NERO: Explainable Out-of-Distribution Detection with Neuron-level Relevance in Gastrointestinal Imaging

Anju Chhetri[1], Jari Korhonen[3], Prashnna Gyawali[2], and Binod Bhattarai[3]

[1]NepAl Applied Mathematics and Informatics Institute for research, Nepal
[2]West Virginia University, USA
[3]University of Aberdeen, Aberdeen, UK
binod.bhattarai@abdn.ac.uk

Abstract. Ensuring reliability is paramount in deep learning, particularly within the domain of medical imaging, where diagnostic decisions often hinge on model outputs. The capacity to separate out-of-distribution (OOD) samples has proven to be a valuable indicator of a model's reliability in research. In medical imaging, this is especially critical, as identifying OOD inputs can help flag potential anomalies that might otherwise go undetected. While many OOD detection methods rely on feature or logit space representations, recent works suggest these approaches may not fully capture OOD diversity. To address this, we propose a novel OOD scoring mechanism, called NERO, that leverages neuron-level relevance at the feature layer. Specifically, we cluster neuron-level relevance for each in-distribution (ID) class to form representative centroids and introduce a relevance distance metric to quantify a new sample's deviation from these centroids, enhancing OOD separability. Additionally, we refine performance by incorporating scaled relevance in the bias term and combining feature norms. Our framework also enables explainable OOD detection. We validate its effectiveness across multiple deep learning architectures on the gastrointestinal imaging benchmarks Kvasir and GastroVision, achieving improvements over state-of-the-art OOD detection methods. Code Available: https://github.com/bhattarailab/NERO

Keywords: OOD · Neuron relevance · Gastrointestinal imaging · Explainable.

1 Introduction

Deep learning-based systems hold significant potential for enhancing computer-aided diagnostics [7] in medical image analysis. However, most of the deep learning models are trained under the closed-world assumption [11,20], where the training and testing data share the same distribution. This assumption limits their effectiveness in real-world deployment, as these models often encounter out-of-distribution (OOD) data and may produce high-confidence predictions

on such data [25,31]. For instance, a deep learning model trained on common gastrointestinal conditions in endoscopic images may misclassify rare diseases like blue rubber bleb nevus syndrome (BRBNS) or gastrointestinal stromal tumors (GISTs), which were absent from its training data. In such scenarios, the model may erroneously assign high confidence to incorrect predictions, leading to potential misdiagnoses. To ensure the trustworthiness of these systems, they must be capable of detecting and flagging OOD data so that the human expert could handle it safely. Similarly, the challenge of obtaining labeled datasets, particularly for rare pathologies, results in a long-tailed class distribution. This has driven interest in applying OOD detection techniques for pathology identification [33,15], where healthy data are modeled as in-distribution (ID), while unhealthy data are treated as OOD samples, making OOD detection increasingly relevant in medical deep learning.

In this work, we focus on post-hoc OOD detection, an unsupervised approach that does not require any model re-training or access to OOD data during training, making it highly practical for real-world deployment. These methods can generally be categorized into three types. Firstly, the *logit-based methods*, uses information from the model's logits. In [14], probabilities from softmax layers are utilized, while [23,13] refine this method by computing energy scores on the logits and using maximum logit value for OOD detection, respectively. Secondly, the *feature-based methods*, focuses on analyzing feature representations extracted from the model. In [21], Mahalanobis distance between feature vectors and their class-wise mean is computed, while [30] introduces activation rectification in the penultimate layer to enhance OOD detection. Similarly, in [28], the OOD score is computed as a weighted difference between the sum of distances to non-nearest centroids and the distance to the nearest centroid. Finally, the *gradient-based methods* include [17], where GradNorm uses vector norm of gradients, backpropagated from the KL divergence between the softmax output and a uniform probability. Recent works also advocate for the integration of multiple information to enhance OOD detection. For instance, [34] introduces a virtual logit that utilizes information from features, logits, and probabilities for OOD scoring.

