) detection tasks, demonstrating its effectiveness and generalization in complex medical imaging scenarios. Our main contributions are summarized as follows:

- We propose an Activation Fluctuation-driven Scaling (AFDS) method, which adapts the scaling intensity based on sample-specific activation fluctuations, effectively amplifying difference between ID and OOD samples.
- We propose a new uncertainty scoring function that addresses the specific challenges of medical OOD detection by considering both the maximum response and its differences from the remaining class responses.
- We conduct extensive experiments on the MEDMNIST and MIDOG benchmark datasets and achieve superior performance on Near-OOD, Far-OOD, and covariate shift (CS-ID) detection tasks, demonstrating the effectiveness and generalization of the proposed method.

## 2. Related Work

In open-world applications, deep learning models often encounter out-of-distribution (OOD) samples that differ significantly from the training distribution, leading to high-confidence erroneous predictions and posing a threat to system reliability. OOD detection enables models to automatically identify and reject predictions on unknown or anomalous samples, which is critical for safety-critical domains such as autonomous driving [5], industrial inspection [6], and medical image diagnosis [7, 15]. In medical image diagnosis OOD detection is particularly critical, as clinical environments may contain rare abnormalities, heterogeneous imaging modalities, device artifacts, or cross-institution data. Misclassifying such OOD samples can lead directly to diagnostic errors and threaten patient safety. Moreover, medical imaging data presents unique challenges such as subtle inter-class differences and significant covariate shifts that make OOD detection both necessary and difficult in this domain.

OOD detection methods can be broadly categorized into training-time [16, 17, 18] and test-time approaches, depending on whether they require intervention during training. Training-time methods enhance a model's OOD awareness by retraining or fine-tuning the network. Representative approaches include Virtual Outlier Synthesis (VOS) [19], LogitNorm [20], Outlier Exposure (OE) [21] and MixOE [22] , and hybrid training paradigms [23]. Existing training-time methods often fall short for clinical deployment due to high computational cost, strong dependence on auxiliary data, or limited adaptability to the characteristics of medical imaging. In medical image diagnosis, OOD detection is particularly crucial due to the potential presence of rare lesions, diverse imaging modalities, device-induced artifacts, and cross-institutional data variations in clinical settings. Misclassifying OOD samples can directly result in misdiagnosis, jeopardizing patient health. Medical imaging presents unique challenges for OOD detection: heterogeneous modalities with distinct statistical properties, subtle inter-class variations requiring fine-grained discrimination and significant covariate shifts across datasets. These factors make OOD detection both an urgent necessity and a technically demanding task in this field. In contrast, test-time methods (also known as post-hoc methods) require no modification to model parameters and operate directly on

pre-trained models, offering advantages in deployment flexibility, computational efficiency, and compatibility. As a post-hoc strategy, test-time methods have become a focal point in OOD detection research, as they improve model robustness without retraining. These methods typically operate on the outputs or activations of pre-trained models and can be grouped into four categories: distance-based [24], output-based [25], gradient-based [26], and activation-based methods [27, 28]. Distance-based methods estimate a sample's deviation from the training distribution by measuring its position in the feature space. These methods rely on the assumption that in-distribution (ID) samples form compact clusters, whereas OOD samples lie near the cluster boundaries or outside them. The Mahalanobis distance [24] is a classical example. It accounts for correlations among features and is insensitive to scale variations, but it requires estimating a covariance matrix, which can lead to numerical instability in high-dimensional feature spaces. However, its performance is highly dependent on feature quality and becomes limited when the model's representational capacity is insufficient. This issue is amplified in medical imaging, where high-dimensional data and subtle inter-class similarities can prevent distance metrics from capturing fine-grained anomalies effectively. Output-based methods derive OOD scores directly from model outputs, such as logits or softmax probabilities, offering the advantage of simple computation. The maximum softmax probability (MSP) [25] uses the largest softmax output as a confidence score, but it suffers from information loss due to the saturation of the softmax function. This limitation is pronounced in medical imaging, where classifiers often exhibit overconfidence on OOD samples. The MaxLogit method [29] uses the maximum logit value directly, avoiding the compression effect of softmax. However, it ignores the inter-class information contained in the non-maximum logits and is limited in medical settings, where logit distributions often exhibit low separability. Existing output-based methods introduce energy functions to improve the separation between ID and OOD samples, but they require hyperparameter tuning and often exhibit limited generalization. Maximum Concept Matching (MCM) [30] leverages all logits through softmax normalization, alleviating some of the limitations of earlier methods, but may lose the magnitude information contained in the raw logits. Overall, these methods fail to fully exploit the structural differences in the logit distribution. most approaches focus

only on the dominant class. They overlook the global shape of the output distribution. This is a critical limitation in medical imaging, where subtle lesions or device-induced shifts can introduce slight but meaningful changes in the logit patterns. Gradient-based methods capture differences in optimization behavior between ID and OOD samples by analyzing the gradients of the loss with respect to model parameters. For ID samples, the model has already learned the relevant features, resulting in small and consistent gradient magnitudes. In contrast, OOD samples require larger parameter adjustments, leading to gradients that are both larger in magnitude and more dispersed in direction. GradNorm [26] is a representative approach that uses the gradient norm as an OOD indicator. However, it incurs high computational cost and is sensitive to model architecture, which can lead to instability in medical imaging due to the complexity of the data. Activation-based methods enhance OOD detection by analyzing or modifying intermediate-layer activations, giving rise to a variety of technical approaches. In activation clipping and pruning strategies, ReAct [27] suppresses abnormal OOD responses by clipping excessively high activations, while ASH [31] amplifies abnormal patterns by pruning low activations and scaling high ones. However, these approaches may introduce distortion or compromise ID accuracy. Recent studies show that pruning operations can impair OOD detection, while scaling plays the central role. Representative typical-set methods such as BATS [32] construct high-probability feature regions using channel-level statistics, but they rely on fixed thresholds and cannot adapt to varying degrees of abnormality. Recent work has shifted toward activation scaling and normalization techniques. For example, the SCALE method [14] enhances robustness by relying solely on scaling operations, thereby avoiding the negative effects of pruning. However, existing methods still face challenges in medical imaging. In particular, ignoring activation fluctuations can lead to uneven feature scaling and limited ability to handle covariate shifts.

