JOURNAL OF
EMERGING INVESTIGATORS

Article

Mitigating open-set misclassification in a colorectal cancer detecting neural network

Olina Mukherjee[1], Benjamin Erichson[2]
[1] Allderdice High School, Pittsburgh, Pennsylvania
[2] International Computer Science Institute, Berkeley, California

SUMMARY

Neural networks typically operate with a closed-world assumption. This implies that they are only designed to classify objects that they were trained on, or in-distribution (ID) objects. However, when deployed in the real world, networks inevitably come across objects that they were not trained on, or out-of-distribution (OOD) objects. In these situations, the closed-world assumption leaves networks ill-prepared, resulting in the misclassification of OOD objects. Neural networks are especially prone to misclassifying open-set objects, a subset of OOD objects that belong to the ID dataset and are therefore very similar in appearance to ID objects. In safety-critical applications, like cancer diagnosis, open-set objects are very common, and their misclassifications can have dire consequences. In this study, we aimed to design a method to mitigate the misclassification of open-set objects. We hypothesized that exposing networks to examples of open-set objects during training via an auxiliary training class would substantially reduce the rate at which open-set objects were misclassified. Our Intra-Dataset Outlier Exposure (IDOE) method outperformed all but one state-of-the-art method in reducing open-set misclassification when evaluated on the MNIST and CIFAR-10 datasets. Additionally, IDOE achieved a test Area Under the Receiver Operating Curve score of 94.8% when tested on the PathMNIST dataset of healthy and cancerous colorectal tissues, demonstrating a high efficacy in medical diagnosis applications. Future work should focus on testing IDOE on more diverse, real-world datasets, such as those used in autonomous vehicles, to better evaluate its applicability in safety-critical environments, where the misclassification of unlabeled open-set objects is costly.

INTRODUCTION

Neural networks have primarily been evaluated based on their ability to classify objects that they were trained to detect. Due to their impressive results in these controlled evaluations, they are being deployed in real-world applications, including autonomous vehicles and medical diagnosis (1). They have seen great success in many of these applications (1). However, they exhibit sensitivities to various factors, including adversarial perturbations (small, deliberate modifications to input data designed to deceive machine learning models), corruptions like blurry or noisy images, and out-of-distribution (OOD) objects (2).

OOD objects are objects that models were not trained on. They pose a particularly challenging threat to neural networks. The term “distribution” refers to the statistical properties of a dataset, including how data points are spread and their overall patterns. Therefore, in-distribution (ID) objects, which networks are trained on, belong to the same distribution as a network’s training dataset, while OOD objects belong to distributions different from the training distribution (2). This leads most networks to accurately classify ID objects and misclassify OOD objects (2). OOD misclassifications can be detrimental in safety-critical applications like medical diagnosis and autonomous vehicle navigation, especially since these applications often involve OOD outliers like emerging diseases and new types of speed bumps (3).

The field of generalized OOD detection includes OOD detection itself as well as open-set recognition (OSR) (2, 4-5). For a cancer-detecting neural network, cancerous and non-cancerous tissues are ID objects, while handwritten digits, fruits, and cars are OOD objects. To prevent the misclassification of these OOD objects, the cancer-detecting network requires OOD detection (2). Additionally, within a given ID dataset, there are known closed-set objects that are seen during training and novel open-set objects that are not seen during training but still belong to the training distribution (5). For the same cancer-detecting network, cancerous and non-cancerous tissues are closed-set objects, while tissue debris—dead and decaying tissue that is neither benign nor malignant—is an open-set object. Filtering out these open-set objects to prevent their misclassification requires OSR (2). This subtler distinction between closed and open-set objects versus ID and OOD objects, makes OSR tasks generally more challenging than OOD detection tasks.