Fundamentally a different perspective, and building on the observation that deep neural networks often assign specialized roles to individual neurons for particular classes [5,26], we present, for the first time, an approach that harnesses neuron-level relevance for OOD detection-*NERO*. Specifically, we cluster neuron-level relevance for samples in each class in the ID data to form representative centroids. We then introduce a novel *relevance distance* metric, which quantifies how far a new sample's relevance signature is from these class-centric centroids, thus serving as an effective OOD separability function. Unlike methods focusing purely on activation magnitudes or feature vectors, our approach considers how neurons collectively contribute to the final prediction. Additionally, we incorporate the scaled relevance in the bias term and combine feature norms to further refine OOD detection performance. Furthermore, since neuron-level relevance has already demonstrated its utility in explainable AI [24,16,6], our framework naturally extends to explainable OOD detection—an especially

critical feature in sensitive domains like medical imaging, where both reliability and interpretability are paramount. We evaluated our method on two different gastrointestinal datasets and two separate classifier backbones, ResNet-18 [12] and Data-efficient Image Transformer (DeiT) [32], demonstrating its superior performance compared to existing methods. Overall, our contributions are as follows:

1. We propose a novel explainable post-hoc OOD detection method, NERO, that leverages neuron-level relevance to analyze prediction-relevant patterns.
2. We demonstrate the effectiveness of our method through extensive empirical evaluations on challenging medical image benchmarks, showing superior performance compared to state-of-the-art OOD detection techniques.

2 Method

Problem Formulation: We consider a multi-class classification task, where we represent training and testing sets as $\mathcal{D}_t = \{(x_i, y_i)\}_{i=0}^{n_t}$ and $\mathcal{D}_e = \{(x_i, y_i)\}_{i=0}^{n_e}$ respectively. x_i represents an image and $y_i \in \{0,...,C\text{-}1\}$ is its corresponding class label, where C is the total number of classes. The data in $\mathcal{D}_t$ and $\mathcal{D}_e$ are assumed to be independent and identically distributed (iid) and are drawn from $\mathbb{P}_{\text{id}}$. We use g to represent a neural network trained on $\mathcal{D}_t$, parameterized by w. The feature extractor $f_w(\cdot)$ is derived from the penultimate layer of g. The aim of OOD detection is to derive a confidence score that allows us to assess whether a given test data (x) is drawn from $\mathbb{P}_{\text{id}}$ or from another distribution $\mathbb{P}_{\text{ood}}$ with unknown OOD class.

Our approach focuses on analyzing the contribution of individual neurons to model predictions when processing ID versus OOD samples. For any input x, we obtain the network predictions, $\hat{y} = g_w(x) \in \mathbb{R}^C$ and features from the penultimate layer, $a = f_w(x) \in \mathbb{R}^d$, where d denotes the total number of features (Fig. 1 (a)). To quantify the contribution of each neuron to the final prediction, we compute relevance scores for each neuron in the penultimate layer.

Relevance Estimation: The relevance is computed by summing over all neurons c in the final layer, where each neuron's contribution in the penultimate layer is weighted by the ratio of the specific connection's strength ($a_j w_{jc}$) to the total input received by neuron c ($\sum a_k w_{kc}$) (Fig.1 (b)). Higher relevance values indicate neurons that substantially influence the prediction outcome. For relevance calculation, we follow the definition of LRP-0 [4,24]:

$$r(x) = \sum_c \frac{a_j w_{jc}}{\sum_{0,k} a_k w_{kc}} \hat{y}_c, \forall j \in \{0, .., d\} \tag{1}$$

Following the approach of [24] an additional neuron $r_\beta(x) \in \mathbb{R}$ is introduced to represent the bias contribution, with a_0=1 and w_{0c} denoting the bias value. When neurons have non-zero bias terms, part of the relevance is injected or absorbed through the bias, inclusion of the bias neuron enables us to attribute relevance to all components influencing the model's output [4]. To quantify how