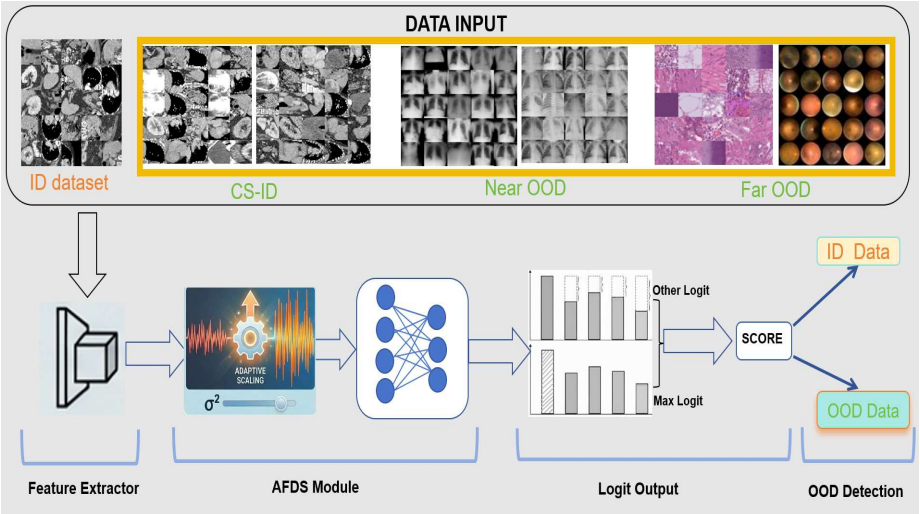



Figure 1: Activation Fluctuation-Driven Scaling (AFDS) General Framework

## 3. AFDS: Activation Flustuation-Driven Scaling for Medical OOD Detection

### *3.1. Statistical Foundations of OOD Detection*

The objective of Out-of-Distribution (OOD) detection is to construct a decision function $D(\mathbf{x})$ that determines whether a test sample $\mathbf{x}$ originates from the same distribution as the training data (ID) or from an unknown distribution (OOD). This problem can be formulated as a statistical hypothesis test, as shown in Eq.(1):

$$\mathcal{H}_0 : \hat{\mathbf{x}} \in P_0 \quad \text{vs.} \quad \mathcal{H}_1 : \hat{\mathbf{x}} \notin P_0, \tag{1}$$

where $P_0$ denotes the marginal input distribution induced by the training set $\mathcal{D}_{\text{in}}^{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, $\hat{\mathbf{x}}$ denotes a test input sample.

A scalar scoring function $S(\hat{\mathbf{x}}; f)$ together with a threshold $\gamma$ produces the decision rule defined in Eq.(2):

$$D(\hat{\mathbf{x}}) = \begin{cases} \text{ID}, & \text{if } S(\hat{\mathbf{x}}) \geq \gamma, \\ \text{OOD}, & \text{otherwise.} \end{cases} \tag{2}$$

The rejection region is correspondingly defined as $\mathcal{R} = \{\hat{\mathbf{x}} : S(\hat{\mathbf{x}}) < \gamma\}$.

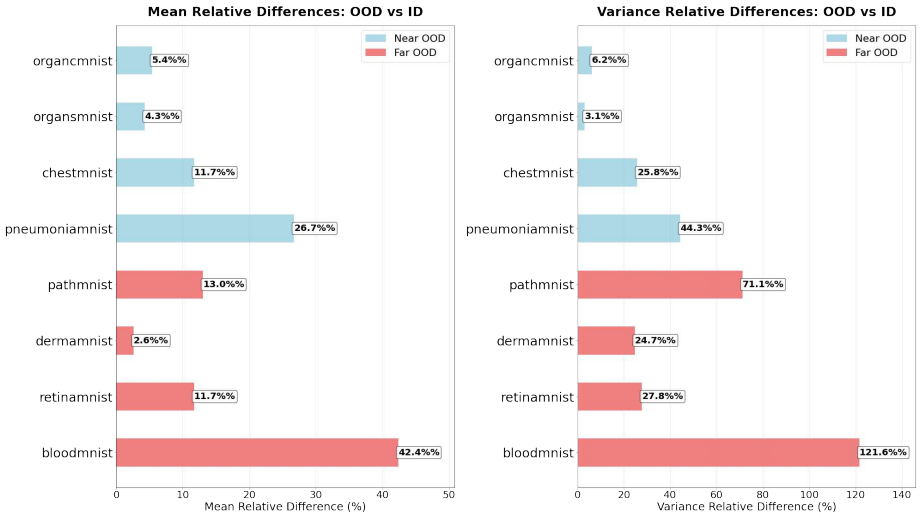



Figure 2: Relative Discrepancy in Activation Statistics: OOD vs ID

### *3.2. Overview of AFDS*

Our proposed Activation Fluctuation-Driven Scaling (AFDS) framework is a hierarchical processing pipeline designed to enhance OOD detection performance in complex medical imaging diagnosis. As illustrated in Figure 1, the system accepts multi-source medical images, including In-Distribution (ID), Covariate-Shift ID (CS-ID), Near-OOD, and Far-OOD samples. The processing pipeline comprises a shared feature extractor, an AFDS module that performs fluctuation-aware adaptive scaling, a logit output stage generating both maximum logit and other logits, and finally an OOD detection scorer that distinguishes ID from OOD samples based on the refined logit values.