Most current OSR approaches compare encountered objects to the closed-set objects they were trained on. One such state-of-the-art approach, OpenMax, does so by fitting trained closed-set objects to a Weibull distribution and classifies encountered objects that lie significantly outside this distribution as open-set objects (6). Similarly, Sun et al. used K-nearest neighbors (KNN) to quantify differences between closed-set objects and encountered objects to classify objects that are significantly different from the closed-set as open-set (7). When applied to the MNIST dataset of handwritten digits, OpenMax and Sun et al.’s implementation of KNN achieved test Area Under the Receiver Operating Characteristic curve (AUROC) scores of 97.3% and 97.5%, respectively (6-9). When applied to the CIFAR-10 dataset of animals and vehicles, OpenMax and KNN achieved test AUROC scores of 84.2% an 86.9%, respectively (6-7, 10).

https://doi.org/10.59720/23-345

We proposed Intra-Dataset Outlier Exposure (IDOE), a method where we expose neural networks to example open-set objects during training via an auxiliary open-set class. We hypothesized that this neural network-based approach would better recognize and generalize the intrinsic characteristics of open-set objects, giving IDOE an advantage over current state-of-the-art OSR methods, which rely on quantifying differences between closed and open-set objects. This is particularly pertinent for datasets with subtle differences between the closed- and open-sets and which, therefore, demand a more nuanced understanding of the data, such as medical datasets. Our study supported these hypotheses, as IDOE outperformed all but one state-of-the-art open-set recognition method across multiple datasets. Additionally, IDOE's high test AUROC score of 94.8% when deployed on the medical PathMNIST dataset of colorectal tissues demonstrates a high efficacy in the real-world application of medical diagnosis where open-set outliers are very visually similar to closed-set trained objects (11).

RESULTS

Comparison of IDOE and state-of-the-art methods

To evaluate the efficacy of the proposed Intra-Dataset Outlier Exposure (IDOE) method, we compared its performance to that of 13 state-of-the-art OSR methods, all of which are post-hoc and therefore focus on learning the distribution of the closed-set. These state-of-the-art AUROC results were found in a survey of the generalized OOD field by Yang, et al. (12). To enable a direct comparison of IDOE to these state-of-the-art results, we used the dataset, network architecture, and metric benchmarks from this survey (12). We evaluated our IDOE method, on two image datasets, each with 10 classes: MNIST, consisting of white handwritten digits 0-9 on black backgrounds, and CIFAR-10, which contains images of vehicles and animals (8, 10, 12). To form the closed- and open-sets, we randomly selected six closed- and four open-set classes each time we trained a network, as was done by Yang, et al. (12).

To create the novel auxiliary open-set class for the IDOE networks, we amalgamated two of the four open-set classes. Thus, the IDOE train set was comprised of images of all six closed-set objects and two open-set objects, while the test set comprised of different images of all six closed-set objects as well as images of the two open-set objects that the network was not trained on (**Figure 1**). We trained five IDOE networks per dataset and evaluated them on the MNIST and CIFAR-10 test sets accordingly. We chose to use the AUROC evaluation metric to compare IDOE's test performance to the test performances of the 13 state-of-the-art methods found in the survey by Yang, et al. (9,12). Accounting for true and false positive rates, this metric holistically describes models' abilities to distinguish between open and closed-sets. Higher AUROC scores are better and random classifiers have an AUROC of 50% (**Figure 2**). For each dataset we reported the average test AUROC scores of the five IDOE networks to maintain consistency with Yang, et al., who reported the average test AUROC score of five runs for each state-of-the-art method (9, 12). We also reported the standard deviation, sensitivity and specificity for these five IDOE networks to gain further insight into the results, though these statistics were not provided by Yang, et al. for the state-of-the-art methods.

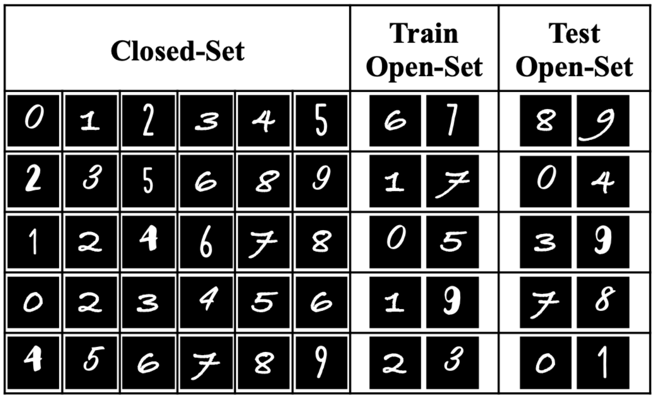

Figure 1: Example MNIST class configurations to train and test IDOE. The train set consists of closed-set and train open-set images while the test set consists of closed-set and test open-set images. Images of numbers were created using computer fonts and are not real MNIST images.