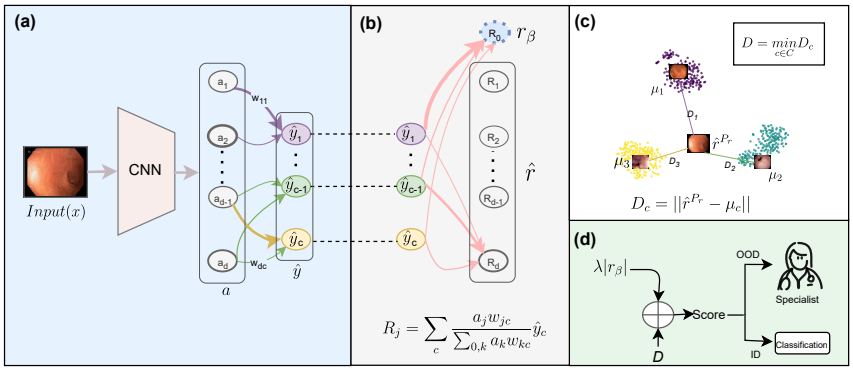

Fig. 1. Illustration of **NERO**. **(a)** Feature extraction: Input images are processed through a CNN to obtain feature (a) in the penultimate layer and class predictions ($\hat{y}$). **(b)** Relevance estimation: Neuron contributions are quantified by calculating relevance scores (R_j) based on connection strengths, with R_0 representing bias contribution. **(c)** Distance computation: The framework calculates distances (D_c) between test sample's relevance and class-wise centroids (μ_c) in principal component space. **(d)** OOD scoring: The detection mechanism combines the minimum distance between sample relevance patterns and class centroids (D) along with the absolute bias relevance term ($|r_\beta|$).

individual samples relate to representative class distributions – a key differentiator for ID examples from OOD examples – we compute class-wise relevance centroids.

Mean Relevance Estimation: For centroid calculations, we utilize only the first d features (i.e., neurons) to ensure neuron-wise comparison, while the bias neuron, r_β, due to their difference in magnitude is handled separately. Let $\hat{r}(x) \in \mathbb{R}^d$ represent the neuron-level relevance score of the input x from the $\mathcal{D}_t$, respectively, and we define a matrix $A \in \mathbb{R}^{n_t \times d}$ containing all $\hat{r}(.)$. We observed that these relevance patterns exhibited distinct class-specific characteristics. To capture these class-specific patterns, we compute class-wise mean relevance scores from the training set. First, to preserve the most informative variations while reducing noise and redundancy, we apply the Principal Component Analysis (PCA) to A and decompose it into principal space, defined by projection of a matrix $P_r \in \mathbb{R}^{d \times z}$, spanned by the first z columns.

$$\mu_c = \frac{1}{n_c} \sum_{i=1}^{n_c} \hat{r}(x)^{P_r} \tag{2}$$

where, $r(.)^{P_r}$ is the projection of $r(.)$ to P_r and n_c is the total number total number of samples belonging to class c.

Proposed OOD Score: With the mean relevance obtained, for any new sample x, we define the OOD score as:

$$S(x) = \min_{c \in \mathcal{C}} ||\hat{r}(x)^{P_r} - \mu_c|| + \lambda |r_\beta(x)| \tag{3}$$

Our scoring function incorporates two complementary components: the minimum distance between projected relevance patterns and class-wise means, and the weighted bias relevance term. The minimum distance measures the deviation of relevance patterns from learned class patterns, with smaller distances for ID samples and larger ones for OOD samples (Fig.1 (c)). Empirical analysis of plotted relevance scores for both feature and bias neurons revealed that the bias neuron also exhibited discriminative signals for ID-OOD separability. Capitalizing on this, we introduce the bias relevance term, weighted by λ, computed as the ratio of their mean values over the training set, to enhance detection capability while ensuring both terms contribute equally and remain comparable in scale (Fig.1 (d)).

$$\lambda = \frac{\mathbb{E}_{j \in \mathcal{D}_t}[||\hat{r}(x_j)^{P_r} - \mu_c||]}{\mathbb{E}_{j \in \mathcal{D}_t}[|r_\beta(x_j)|]} \tag{4}$$