### *3.3. Activation Fluctuation-Driven Scaling*

The first core innovation is an activation fluctuation scaling mechanism that addresses limitations of conventional feature scaling methods, particularly their inability to dynamically adapt to the varying impact of activation fluctuations between ID and OOD samples. This limitation is particularly salient given the significant statistical discrepancies observed in the activation patterns. As evidenced in Figure 2, the relative differences in both mean and variance of the activations can be substantial for OOD

data. For instance, while a near-OOD sample from organcmnist shows a modest 5.4% shift in mean activation, a far-OOD sample from bloodmnist exhibits a dramatic 42.4% shift. Similarly, variance discrepancies range from 6.2% for near-OOD organcmnist to a striking 71.1% for far-OOD pathmnist. Our proposed mechanism directly leverages these statistical insights to dynamically calibrate feature scaling, enabling more sensitive and adaptive detection of distributional shifts.

#### *3.3.1. Feature Representation*

Consider a standard classification model with a feature extractor $f(\cdot)$ that maps input $\mathbf{x}$ to a D-dimensional activation vector $\mathbf{a} = f(\mathbf{x}) \in \mathbb{R}^D$. A linear classifier parameterized by weight matrix W and bias vector $\mathbf{b}$ then produces the logit vector as shown in Eq.(3):

$$\mathbf{z} = W(\mathbf{a}) + \mathbf{b}, \tag{3}$$

where $\mathbf{z} \in \mathbb{R}^K$ denotes the logit vector.

#### *3.3.2. Construction of the Variance-Aware Scaling Factor r*

AFDS introduces a dynamic scaling factor defined in Eq.(4):

$$r = \frac{Q}{Q_p} \times (1 + \beta \cdot \mathrm{Var}(\mathbf{a})), \tag{4}$$

The key parameters in Eq. (5) are defined by the following mathematical expressions (5) and (5):

$$Q = \sum_{j=1}^{D} a_j \quad \text{and} \quad Q_p = \sum_{a_j > P_p(\mathbf{a})} a_j, \tag{5}$$

$$\mathrm{Var}(\mathbf{a}) = \frac{1}{D} \sum_{j=1}^{D} (a_j - \bar{a})^2, \tag{6}$$

where $\bar{a} = Q/D$ denotes the mean activation. Here, $\beta$ is the variance sensitivity coefficient that controls the influence of variance and $p$ is the percentile parameter that determines the activation range to calculate $Q_p$.

The combination enables adaptive scaling: when $\mathrm{Var}(\mathbf{a}) \to 0$ (characteristic of ID samples), $r \approx Q/Q_p$, providing stable scaling similar to conventional methods;

when Var(**a**) increases (characteristic of OOD samples), the scaling factor increases accordingly, enhancing sensitivity to anomalous patterns.

### *3.3.3. Construction of the Mean-Aware Scaling Factor $r_{mean}$*

Inspired by the design of the variance-aware scaling factor in AFDS, we propose a mean-aware dynamic scaling factor method. This method enhances the model's sensitivity to distributional shifts by incorporating the statistical properties of activation means.

The mean-aware scaling factor $r_{\text{mean}}$ is defined in Eq.(7):

$$r_{\text{mean}} = \frac{Q}{Q_p} \times (1 + \alpha \cdot (\bar{a} - \mu_{\text{ref}})), \tag{7}$$

The key parameters in Eq. (7) are defined by the following mathematical expressions (8) and (9):

$$Q = \sum_{j=1}^{D} a_j \quad \text{and} \quad Q_p = \sum_{a_j > P_p(\mathbf{a})} a_j, \tag{8}$$

$$\bar{a} = \frac{1}{D} \sum_{j=1}^{D} a_j = \frac{Q}{D}, \tag{9}$$

$$\mu_{\text{ref}} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \bar{a}_i, \tag{10}$$

where $\bar{a}$ denotes the mean activation of the current sample, and $\mu_{\text{ref}}$ represents the average of the mean activations across all samples in the training set, serving as a reference baseline. Here, $\alpha$ is the *mean sensitivity coefficient* that controls the influence of the mean deviation, $p$ is the percentile parameter that determines the activation range for calculating $Q_p$, and $\cdot$ denotes the absolute value operation, ensuring the positivity of the deviation measure.

This design enables adaptive scaling: when $\bar{a} - \mu_{\text{ref}} \to 0$ (characteristic of ID samples), $r_{\text{mean}} \approx Q/Q_p$, providing stable scaling similar to conventional methods; when $\bar{a} - \mu_{\text{ref}}$ increases (characteristic of OOD samples where the overall activation level shifts systematically), the scaling factor increases accordingly, thereby enhancing sensitivity to anomalous patterns. The coefficient $\alpha$ is tunable to balance the contribution

of the mean deviation against the base scaling factor.

#### *3.3.4. Scaled Logit Vector*

The scaled logit vector is computed through the transformation defined in Eq.(11):

$$\begin{aligned} \mathbf{z}' &= W \cdot (\mathbf{a} \circ s_f(\mathbf{a})) + \mathbf{b}, \\ \text{where } s_f(\mathbf{a})_j = \exp(r) \quad \text{and} \quad r &= \frac{Q}{Q_p} \times (1 + \beta \cdot \text{Var}(\mathbf{a})), \end{aligned} \tag{11}$$

This formulation applies the scaling through the activation modulation term $s_f(\mathbf{a})_j$, which preserves the relative relationships between logit components while adaptively adjusting their magnitudes based on the fluctuation characteristics.

### *3.4. Adaptive Weighted Logit Scoring*

The second innovation addresses limitations of existing logit-based scoring functions, particularly in medical imaging where subtle inter-class differences and significant covariate shifts make single-logit methods like MaxLogit inadequate. Our approach incorporates both the maximum logit (reflecting confidence in the dominant class) and information from other logits (quantifying prediction consistency).

#### *3.4.1. Score Components*

Based on the scaled logit vector $\mathbf{z}'$, we define two key scoring components. The first is the Maximum Logit, serving as a confidence measure, defined in Eq.(12):

$$\text{MaxLogit} = \max(\mathbf{z}'), \tag{12}$$

representing the confidence of the model in its primary prediction. The second is the OtherLogit, serving as a consistency measure, defined in Eq.(13):

$$\text{OtherLogit} = \frac{1}{N-1} \sum_{j=2}^{N} \left( z'_{(1)} - z'_{(j)} \right), \tag{13}$$

where $z'_{(1)}$ denotes the largest component (i.e., the maximum logit), $z'_{(j)}$ represents the j-th largest component in the sorted vector, and N (with $2 \leq N \leq K$) controls the number

of top logits included in the calculation.