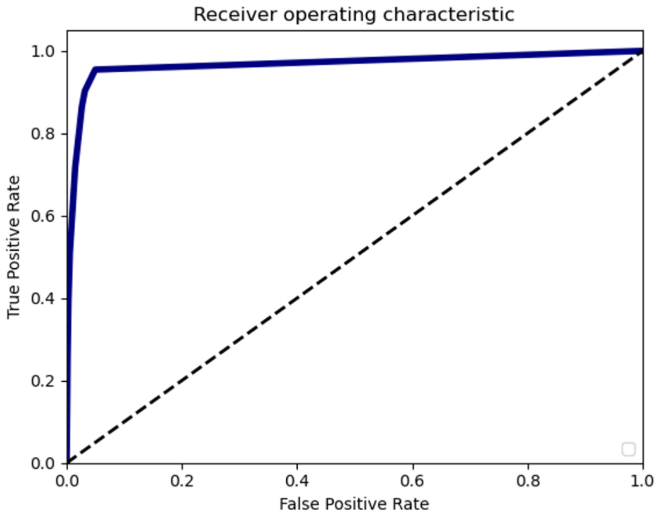

Figure 2: Area Under the Receiver Operating Characteristic (AUROC) curve evaluation metric. The dashed diagonal line represents a random classifier (50%), and the dark blue line is the curve of an IDOE network trained on MNIST that achieved a 97%.

IDOE outperformed all of the state-of-the-art methods on the MNIST dataset, with an average test AUROC score of 98.5%. On CIFAR-10, the KNN implementation by Sun et al. recorded the top average test AUROC score at 86.9%, while IDOE scored 0.6% lower (**Table 1**) (7). Almost all models performed worse on CIFAR-10 compared to MNIST, likely because CIFAR-10 contains a diverse range of animal and vehicle images, whereas MNIST consists of uniform, white handwritten digits on black backgrounds. It is harder for networks to generalize from limited examples to accurately classify diverse test objects. Additionally, IDOE had standard deviations of 1.57% on MNIST and 3.46% on CIFAR-10 (**Table 2**). It achieved a sensitivity of 97.8% with a specificity of 99.3% on MNIST and a sensitivity of 94.5% with a specificity of 75.9% on CIFAR-10 (**Table 2**). The higher standard deviation and lower specificity on CIFAR-10 suggest that IDOE was less consistent on this dataset than on MNIST. Since CIFAR-10 is more diverse, there is also opportunity for greater variation between the five networks since the closed- and open sets were randomly selected for each network.

Method Implemented for OSR	Source	Mean AUROC Score	
		MNIST	*CIFAR-10*
State-of-the-Art Methods:			
OpenMax	Bendale and Boult (6)	97.3%	84.2%
Maximum Softmax Probability	Hendrycks and Gimpel (2)	96.2%	85.3%
Out-of-distribution Image Detection	Liang, et al. (17)	98.0%	72.1%
Mahalanobis Distance-based Score	Lee, et al. (18)	89.8%	42.9%
Gram	Sastry and Oore (19)	82.3%	61.0%
Energy Based Out-of-distribution detection	Liu, et al. (20)	98.1%	84.9%
Gradient Norms	Huang, et al. (21)	94.5%	64.8%
Rectified Activations	Sun, et al. (22)	82.9%	85.9%
Maximum Logits Scores	Hendrycks, et al. (23)	98.0%	84.8%
Kullback–Leibler-divergence Matching	Hendrycks, et al. (23)	85.4%	73.7%
Virtual-logit Matching	Wang, et al. (24)	88.8%	83.5%
K-nearest neighbors	Sun, et al. (7)	97.5%	**86.9%**
Directed Sparisification	Sun and Li (25)	66.3%	79.3%
Proposed Method:			
Intra-Dataset Outlier Exposure (IDOE)		**98.5%**	86.3%

Table 1: Performance of implementations of state-of-the-art and proposed IDOE open-set recognition methods on MNIST and CIFAR-10. Reported AUROC scores are averages of five networks (n=5).