Building on the insights of previous works like [9], which utilizes information from the null space of the weight matrix for outlier detection, we analyze neurons with lower relevance scores to capture subtle OOD attributes. A model trained on ID data primarily captures ID-specific features, with highly relevant neurons focusing on these characteristics. Introducing a scaling term that considers the norm of normalized features ($\hat{f}_w(.)$) from the less relevant neurons ensures that informative OOD signals are not overshadowed by dominant ID features. This refinement enhances the model's ability to capture OOD attributes, improving performance in distinguishing OOD from ID data.

$$S(x) = (\min_{c \in \mathcal{C}} ||\hat{r}(x)^{P_r} - \mu_c|| + \lambda |r_\beta(x)|) \cdot (\sum_{j \in B_k} |\hat{f}_w(x)_j|) \tag{5}$$

where $B_k = \{j \mid j \in \text{argmin}_{J \subseteq [d], |J|=k} |\hat{r}(x_i)_J|\}$ is the set of indices corresponding to the k least relevant neurons. With $S(\cdot)$ defined in Eqn 5, OOD detection is framed as a binary classification task with a threshold λ_S, that guides the decision process with those exceeding the threshold as OOD and the rest as ID. The threshold λ_S is often set for a 95% true positive rate on test set.

3 Experiments and Results

Datasets and Setup: We evaluated our method using two publicly available multi-class endoscopy datasets: Kvasir-v2 [29] and GastroVision [18]. These datasets provide comprehensive benchmarking in real-world settings with both balanced and unbalanced class distributions. The Kvasir-v2 dataset comprises 8,000 gastrointestinal tract images equally distributed across 8 classes (1,000 per class), representing pathological findings, anatomical landmarks, and endoscopic procedures. We designated three healthy anatomical landmarks-Pylorus, Z-line,

and Cecum-as ID classes, with the remaining five serving as OOD classes corresponding to abnormalities. The GastroVision dataset encompasses 27 classes across upper and lower GI categories, totaling 8,000 images. The class distribution is imbalanced, with samples per class ranging from 6 to 1,467. We categorized 11 normal and anatomical findings as ID data, while pathological and therapeutic findings were designated as OOD samples. For both datasets, we used an 80:20 train-test split, resulting in 800 training and 200 test images per class for Kvasir-v2 (2,400 training, 600 test total). For GastroVision, this yielded 3,804 training and 955 test images across all classes.

Implementation Details: We experiment with both CNN and transformer architectures, using ResNet-18 and DeiT [32] for image classification. For ResNet-18, we extract 512-dimensional feature vectors from the global average pooling layer before the final classification layer. For the transformer, we use the small DeiT model (DeiT-S) with a 16×16 patch size, where the [CLS] token embedding from the final transformer block (384-dimensional) serves as the feature representation. We initialize both models with ImageNet pre-trained weights and fine-tune them for classification. The models are optimized using Adam [19] with a $1e^{-4}$ learning rate. Input images from Kvasir-v2 and GastroVision are resized to 224×224 pixels. Training was performed on NVIDIA A100 GPUs for 20 epochs on Kvasir-v2 and 50 on GastroVision.

Evaluation Metrics: We evaluate our model using AUROC and FPR95. AUROC measures the overall performance of the model in distinguishing between classes. It is threshold-independent, with higher values indicating better performance. FPR95 quantifies the false positive rate when the true positive rate reaches 95%, with lower values indicating superior performance.

Baselines: We compared our method with 10 competitive methods MSP [14], ODIN [22], Energy [23], Entropy [8], MaxLogit [13], NECO [3], ViM [34], ReAct[30], GradNorm [17], and Mahalanobis following this setting[21].