### 3.4.2. Adaptive Weighted Scoring Function

The final AFDS score combines both components through adaptive weighting as formalized in Eq.(14):

$$S_{\text{AFDS}}(\mathbf{x}) = \alpha \cdot \max(\mathbf{z}') + (1 - \alpha) \cdot \frac{1}{N-1} \sum_{j=2}^{N} \left( z'_{(1)} - z'_{(j)} \right), \tag{14}$$

where $\alpha \in [0, 1]$ is an adaptive weighting coefficient that balances the contributions of confidence and consistency measures, tuned on a validation set.

The scoring function is designed with the following rationale. The MaxLogit component captures the model's confidence in its primary prediction, with ID samples typically generating higher values than OOD samples. The OtherLogit component quantifies the consistency of the prediction by measuring the average gap between the dominant logit and the other top logits. ID samples usually show tighter logit distributions, resulting in smaller gaps that indicate coherent predictions, while OOD samples tend to exhibit larger disparities. Adaptive weighting through $\alpha$ allows the final score to emphasize either confidence or consistency based on specific domain requirements. The OOD detection decision is finalized using the rule defined in Eq.(2).The corresponding rejection region is $\mathcal{R} = \{\hat{\mathbf{x}} : S_{\text{AFDS}}(\hat{\mathbf{x}}) < \gamma\}$.

The AFDS framework frushes activation flustuation scaling with adaptive logit weighting to create a coherent OOD detection system particularly suited for medical imaging applications that demand high reliability and robustness to distributional shifts.

## 4. Experiments

### 4.1. Set Up

**Dataset**. For out-of-distribution (OOD) detection in medical imaging, we follow the MedMNIST benchmark protocol, using OrganAMNIST as the in-distribution (ID) dataset. OrganAMNIST consists of abdominal CT axial view images employed for

Table 1: Overview of MedMNIST 2D benchmark datasets used in this study

| **Category** | **Dataset** | **Description** | **# of Samples** |
|---|---|---|---|
| ID | OrganAMNIST [33, 34] | Abdominal CT axial view | 17,778 |
| Near OOD | OrganCMNIST [33, 34] | Abdominal CT coronal view | 8,268 |
| | OrganSMNIST [33, 34] | Abdominal CT sagittal view | 8,829 |
| | ChestMNIST [35] | Chest X-ray of patients having various thoracic diseases | 22,433 |
| Far OOD | PneumoniaMNIST [36] | Pediatric chest X-ray healthy or with pneumonia | 624 |
| | PathMNIST [37] | Histological images of colorectal cancer tissues | 7,180 |
| | DermaMNIST [38] | Dermatoscopic images of common pigmented skin lesions | 2,005 |
| | RetinaMNIST [39] | Retina fundus images with different levels of diabetic retinopathy severity | 400 |
| | BloodMNIST [40] | Microscopic images of different blood cell types | 3,421 |

multi-class classification, covering 11 distinct categories of abdominal organs with a total of 17,778 images. The OOD data are divided into "Near-OOD" and "Far-OOD" categories, enabling evaluation of model performance under varying levels of distributional shift. Table 1 provides an overview of OrganAMNIST, detailing its 11 abdominal organ categories. This dataset serves as the primary reference for model training and evaluation, representing the core domain where optimal performance is expected. Its standardized divisions for training, validation, and testing ensure consistent benchmarking under controlled conditions. Focusing on CT axial imaging of abdominal organs, it provides the context for comparative assessment against near- and far-out-of-distribution (OOD) datasets.

The near-OOD category includes four datasets that introduce moderate domain shifts from the ID data (OrganAMNIST). These datasets share similarities in anatomical context or imaging technology but introduce controlled variations in imaging perspective, patient demographics, or clinical pathology. This category tests the model's robustness in handling clinically relevant distribution shifts, where semantic features partially overlap with the ID domain. Specifically, it includes: (1) OrganCMNIST and OrganSMNIST, which provide abdominal CT images from coronal and sagittal planes, respectively, differing only in imaging perspective from the ID's axial view; and (2) ChestMNIST and PneumoniaMNIST, which switch the imaging modality to X-ray while maintaining thoracic anatomical relevance. PneumoniaMNIST focuses specifically on pediatric pneumonia cases. These datasets challenge the model to generalize across clinically distinct imaging scenarios.

The far-OOD category includes four datasets that differ significantly from the ID data in terms of imaging modality, anatomical region, tissue type, and clinical application. These datasets represent extreme domain shifts, testing the model's robustness against substantial changes. The far-OOD category comprises: (1) PathMNIST, which transitions from CT imaging to histopathology of colorectal cancer tissue; (2) DermaMNIST, featuring dermoscopic images of pigmented skin lesions; (3) RetinaMNIST, focused on retinal fundus images for diabetic retinopathy assessment; and (4) BloodMNIST, consisting of microscopic images of blood cells. The large domain gap between these datasets and the ID data rigorously tests model generalization in diverse

clinical contexts.

The collection includes approximately 110,000 samples and follows standardized evaluation protocols for consistent performance assessment. This hierarchical structure enables systematic analysis of model performance degradation as domain shifts progress from ID to near-OOD to far-OOD.

**Baseline and metrics**. Baseline methods include SCALE [14], ReAct [27], ASH [31], ODIN [41], MSP [25], TempScale [42], EBO [43], MaxLogit [29] and LogitGap [44]. The OOD detection task uses AUROC and FPR95 as primary evaluation metrics. AUROC measures the general ability of the model to distinguish between ID and OOD data, while FPR95 characterizes the proportion of OOD samples misclassified as ID when the ID recall rate reaches 95%. A lower value of FPR95 indicates a better suppression of false-positive OOD signals under high recall constraints. These metrics collectively capture the model's robustness to distribution shifts.

