MIDL
Nashville 2023

Paper ID: 69

Know Your Space: Inlier and Outlier Construction for Calibrating Medical OOD Detectors






Vivek Narayanaswamy
LLNL

Yamen Mubarka
LLNL

Rushil Anirudh
LLNL

Deepta Rajan
Microsoft

Andreas Spanias
ASU

Jay Thiagarajan
LLNL

Lawrence Livermore National Laboratory

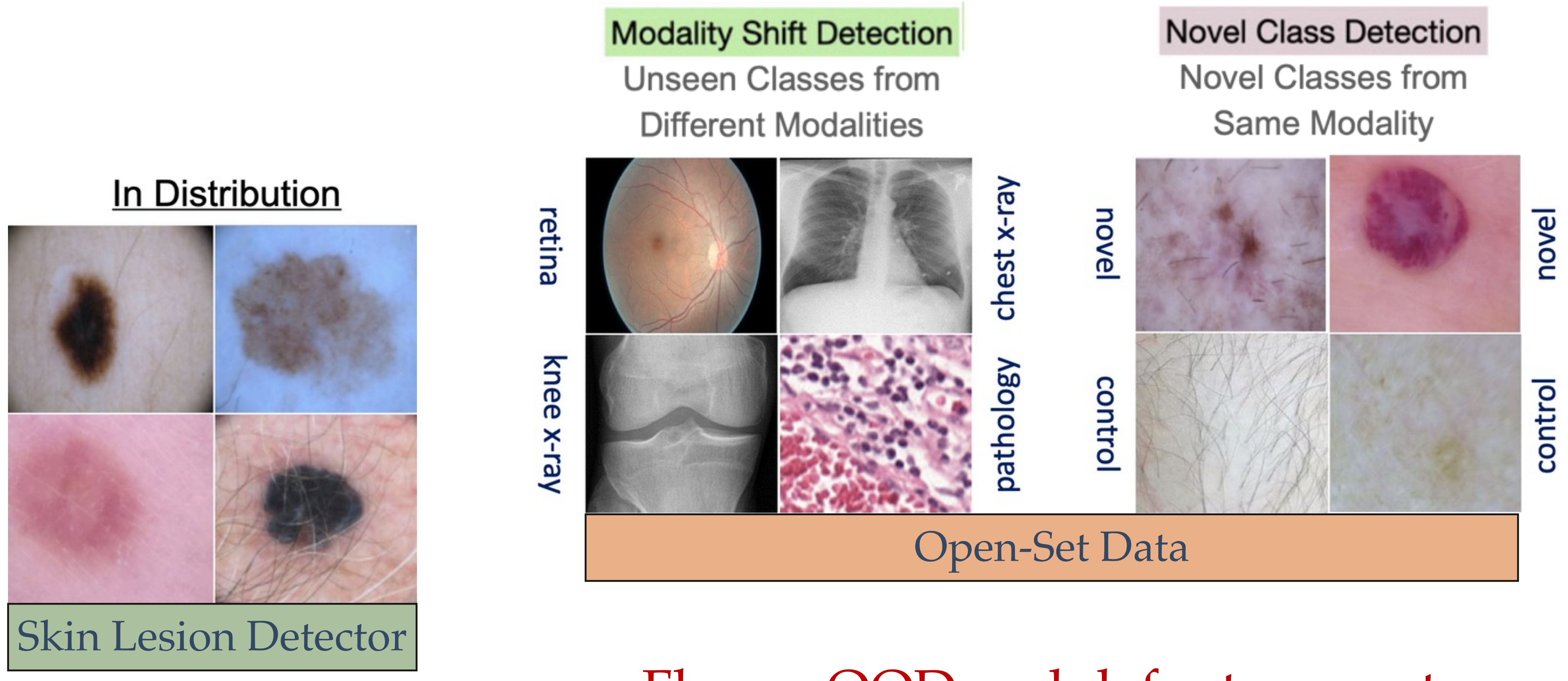
ASU

Safe Deployment of AI Models Requires us to Monitor Predictions and Detect Unexpected Model Functionality

Ability to flag open-set samples with diverse semantic characteristics w.r.t the training data is a critical aspect of safety in medical AI