3.1 Quantitative Results

NERO vs. Baselines: Table 1 presents our main results on ResNet-18 and DeiT for the Kvasir and GastroVision datasets. The best results are highlighted in bold, while the second-best results are underlined. NERO demonstrates competitive, or even superior, performance compared to state-of-the-art approaches. On Kvasir-v2, it achieves the lowest FPR95 (28.84% with ResNet-18, 18.96% with DeiT), outperforming all baselines. It also ranks among the top three in AUROC (90.76% with ResNet-18, 92.73% with DeiT). On GastroVision, NERO delivers the best results on both metrics with DeiT and second best with ResNet-18. Its consistent performance across different architectures and datasets highlights its robustness in reducing false positives while maintaining high detection accuracy.

The effect of hyperparameter: Our results show that performance remains robust across different numbers of bottom channels selected for feature scaling. As seen in Fig. 2, metrics stay stable over a broad range, though using very

Table 1. Performance comparison of different methods using ResNet-18 and DeiT on Kvasir-v2 and GastroVision. Metrics include AUROC (higher is better, ↑) and FPR95% (lower is better, ↓), both reported as percentages.

	ResNet-18				DeiT			
	Kvasir-v2		GastroVision		Kvasir-v2		GastroVision	
Method	AUC	FPR95	AUC	FPR95	AUC	FPR95	AUC	FPR95
MSP	90.3	41.72	66.93	90.56	87.05	40.18	70.0	90.74
ODIN	**91.77**	35.44	69.79	79.27	88.41	36.4	73.37	83.68
Energy	88.85	52.36	70.31	79.79	85.77	44.02	75.35	83.68
Entropy	90.38	41.86	67.37	87.32	87.2	39.94	70.34	90.19
MaxLogit	88.9	52.38	70.08	80.44	85.77	44.02	75.11	84.2
Mahalanobis	84.05	54.06	65.93	89.69	**94.50**	21.86	75.68	81.43
ViM	90.62	41.1	72.70	76.98	93.88	24.38	76.69	78.37
NECO	89.64	47.90	**79.81**	**71.61**	88.31	37.60	76.95	81.92
Energy+ReAct	86.57	53.78	61.93	83.86	83.49	46.84	73.42	83.22
GradNorm	85.33	54.68	62.55	90.5	71.33	57.8	54.85	88.68
NERO (ours)	90.76	**28.84**	75.95	74.33	92.73	**18.96**	**82.03**	**76.74**

few low-relevance channels can be suboptimal, as indicated by the relevance distribution.

3.2 Qualitative Results

Explainability Analysis: In this section, we demonstrate the explainability of our proposed OOD framework. To this end, we consider two cases comparing ID and OOD samples. For each category, we visualize the specific features that neurons attend to when processing ID and OOD samples. To provide visual insights into how the OOD detector distinguishes between ID and OOD data, we select the top four channels with the highest relevance scores for each ID class and visualize their relevance maps on both correctly classified ID samples and OOD samples. For visualization, we use Concept Relevance Propagation (CRP) [1] to compute channel-conditional relevance maps $R(\mathrm{x}|\mathrm{y})$, where x represents the input image and y represents a set of conditions.

The results are presented in Fig. 3. The left panel illustrates Case I, where we analyze Pylorus (ID class) and Polyp (OOD data), while the right panel depicts Z-line (ID class) and Ulcerative Colitis (OOD data). The top two rows show the corresponding relevance maps for ID and OOD samples, respectively. Notably, our framework successfully flags the second-row samples as OOD, and the visualizations reveal distinct patterns: ID samples exhibit focused, high-intensity activations around key anatomical landmarks, whereas OOD samples display markedly different attribution patterns, with lower and more diffuse intensity values across the channels. Further, to demonstrate the consistency of this pattern across our dataset, we present the mean relevance scores for both ID and OOD class plots for each case (Fig. 3, bottom), with channel ind‘ices sorted according to the mean relevance score of the ID class. We applied a moving average