**Hyperparameter Settings and Backbone Network Configuration**. In our experiments, we tuned key hyperparameters to optimize model performance. These include the parameter $N$ controlling the number of logits in the LogitGap calculation, the weight threshold $\tau$ for adaptive weighting, the percentile parameter $p$ for scaling, the mean-aware and variance sensitivity coefficient $\beta$. All experiments utilized a ResNet-18 backbone network, ensuring consistent and comparable feature extraction in medical imaging datasets. The ResNet-18 architecture provided a robust foundation, while optimized hyperparameters ensured the effectiveness of our mean-aware and variance-aware scaling approach.

**Main Results**. We used OrganAMNIST as the ID dataset and ResNet18 as the backbone network for the experiments. Table 2 presents the detailed comparison results of all methods in terms of near-OOD, far-OOD, and average performance. Based on the results presented in Table 2, our proposed AFDS method demonstrates a better overall performance compared to all baseline methods. It exhibits significant advantages in both Far OOD and Near OOD detection tasks .

The AFDS method demonstrates particularly strong performance in Far OOD detection, achieving an FPR95 of 19.78% and an AUROC of 92.37%. Compared to the SCALE method (FPR95: 49.00%, AUROC: 84.20%), which represents one of

Table 2: Performance comparison of AFDS with other baselines on the MedMNIST dataset

| | Near OOD | | Far OOD | | Average | |
|---|---|---|---|---|---|---|
| Methods | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| SCALE [14] | 23.05 | 89.63 | 49.00 | 84.20 | 36.03 | 86.92 |
| ReAct [27] | 46.60 | 84.65 | 26.05 | 88.97 | 36.33 | 86.81 |
| ASH [31] | 49.00 | 84.20 | 27.17 | 88.53 | 38.09 | 86.37 |
| ODIN [41] | 48.55 | 84.46 | 27.06 | 89.28 | 37.81 | 86.87 |
| MSP [25] | 64.45 | 83.01 | 22.73 | 90.13 | 43.59 | 86.57 |
| TempScale [42] | 91.24 | 54.73 | 62.35 | 81.38 | 76.80 | 68.06 |
| EBO [43] | 90.55 | 56.02 | 62.75 | 84.45 | 76.65 | 70.24 |
| MaxLogit [29] | 49.01 | 84.18 | 30.48 | 87.30 | 39.75 | 85.74 |
| LogitGap [44] | 48.72 | 84.33 | 27.70 | 88.93 | 38.21 | 86.63 |
| AFDS(ours) | **42.89** | **85.16** | **17.51** | **93.45** | **30.20** | **89.35** |

the better-performing baselines, AFDS achieves a significant reduction in FPR95 by 29.22% and an improvement in AUROC by 8.17%. This indicates that AFDS has a stronger ability to identify abnormal samples that exhibit substantial distributional differences from the ID data.

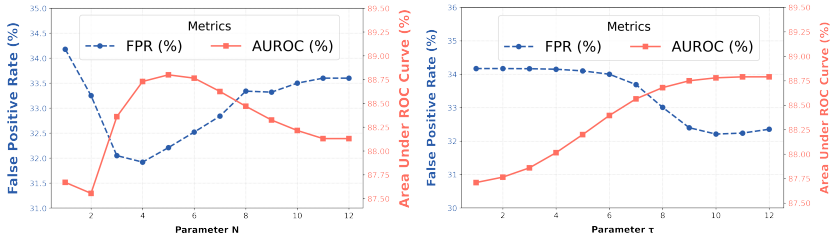
In the more challenging Near-OOD detection task, several baseline methods show considerable limitations. TempScale and EBO, for instance, report FPR95 values above 90% (91.24%, 90.55%, and 90.78%, respectively), while their AUROC values remain below 57% (54.73%, 56.02%, and 55.60%, respectively). These methods struggle with subtle distribution shifts, a common challenge in complex data scenarios like medical imaging. In contrast, AFDS consistently outperforms them, achieving an FPR95 of 43.58% and an AUROC of 85.75%, significantly surpassing the baseline methods.

Overall, AFDS shows a clear advantage in average performance. Its combined average FPR95 for both Far-OOD and Near-OOD tasks is 31.68%, and its average AUROC is 89.06%, both well above the performance of the baseline methods.

### *4.2. Sensitivity Analysis*

Figure 3(a) shows the impact of the $N$ on detection performance. A threshold of 3 strikes the optimal balance between reducing FPR95 and improving AUROC. Both higher and lower thresholds lead to performance degradation, emphasizing the importance of selecting an appropriate logit weighting threshold.

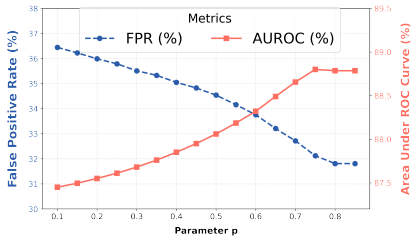
Figure 3(b) illustrates the effect of the adaptive weighting. A threshold value of



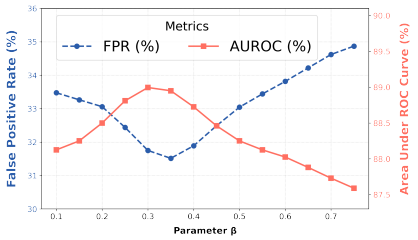

(a) sensitivity of $N$




(b) sensitivity of $\tau$




(c) sensitivity of $p$




(d) sensitivity of $\beta$

Figure 3: The sensitivity analysis of $N$,$\tau$,$p$ and $\beta$.

17 achieves the best performance, highlighting that a well-chosen coefficient is crucial for balancing the contributions of different logit components in the adaptive weighting scheme.

Figure 3(c) explores the influence of the percentile parameter. The optimal value of 0.85 achieves the highest AUROC of 88.80%, effectively filtering out noisy confidence estimates while retaining discriminative information.