Standard deviation, sensitivity, and specificity values were not reported for the state-of-the-art methods by Yang, et al. (12).

Evaluating the efficacy of IDOE on PathMNIST

Since the proposed IDOE method performed similarly or better than the 13 state-of-the-art methods when evaluated on the standard MNIST and CIFAR-10 datasets, we decided to test its efficacy in a medical diagnostic application. We deployed IDOE on the PathMNIST dataset of histological slides of healthy and cancerous colorectal tissues (11). PathMNIST consists of two cancerous tissue classes, five healthy tissue classes, and two inconsequential classes (background and tissue debris) (11). We amalgamated the background and tissue debris classes to form the auxiliary open-set class during training since they are of little value when it comes to diagnosing cancerous tissue. Five randomly selected classes served as closed-set objects and the remaining two classes were used as novel open-set objects for testing (**Figure 3**). We trained five neural networks with the same auxiliary open-set class but different random selections of the five closed-set classes. Then we tested the networks on both unseen closed-set images and images of the novel open-set objects. Lastly, we calculated the average test AUROC score, standard deviation, sensitivity, and specificity of the five IDOE networks.

IDOE achieved a mean test AUROC score of 94.8% with a standard deviation of 0.63% (**Table 2**). This standard deviation is lower than the standard deviations achieved on both MNIST and CIFAR-10, likely due to the fact that the auxiliary open-set class did not vary for the five different IDOE networks deployed on PathMNIST. IDOE also recorded sensitivity and specificity scores of 94.7% and 91.6%, respectively, on PathMNIST, surpassing its performance on CIFAR-10 but not surpassing its performance on MNIST (**Table 2**). This is likely due to the fact that MNIST is the least diverse dataset, only consisting of white digits on black backgrounds. PathMNIST is slightly more diverse, consisting of only images of histological slides which vary more than grayscale digits since different tissues have different color tones and detailed cell structures. Lastly, CIFAR-10 is the most diverse since it contains different objects ranging from birds to trucks.

DISCUSSION

IDOE outperformed all state-of-the-art methods when evaluated on the MNIST dataset (8). When we evaluated IDOE on the CIFAR-10 dataset, it outperformed all but one of the state-of-the-art methods, with only the KNN model implemented by Sun, et al. performing better (7). Additionally, for CIFAR-10, IDOE's test AUROC score of 86.3% was only marginally lower than the 86.9% achieved by Sun, et al. (7, 10). Additionally, the standard deviations were 1.57% and 3.46% on the MNIST and CIFAR-10 datasets, respectively, indicating minimal differences in performance between runs and therefore minimal dependance on the selection of open and closed-sets.

In comparison to state-of-the-art methods, which rely on quantifying the distribution of the closed-set to identify outliers as open-set, IDOE's auxiliary open-set class enabled the model to learn the common attributes of the in-distribution. As a result, when IDOE encountered an object that shared the dataset's common attributes but did not resemble a specific closed-set object closely enough, it usually classified that object into the open-set class. For example, suppose IDOE

Method	AUROC Data		Sensitivity	Specificity
	μ	σ		
MNIST	98.5%	1.57%	97.8%	99.3%
CIFAR-10	86.3%	3.46%	94.5%	75.9%
PathMNIST	94.8%	0.63%	94.7%	91.6%

Table 2: Average performance of 5 proposed IDOE networks on MNIST, CIFAR-10, and PathMNIST (n=5).