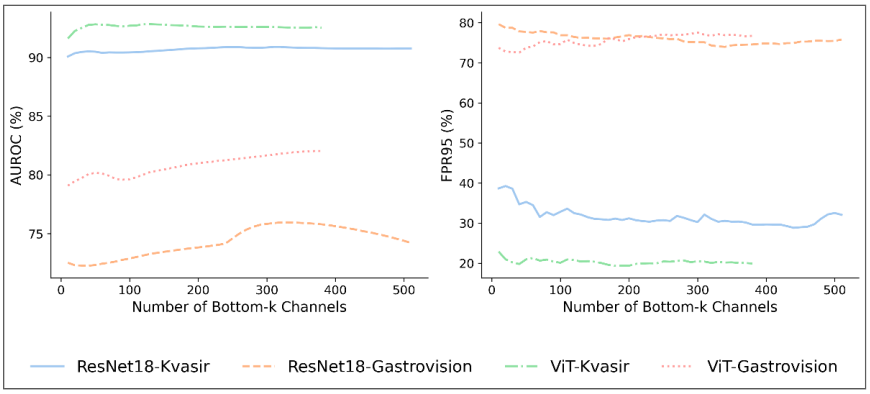

Fig. 2. Robustness to the number of bottom channels selected, with AUROC shown on the left and FPR95 on the right.

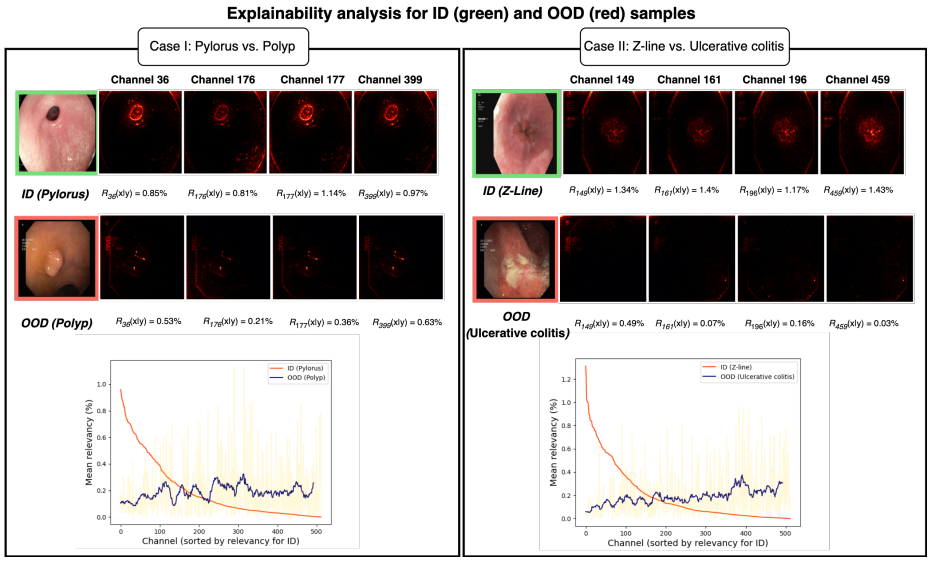

Fig. 3. Explainability analysis of ID and OOD samples for two cases. In each case, the top two rows show attribution maps for ID and OOD samples across the top four channels with the highest relevance scores (per ID class). The bottom plot depicts the distribution of mean relevance scores across all channels for both ID and OOD samples.

to the OOD data to highlight the general trend. These plots illustrate a clear distinction between ID and OOD relevance scores across channels.

Penultimate Layer Selection: The use of the penultimate layer for obtaining relevance scores is motivated by the presence of skip connections and attention layers in widely used architectures [12,10,32], which often disrupt the propa-

gation of relevance scores and violate the conservation property of the LRP algorithm [4]. While recent methods have adapted LRP for residual and attention mechanisms [27,2], operating on the penultimate layer, which consists of a fully connected network, ensures broader applicability across architectures.

4 Conclusion

In this paper, we introduced Neural Relevance-based OOD detection (NERO), a novel post-hoc OOD detection method that leverages neuron-level relevance patterns to distinguish OOD samples. By clustering neuron-level relevance signatures for each ID class and quantifying the relevance distance of test samples from these class-centric centroids, our approach provides a robust framework for OOD detection. Our evaluations on challenging gastrointestinal datasets demonstrated its effectiveness, achieving consistently superior performance across diverse model architectures, including CNN-based (ResNet-18) and transformer-based (DeiT) networks.