Figure 3(d) presents the effect of the scaling factor. A scaling factor of 0.3 yields the best performance, with an FPR95 of 31.75% and an AUROC of 89.00%. Performance remains stable across a broad range of values, demonstrating the robustness of our method to parameter variations.

### *4.3. Ablation Experiments*

Table 3 summarizes the results of the ablation analysis, demonstrating the contributions of each component in our AFDS framework.

Ablation studies demonstrate the individual contribution of each component to overall performance. The removal of the Max-Item component leads to substantial per-

Table 3: Ablation analysis of AFDS components on the MEDMNIST dataset

| | Near OOD | | Far OOD | | Average | |
|---|---|---|---|---|---|---|
| Method | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| Without Max-Item | 46.57 | 84.55 | 22.15 | 90.87 | 34.36 | 87.71 |
| Without Gap-Item | 44.41 | 85.16 | 20.93 | 91.42 | 32.67 | 88.29 |
| Without Fluctuation-Item | 48.77 | 84.47 | 23.45 | 90.15 | 36.11 | 87.31 |
| Without Mean-Item | 43.58 | 85.75 | 19.78 | 92.37 | 31.68 | 89.06 |
| Without Variance-Item | 43.58 | 85.33 | 18.67 | 93.11 | 31.13 | 89.22 |
| AFDS (Ours) | **42.89** | **85.16** | **17.51** | **93.45** | **30.20** | **89.35** |

formance degradation, increasing the average FPR95 by 4.16 percentage points (from 30.20 to 34.36) and decreasing the average AUROC by 1.61 percentage points (from 89.32 to 87.71). Similarly, excluding the Gap-Item component results in an average increase in FPR95 of 2.47 percentage points and an average decrease in AUROC of 1.03 percentage points. Most notably, the absence of the Fluctuation-Item component produces the most severe performance decline, elevating the average FPR95 by 5.91 percentage points and diminishing the average AUROC by 2.01 percentage points. The removal of either the Mean-Item or Variance-Item components, which are essential for capturing statistical discrepancies, also leads to measurable performance drops. Specifically, omitting the Mean-Item increases the average FPR95 by 1.48 percentage points, while excluding the Variance-Item raises it by 0.93 percentage points. These findings collectively highlight the critical importance of all components within our proposed framework, with the adaptive weighting mechanism and activation fluctuation scaling being particularly vital to achieving optimal OOD detection performance in medical imaging applications.

### *4.4. Generalization Analysis*

**Dataset**. Table 4 provides a systematic domain division of the MIDOG dataset [48, 49], a multi-domain collection for mitotic figure detection. The dataset includes region-of-interest images from 503 histological samples across seven tumor types. The categorization is designed to assess model robustness to different types of data distribution shifts. The dataset is divided into four main categories: ID: Represented by Domain 1a, this category serves as the benchmark. It includes images from human breast carcinoma, captured with the same imaging hardware and staining protocol, exhibiting no

Table 4: MIDOG Dataset Domain Division

| Category | Dataset(s) | Description |
|---|---|---|
| **ID** | **Domain 1a** | The images are Extracted from 50 whole slide images: 451 mitotic cell crops, 724 imposter cell crops, and 1153 additional crops. derived from **human breast carcinoma**. The images were generated using the same imaging hardware as the In-Distribution (ID) dataset and originate from the same institute, employing the same staining protocol, and showing no apparent covariate shifts. Serves as the in-distribution benchmark dataset. |
| **CS-ID** | **Domains 1b, 1c** | In-Distribution data with covariate shifts. Compared to Domain 1a, these domains originate from the same cancer type (human breast carcinoma) but exhibit variations in imaging hardware or staining protocols, leading to covariate shifts. |
| **Near-OOD** | **Domains 2, 3, 4, 5, 6a, 6b, 7** | Near out-of-distribution data with **semantic shift**. These domains encompass seven different cancer types and two species (human and canine), resulting in semantic cell type shifts. Among them, Domains 2, 3, 4, 6a, and 6b also exhibit covariate shifts; Domains 5 and 7 show no apparent covariate shifts. |
| **Far-OOD** | **CCAgT** [45, 46]<br>**FNAC** [47] | Far out-of-distribution datasets, typically referring to datasets that differ significantly from the in-distribution data in task, modality, or scenario. |

apparent covariate shifts. CS-ID (In-Distribution with Covariate Shift): Comprising Domains 1b and 1c, this category introduces covariate shifts. Although the type of cancer (human breast carcinoma) remains the same as the ID domain, these domains exhibit variations in imaging hardware or staining protocols. Near-OOD (Near Out-of-Distribution): This category, encompassing Domains 2 through 7, introduces semantic shifts. The domains represent seven different types of cancer (e.g. breast carcinoma,

Table 5: Performance comparison of AFDS with other baselines on the MIDOG dataset

| | cs-ID | | Near OOD | | Far OOD | | Average | |
|---|---|---|---|---|---|---|---|---|
| Methods | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ |
| SCALE [14] | 97.01 | 53.4 | 91.52 | 54.88 | 72.11 | 82.03 | 86.88 | 63.44 |
| ReAct [27] | 94.02 | 53.61 | 93.97 | 55.38 | 52.19 | 84.86 | 80.06 | 64.62 |
| ASH [31] | 96.02 | 54.39 | 90.15 | 56.29 | 67.53 | 84.32 | 84.57 | 65.00 |
| ODIN [41] | 94.22 | 56.49 | 86.45 | 61.01 | 70.32 | 85.08 | 83.66 | 67.53 |
| MSP [25] | 93.82 | 53.26 | 91.24 | 54.53 | 62.75 | 80.91 | 82.60 | 62.90 |
| TempScale [42] | 93.82 | 53.47 | 91.24 | 54.73 | 62.35 | 81.38 | 82.47 | 63.19 |
| EBO [43] | 94.02 | 54.39 | 90.55 | 56.02 | 62.75 | 84.45 | 82.44 | 64.95 |
| MaxLogit [29] | 94.42 | 54.15 | 90.78 | 55.60 | 62.95 | 82.98 | 82.72 | 64.24 |
| LogitGap [44] | 93.82 | 54.09 | 91.01 | 55.38 | 61.16 | 83.02 | 82.00 | 64.16 |
| **AFDS (ours)** | **91.63** | **55.38** | **89.36** | **56.11** | **42.63** | **84.56** | **74.54** | **65.35** |

lung carcinoma, lymphosarcoma) and two species (human and canine). In particular, Domains 2, 3, 4, 6a and 6b also exhibit covariate shifts, while Domains 5 and 7 do not. Far-OOD (Far Out-of-Distribution): This category includes datasets (CCAgT and FNAC 2019) that differ significantly from the in-domain data in task, modality, or scenario, representing a more substantial domain shift. This structured division facilitates the study of domain generalization in computational pathology, helping to address performance degradation caused by domain shifts originating from different types of tumor, laboratories, and scanners.