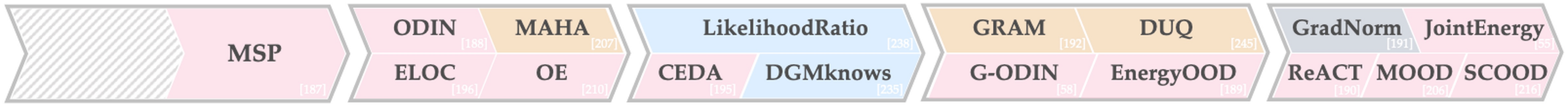
$$\mathcal{Y}_{\text{OOD}} \neq \mathcal{Y}_{\text{ID}}$$

Unseen semantic concepts



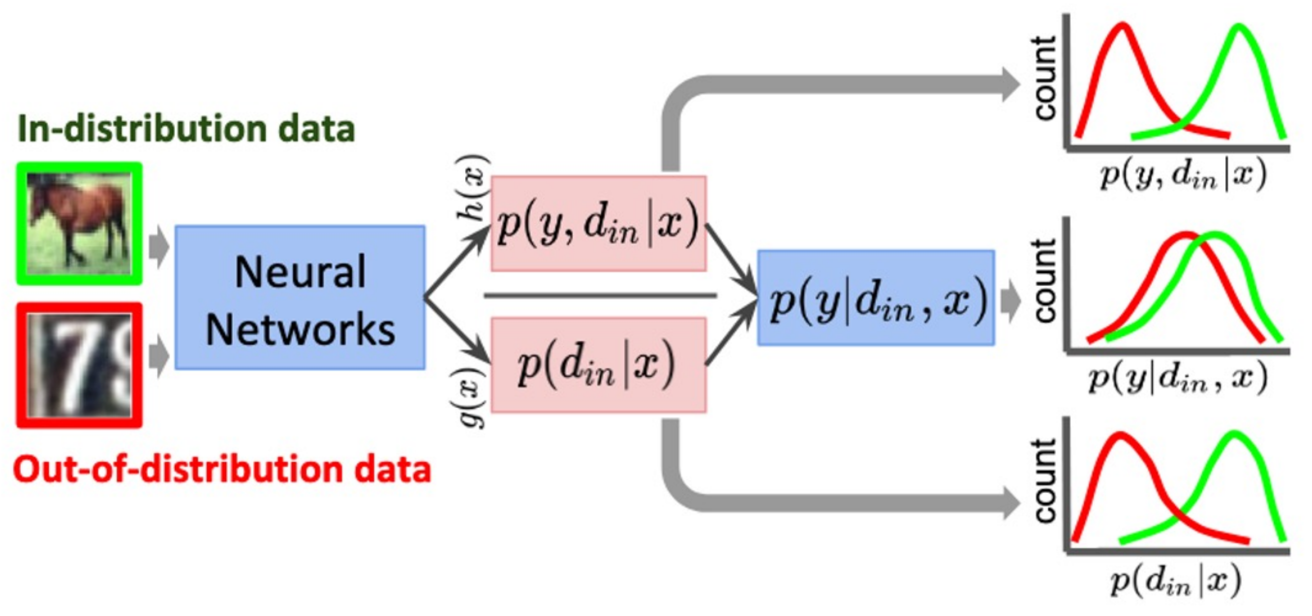
Flag as OOD and defer to experts

Constructing model scores for open-set detection is an active area of research in the vision community