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

References

1. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence **5**(9), 1006–1019 (2023)
2. Achtibat, R., Hatefi, S.M.V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S., Samek, W.: Attnlrp: attention-aware layer-wise relevance propagation for transformers. arXiv preprint arXiv:2402.05602 (2024)
3. Ammar, M.B., Belkhir, N., Popescu, S., Manzanera, A., Franchi, G.: Neco: Neural collapse based out-of-distribution detection. arXiv preprint arXiv:2310.06823 (2023)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
5. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
6. Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., Stenzinger, A., Parlow, L., Budczies, J., Goeppert, B., et al.: Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. arXiv preprint arXiv:1805.11178 (2018)
7. Chan, H.P., Samala, R.K., Hadjiiski, L.M., Zhou, C.: Deep learning in medical image analysis. Deep learning in medical image analysis: challenges and applications pp. 3–21 (2020)
8. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: Proceedings of the ieee/cvf international conference on computer vision. pp. 5128–5137 (2021)

9. Cook, M., Zare, A., Gader, P.: Outlier detection through null space analysis of neural networks. arXiv preprint arXiv:2007.01263 (2020)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. pp. 1026–1034 (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. 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. arXiv preprint arXiv:1911.11132 (2019)
14. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
15. Hong, Z., Yue, Y., Chen, Y., Cong, L., Lin, H., Luo, Y., Wang, M.H., Wang, W., Xu, J., Yang, X., et al.: Out-of-distribution detection in medical image analysis: A survey. arXiv preprint arXiv:2404.18279 (2024)
16. Horst, F., Lapuschkin, S., Samek, W., Müller, K.R., Schöllhorn, W.I.: Explaining the unique nature of individual gait patterns with deep learning. Scientific reports **9**(1), 2391 (2019)
17. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. Advances in Neural Information Processing Systems **34**, 677–689 (2021)
18. Jha, D., Sharma, V., Dasu, N., Tomar, N.K., Hicks, S., Bhuyan, M.K., Das, P.K., Riegler, M.A., Halvorsen, P., Bagci, U., et al.: Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection. In: Workshop on Machine Learning for Multimodal Healthcare Data. pp. 125–140. Springer (2023)
19. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS **25** (2012)
21. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
22. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
23. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
24. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019)
25. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
26. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A.: The building blocks of interpretability. Distill **3**(3), e10 (2018)

27. Otsuki, S., Iida, T., Doublet, F., Hirakawa, T., Yamashita, T., Fujiyoshi, H., Sugiura, K.: Layer-wise relevance propagation with conservation property for resnet. In: European Conference on Computer Vision. pp. 349–364. Springer (2024)
28. Pokhrel, S., Bhandari, S., Ali, S., Lambrou, T., Nguyen, A., Shrestha, Y.R., Watson, A., Stoyanov, D., Gyawali, P., Bhattarai, B.: Ncdd: Nearest centroid distance deficit for out-of-distribution detection in gastrointestinal vision. arXiv preprint arXiv:2412.01590 (2024)
29. Sharma, A., Kumar, R., Garg, P.: Deep learning-based prediction model for diagnosing gastrointestinal diseases using endoscopy images. International Journal of Medical Informatics **177**, 105142 (2023)
30. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in Neural Information Processing Systems **34**, 144–157 (2021)
31. Tang, K., Miao, D., Peng, W., Wu, J., Shi, Y., Gu, Z., Tian, Z., Wang, W.: Codes: Chamfer out-of-distribution examples against overconfidence issue. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1153–1162 (2021)
32. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
33. Tschuchnig, M.E., Gadermayr, M.: Anomaly detection in medical imaging-a mini review. In: International Data Science Conference. pp. 33–38. Springer (2021)
34. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022)