**Experiment results**. As shown in Table 5 presents the detailed performance of AFDS on the MIDOG dataset. Our method achieved an F1-Score of 81.88%, demonstrating strong discriminative power. In OOD detection, AFDS achieved an AUROC of 89.36% and an FPR95 of 56.11% in near-OOD detection, and an AUROC of 84.56% and an FPR95 of 42.63% in far-OOD detection, significantly outperforming all baseline methods.

AFDS consistently outperformed the baseline methods across multiple challenging scenarios—CS-ID, Near-OOD, and Far-OOD—indicating its superior generalization capability. This balanced performance highlights the method's robustness to various distribution shifts.

From a clinical perspective, this generalization is critical. Real-world clinical data often exhibit complex and unpredictable distribution shifts. A reliable model must handle internal covariate shifts, detect subtle semantic differences, and maintain strong discriminative power against significant OOD samples. The superior performance of

AFDS across these metrics demonstrates its effectiveness in open-world clinical environments, where multiple distribution shifts can coexist. This capability reduces diagnostic risks associated with model failure in specific scenarios, providing a solid foundation for clinical adoption.

## 5. Conclusion

This study addressed critical challenges in post-hoc Out-of-Distribution (OOD) detection for medical imaging, particularly the inability of existing methods to capture activation fluctuation patterns and fully utilize model output information. We introduced Activation Fluctuation-Driven Scaling (AFDS), a novel method designed for medical imaging. The core innovation of AFDS was its dual-mechanism approach. First, we proposed a sample-level activation fluctuation-driven scaling mechanism that adaptively adjusted scaling intensity based on each sample's activation fluctuation. This method improved distribution separability, overcoming the limitations of traditional global scaling techniques, which had failed to capture discriminative features. Second, we introduced a discriminative scoring function that combined the confidence of the dominant class, represented by the maximum logit, with inter-class discriminative information, captured by the difference between the maximum logit and the remaining logits. This approach leveraged structural differences in the model's output distribution more effectively.

Extensive experiments on medical imaging benchmarks, including MEDMNIST and MIDOG, demonstrated that AFDS achieved leading performance in Near-OOD, Far-OOD, and the challenging covariate shift (CS-ID) detection tasks, significantly improving key metrics such as AUROC. These results confirmed the method's robustness and generalizability. Looking forward, the AFDS framework offers a reliable and efficient path for deep learning models in medical image analysis, particularly in open-world environments. Future work will explore integrating this method with more complex architectures, such as visual-language models, and adapting it to evolving clinical scenarios.

## Acknowledgment

This work is supported by the National Natural Science Foundation of China (U23A20389).

## References

[1] Y. Sun, X. Wang, D. Peng, Z. Ren, X. Shen, Hierarchical hashing learning for image set classification, IEEE Transactions on Image Processing 32 (2023) 1732–1744.

[2] Y. Sun, D. Peng, H. Huang, Z. Ren, Feature and semantic views consensus hashing for image set classification, in: Proceedings of the 30th ACM International conference on multimedia, 2022, pp. 2097–2105.

[3] Z. Huang, M. Zhu, X. Xia, L. Shen, J. Yu, C. Gong, B. Han, B. Du, T. Liu, Robust generalization against photon-limited corruptions via worst-case sharpness minimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16175–16185.

[4] Y. Yang, D. Zhou, D. Zhan, H. Xiong, Y. Jiang, J. Yang, Cost-effective incremental deep model: Matching model capacity with the least sampling, IEEE Trans. Knowl. Data Eng. 35 (4) (2023) 3575–3588.

[5] K. Wang, Q. Ma, C. Shen, J. Lu, Application of uncertainty to out-of-distribution detection for autonomous driving perception safety, IEEE Transactions on Intelligent Transportation Systems.

[6] Y. Hahn, J. Voets, A. Königsfeld, H. Tercan, T. Meisen, Out of distribution detection for efficient continual learning in quality prediction for arc welding, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 5699–5706.

[7] Z. Han, X.-J. Gui, H. Sun, Y. Yin, S. Li, Towards accurate and robust domain adaptation under multiple noisy environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (5) (2022) 6460–6479.

[8] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, International Journal of Computer Vision 132 (12) (2024) 5635–5662.

[9] V. Narayanaswamy, Y. Mubarka, R. Anirudh, D. Rajan, J. J. Thiagarajan, Exploring inlier and outlier specification for improved medical ood detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4589–4598.

[10] D. Karimi, A. Gholipour, Improving calibration and out-of-distribution detection in deep models for medical image segmentation, IEEE transactions on artificial intelligence 4 (2) (2022) 383–397.

[11] Y. Li, X. Kong, J. Xie, G. Ver Steeg, Y. Zhang, Denoising diffusion wavelet models for zero-shot medical image translation, Knowledge-Based Systems 324 (2025) 113800. doi:10.1016/j.knosys.2025.113800.

[12] Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, F. Liu, Is out-of-distribution detection learnable?, Advances in Neural Information Processing Systems 35 (2022) 37199–37213.

[13] Y.-C. Hsu, Y. Shen, H. Jin, Z. Kira, Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10951–10960.