https://doi.org/10.59720/23-345

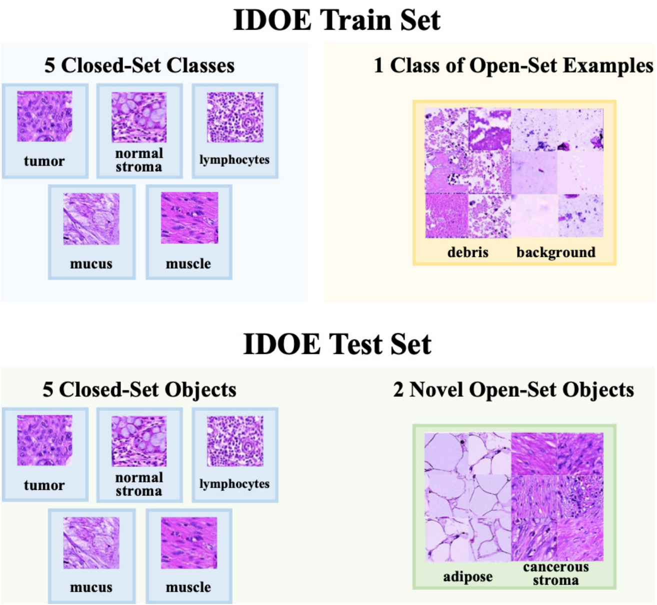

Figure 3: Example configuration of IDOE train and test sets on PathMNIST. IDOE trains networks using debris and background in its auxiliary open-set class in addition to five closed-set tissues. IDOE networks are tested on images of the two novel open-set tissues in addition to different images of the closed-set tissues. Images are from Yang, Jiancheng, et al (12).

was trained with digits 0, 1, 2, 3, 4, and 5 from MNIST as its closed-set objects and digits 6 and 7 in its auxiliary open-set class. When it encounters a novel open-set MNIST digit, such as 8 or 9, these images share the same black background and white text as the closed and open-set objects. However, since the 8 and 9 do not closely resemble any of the closed-set digits, they are likely to be correctly placed in the more general open-set class.

However, we found a large decline in specificity from 99.3% on MNIST to 75.9% on CIFAR-10. CIFAR-10 is inherently more diverse than MNIST, comprising images of vehicles and animals with varying backgrounds and color schemes. When test closed-set images very different from those in the train set were encountered, IDOE networks likely classified them in the more general open-set class, resulting in a higher false positive rate indicative of the decline in specificity. As a result of this limitation, this current IDOE framework should be implemented in applications where images are relatively consistent in terms of background, like MNIST. For example, in medical applications, all encountered images will be of tissues on histological slides and therefore they will have very similar backgrounds. In the future, experiments can be conducted to determine whether increasing the size and diversity of the auxiliary open-set class in such diverse datasets improves IDOE's ability to distinguish between open- and closed-set objects.

When we deployed IDOE on the medical PathMNIST dataset of colorectal tissues, we found that it was consistently accurate, with a mean test AUROC score of 94.8% and standard deviation of 0.63%. IDOE also had high sensitivity and specificity scores of 94.7% and 91.6%, respectively. These scores are higher than IDOE's scores on CIFAR-10, which means that IDOE had a lower false positive rate on PathMNIST. This may be due to the fact that all images in the PathMNIST dataset are tissues, which are less diverse in appearance than the animal and vehicle images in CIFAR-10. As a result, the networks likely misclassified less closed-set images into the open-set class. Since PathMNIST is a medical dataset, these results also suggest that IDOE may reduce misdiagnosis of open-set tissues in similar medical applications where open-set and closed-set objects are very similar and more difficult to classify. With the findings from our study, physicians can see that open-set objects pose a great challenge for automated diagnostic assistants. By leveraging IDOE as a diagnostic assistant that filters out open-set objects, physicians can focus on diagnosing known closed-set conditions with greater confidence, reducing the risk of misdiagnosis. Additionally, by analyzing the detected open-set objects, physicians could identify previously undiscovered conditions, potentially aiding in the early detection and diagnosis of emerging medical issues. Implementing IDOE in clinical practice could thus help minimize misdiagnosis of open-set tissues and enhance patient care, without requiring physicians to check that their diagnostic neural network assistants are only diagnosing tissues that they were trained on.

MATERIALS AND METHODS

IDOE Framework

We implemented a k-class neural network that seeks to learn a nonlinear function *h: X* → *Y* where *X* is a set of input images and *Y* = *{1, …, k, o}* is a set of class labels where *{1, …, k}* represents closed-set classes and *o* represents an auxiliary open-set class. We trained the network on a dataset consisting of m examples $(x_1, y_1), \ldots, (x_m, y_m)$, where each x_i is an element of the set *X* is an input image, and y_i is an element of the set *Y* is the corresponding output label. During training, the network learned the parameters θ by minimizing

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i),\ y_i) \qquad \textbf{(Eqn 1)}.$$

the empirical risk with the loss function *l*:

After training, the network had learned *h: X* → *Y'* and was able to classify test images, *xtest*, as belonging to one of the closed-set classes *{1, …, k}* if *h(xtest)* is an element of the set *{1, …, k}* or to the open-set class, *o*, otherwise.