[14] K. Xu, R. Chen, G. Franchi, A. Yao, Scaling for training time and post-hoc out-of-distribution detection enhancement, arXiv preprint arXiv:2310.00227.

[15] Z. Hong, Y. Yue, Y. Chen, L. Cong, H. Lin, Y. Luo, M. H. Wang, W. Wang, J. Xu, X. Yang, et al., Out-of-distribution detection in medical image analysis: A survey, arXiv preprint arXiv:2404.18279.

[16] T. Ouyang, Y. Seo, I. Echizen, Textual out-of-distribution (ood) detection for llm quality assurance, Knowledge-Based Systems 310 (2025) 112975.

[17] T. Yin, C. Zhao, X. Liu, M. Shao, Out-of-distribution detection in heterogeneous graphs via energy propagation, Knowledge-Based Systems (2025) 113667.

[18] Y. Chen, Y. Wu, T. Wang, S. Zhang, D. Zhu, C. Liu, Mitigating covariance overfitting in out-of-distribution detection through intrinsic parameter learning, Knowledge-Based Systems 316 (2025) 113318.

[19] X. Du, Z. Wang, M. Cai, Y. Li, Vos: Learning what you don't know by virtual outlier synthesis, arXiv preprint arXiv:2202.01197.

[20] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, Y. Li, Mitigating neural network overconfidence with logit normalization, in: International conference on machine learning, PMLR, 2022, pp. 23631–23644.

[21] D. Hendrycks, M. Mazeika, T. Dietterich, Deep anomaly detection with outlier exposure, arXiv preprint arXiv:1812.04606.

[22] J. Zhang, N. Inkawhich, R. Linderman, Y. Chen, H. Li, Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 5531–5540.

[23] S. Cao, Z. Zhang, Deep hybrid models for out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4733–4743.

[24] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in neural information processing systems 31.

[25] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv preprint arXiv:1610.02136.

[26] R. Huang, A. Geng, Y. Li, On the importance of gradients for detecting distributional shifts in the wild, Advances in Neural Information Processing Systems 34 (2021) 677–689.

[27] Y. Sun, C. Guo, Y. Li, React: Out-of-distribution detection with rectified activations, Advances in Neural Information Processing Systems 34.

[28] Y. Yuan, R. He, Y. Dong, Z. Han, Y. Yin, Discriminability-driven channel selection for out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26171–26180.

[29] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, D. Song, Scaling out-of-distribution detection for real-world settings, arXiv preprint arXiv:1911.11132.

[30] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, Y. Li, Delving into out-of-distribution detection with vision-language representations, Advances in neural information processing systems 35 (2022) 35087–35102.

[31] A. Djurisic, N. Bozanic, A. Ashok, R. Liu, Extremely simple activation shaping for out-of-distribution detection, arXiv preprint arXiv:2209.09858.

[32] Y. Zhu, Y. Chen, C. Xie, X. Li, R. Zhang, H. Xue, X. Tian, bolun zheng, Y. Chen, Boosting out-of-distribution detection with typical features (2022). arXiv:2210.04200.

[33] X. Xu, F. Zhou, B. Liu, D. Fu, X. Bai, Efficient multiple organ localization in ct image using 3d region proposal network, IEEE transactions on medical imaging 38 (8) (2019) 1885–1898.

[34] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, et al., The liver tumor segmentation benchmark (lits), Medical image analysis 84 (2023) 102680.

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[36] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, cell 172 (5) (2018) 1122–1131.

[37] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, et al., Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study, PLoS medicine 16 (1) (2019) e1002730.

[38] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific data 5 (1) (2018) 1–9.

[39] R. Liu, X. Wang, Q. Wu, L. Dai, X. Fang, T. Yan, J. Son, S. Tang, J. Li, Z. Gao, et al., Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge, Patterns 3 (6).

[40] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, J. Rodellar, A dataset of microscopic peripheral blood cell images for development of automatic recognition systems, Data in brief 30 (2020) 105474.

[41] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, arXiv preprint arXiv:1706.02690.

[42] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks (Jun. 2017). arXiv:1706.04599, doi:10.48550/ARXIV.1706.04599.
URL https://arxiv.org/abs/1706.04599

[43] W. Liu, X. Wang, J. D. Owens, Y. Li, Energy-based out-of-distribution detection (2021). arXiv:2010.03759.
URL https://arxiv.org/abs/2010.03759

[44] J. Liang, R. Hou, M. Hu, H. Chang, S. Shan, X. Chen, Revisiting logit distributions for reliable out-of-distribution detection, arXiv preprint arXiv:2510.20134.

[45] J. G. Atkinson Amorim, L. A. Buschetto Macarini, A. Victória Matias, A. Cerentini, F. Botelho De Miranda Onofre, A. Sherlley Casimiro Onofre, A. Von Wangenheim, A novel approach on segmentation of agnor-stained cytology images using deep learning, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2020, pp. 552–557. doi:10.1109/CBMS49503.2020.00110.

[46] João Gustavo Atkinson Amorim, Ccagt: Images of cervical cells with agnor stain technique (2022). doi:10.17632/WG4BPM33HJ.2.

[47] A. R. Saikia, K. Bora, L. B. Mahanta, A. K. Das, Comparative assessment of cnn architectures for classification of breast fnac images, Tissue and Cell 57 (2019) 8–14, eM in cell and tissues. doi:https://doi.org/10.1016/j.tice.2019.02.001.
URL https://www.sciencedirect.com/science/article/pii/S0040816618304397

[48] M. Aubreville, F. Wilm, N. Stathonikos, K. Breininger, T. A. Donovan, S. Jabari, M. Veta, J. Ganz, J. Ammeling, P. J. van Diest, R. Klopfleisch, C. A. Bertram, A comprehensive multi-domain dataset for mitotic figure detection, Scientific Data 10 (1). doi:10.1038/s41597-023-02327-4.

[49] M. Gutbrod, D. Rauber, D. W. Nunes, C. Palm, Openmibood: Open medical imaging benchmarks for out-of-distribution detection, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 25874–25886.