IDOE Evaluation

In evaluating IDOE, we followed the benchmarks described by Yang, et al., as these were the benchmarks that they used to evaluate the 13 state-of-the-art OSR methods cited above (12). We deployed IDOE on two standard 10-class datasets: MNIST, containing of 70,000 28 by 28 pixel images of white handwritten digits 0-9 on black backgrounds, and CIFAR-10, containing 60,000 32 by 32 pixel images of airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks (8, 10). As suggested by Yang, et al., we randomly selected six classes to be closed-set classes and four classes to be open-set classes each time a network was trained (12). We amalgamated two of the open-set to form the IDOE auxiliary class of open-set examples during training and the number of images in this class was limited to the average number of images in the closed-set classes to prevent overfitting. IDOE networks were trained with all six closed-set classes as well as the auxiliary open-set class. We then evaluated them on a test set consisting of different

https://doi.org/10.59720/23-345

images of the same closed-set objects as well as images of the two open-set objects not used during training. The training-set of MNIST consists of six closed-set classes and one open-set class each with 6,007 images (8). The test-set of MNIST consists of 1,010 images of each closed-set object and 1,004 images of each of the two test open-set objects (8). For CIFAR-10, the training-set had six closed-set classes and one open-set class each with 5,000 images (10). The CIFAR-10 test-set had 1,000 images of each closed- and test open-set object (10). Lastly, the PathMNIST training-set had five closed-set classes and one open-set class each with 5,241 images (11). The PathMNIST test-set had 5,241 images of each closed-set object and 5,243 images of each of the two test open-set objects (11).

Besides the benchmarked MINST and CIFAR-10 datasets, we also deployed IDOE on the PathMNIST dataset of 28 by 28 pixel images of colorectal tissues on histological slides (8, 10-11). PathMNIST has nine classes and 107,180 total images (11). We amalgamated the inconsequential background (empty slides) and tissue debris classes to create the auxiliary open-set class for training each IDOE network. Then we randomly selected five of the remaining objects to be used as closed-set objects and we used the last two as open-set objects during testing. We trained IDOE networks with all five closed-set classes as well as the auxiliary open-set class. Then we tested the networks on different images of the closed-set objects as well as images of the novel open-set objects not included in the train set.

Following the benchmarks from Yang, et al., we used the LeNet-5 and ResNet18 models as base architectures for the IDOE networks classifying the MNIST and CIFAR-10 datasets, respectively (12-14). We also used LeNet-5 as the base architecture when we deployed IDOE on PathMNIST since the PathMNIST images are the same size as MNIST images, while CIFAR-10 images are bigger. To prevent overfitting, we trained each network using the Stochastic Gradient Descent optimizer, which updates model parameters iteratively, with a learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005 for 100 epochs, as per Yang, et al. (12). Lastly, we calculated the AUROC, sensitivity, and specificity of each network. We coded, trained, and calculated the scores of these networks on Google Colaboratory using Python, Keras, and TensorFlow (15-16).

ACKNOWLEDGMENTS

The primary author would like to thank her high school physics teacher for supporting her interest in research and giving her opportunities to participate in science fairs throughout high school.

Received: December 27, 2023
Accepted: July 5, 2024
Published: July 14, 2025

REFERENCES

1. Samek, Wojciech, et al. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications." *Proc. IEEE*, vol. 109, no. 3, pp. 247-278, 4 Mar. 2021, https://doi.org/10.1109/JPROC.2021.3060483.
2. Hendrycks, Dan, and Kevin Gimpel, "A Baseline for Detecting Misclassified and Out-of-distribution Examples in Neural Networks." *ICLR,* 23 Mar. 2017, https://doi.org/10.48550/arXiv.1610.02136.
3. Zadorozhny, Karina, et al. "Out-of-distribution detection for medical applications: Guidelines for practical evaluation." *Multimodal AI in Healthcare*, 29 Nov. 2022, pp. 137-153, https://doi.org/10.1007/978-3-031-14771-5_10.
4. Yang, Jingkang, et al., "Generalized Out-of-distribution Detection: A Survey." *arXiv*, 21 Oct. 2021, https://doi.org/10.48550/arXiv.2110.11334. Accessed 24 Dec. 2023. Preprint.
5. Scheirer, Walter J., et al., "Toward Open Set Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757-1772, 29 Nov. 2012, https://doi.org/10.1109/TPAMI.2012.256.
6. Bendale, Abhijit, and Terrance E. Boult, "Towards Open Set Deep Networks." *CVPR*, Las Vegas, NV, USA, 12 Dec. 2016, pp. 1563-1572, https://doi.org/10.1109/CVPR.2016.173.
7. Sun, Yiyou, et al., "Out-of-Distribution Detection with Deep Nearest Neighbors." *ICML*, 13 Apr. 2022, https://doi.org/10.48550/arXiv.2204.06507.
8. LeCun, Yann, et al., "The MNIST Database of Handwritten Digits." *YannLeCun*, 2010, https://yann.lecun.com/exdb/mnist/.
9. Bradley, Andrew P., "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1 July 1997, https://doi.org/10.1016/S0031-3203(96)00142-2.
10. Krizhevsky, Alex, "Learning Multiple Layers of Features from Tiny Images." *Tech Report*, 8 Apr. 2009, www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.
11. Yang, Jiancheng, et al., "MedMNIST v2 - A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification." *Sci Data*, vol. 10, no. 41, 19 Jan. 2023, https://doi.org/10.1038/s41597-022-01721-8.
12. Yang, Jingkang, et al., "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection." *arXiv*, 13 Oct. 2022, https://doi.org/10.48550/arXiv.2210.07242. Accessed 24 Dec. 2023. Preprint.
13. LeCun, Yann, et al., "Gradient-Based Learning Applied to Document Recognition." *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 6 Aug. 2002, https://doi.org/10.1109/5.726791.
14. He, Kaiming, et al., "Deep Residual Learning for Image Recognition." *CVPR*, Las Vegas, NV, USA, 12 Dec. 2016, pp. 770-778, https://doi.org/10.1109/CVPR.2016.90.
15. "Welcome to Colab." *Google Colab*, Google, 2017, colab.research.google.com/.
16. Martín Abadi, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems." 14 Mar. 2016. Software available from tensorflow.org.
17. Liang, Shiyu, et al., "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks." *ICLR*, 30 Apr. 2018, https://doi.org/10.48550/arXiv.1706.02690.
18. Lee, Kimin, et al., "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks." *NeurIPS*, 27 Oct. 2018, https://doi.org/10.48550/arXiv.1807.03888.
19. Sastry, Chandramouli S., and Sageev Oore, "Detecting Out-of-Distribution Examples with Gram Matrices." *ICML*, 9 Jan. 2020, https://doi.org/10.48550/arXiv.1912.12510.
20. Liu, Weitang, et al., "Energy-Based Out-of-Distribution Detection." *NeurIPS*, 8 Oct. 2020, https://doi.

org/10.48550/arXiv.2010.03759.

21. Huang, Rui, et al., “On the Importance of Gradients for Detecting Distributional Shifts in the Wild.” *NeurIPS*, 9 Oct. 2021, https://doi.org/10.48550/arXiv.2110.00218.
22. Sun, Yiyou, et al., “ReAct: Out-of-Distribution Detection with Rectified Activations.” *NeurIPS*, 24 Nov. 2021, https://doi.org/10.48550/arXiv.2111.12797.
23. Hendrycks, Dan, et al., “Scaling Out-of-Distribution Detection for Real-World Settings.” *ICML*, 15 May 2022, https://doi.org/10.48550/arXiv.1911.11132.
24. Wang, Haoqi, et al., “VIM: Out-of-Distribution with Virtual-Logit Matching.” *CVPR*, 21 Mar. 2022, https://doi.org/10.48550/arXiv.2203.10807.
25. Sun, Yiyou and Yixuan Li, “DICE: Leveraging Sparsification for Out-of-Distribution Detection,” *ECCV*, 17 July 2022, https://doi.org/10.48550/arXiv.2111.09805.