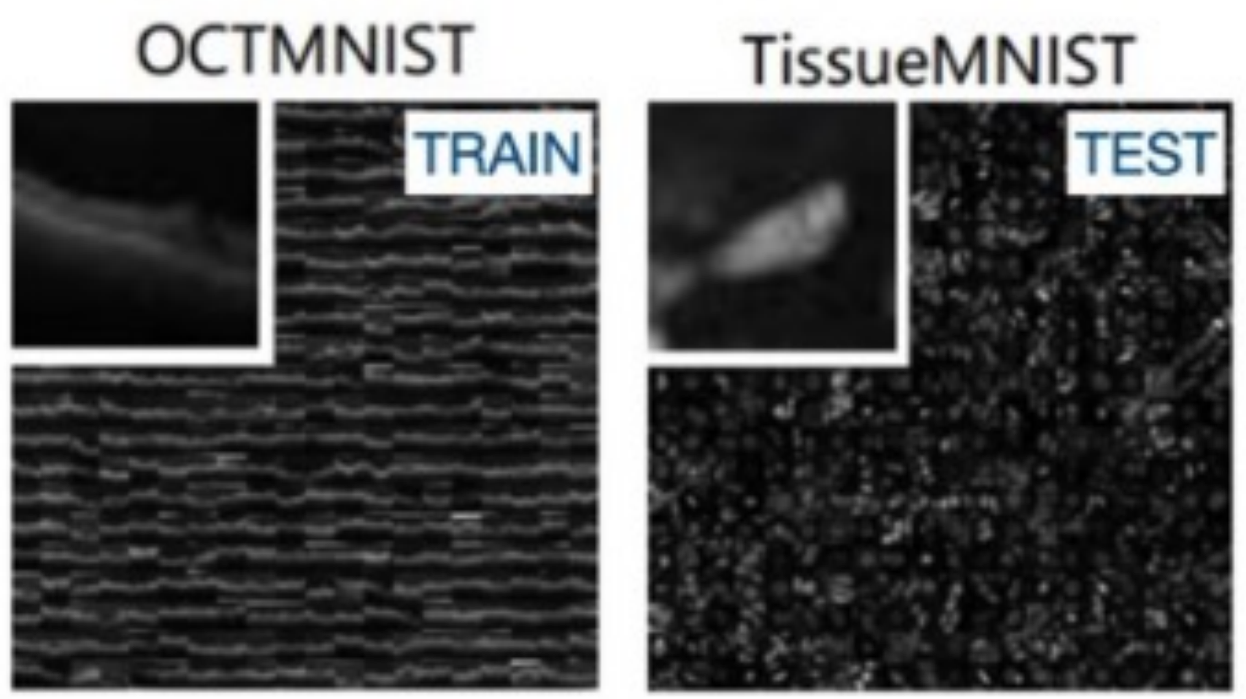
Key question: Can Off-the-shelf OOD Detectors Flag Open-Set Medical Image Data?

Generalized-ODIN



Despite achieving high accuracies on test data, conventional OOD detectors struggle with open data settings in medical imaging!

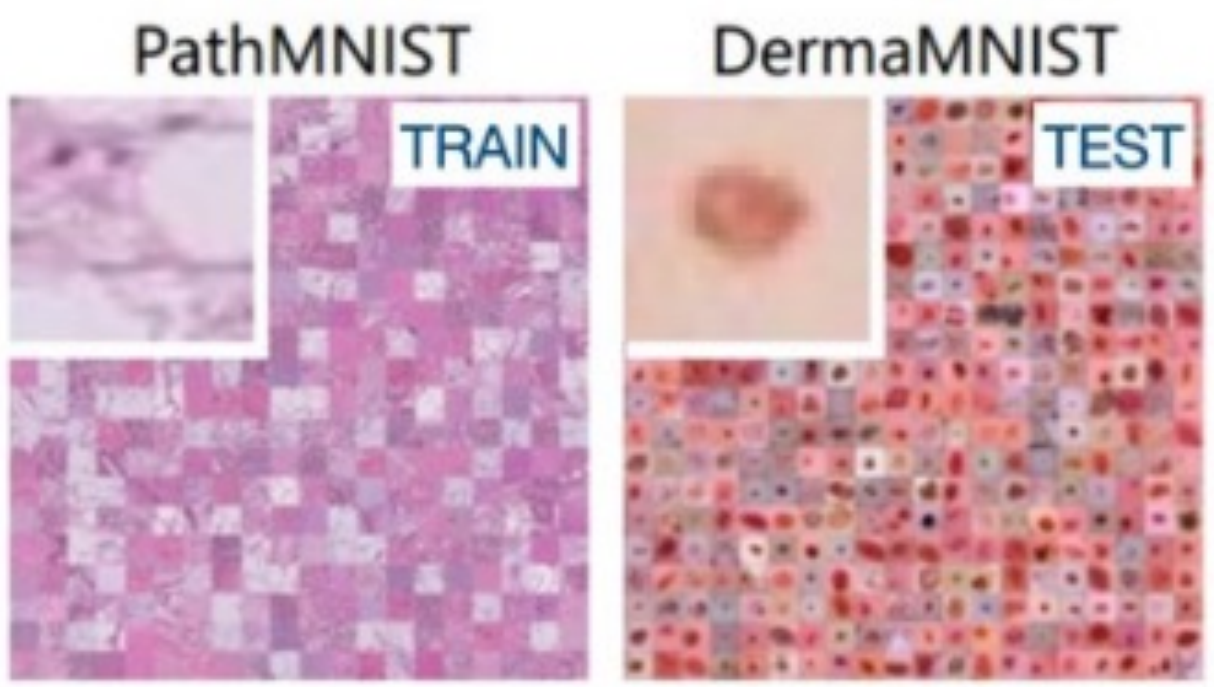
Modality Shift Detection



Balanced Accuracy: 95.9 ✓

OOD Rejection AUROC: 49.7 ✗

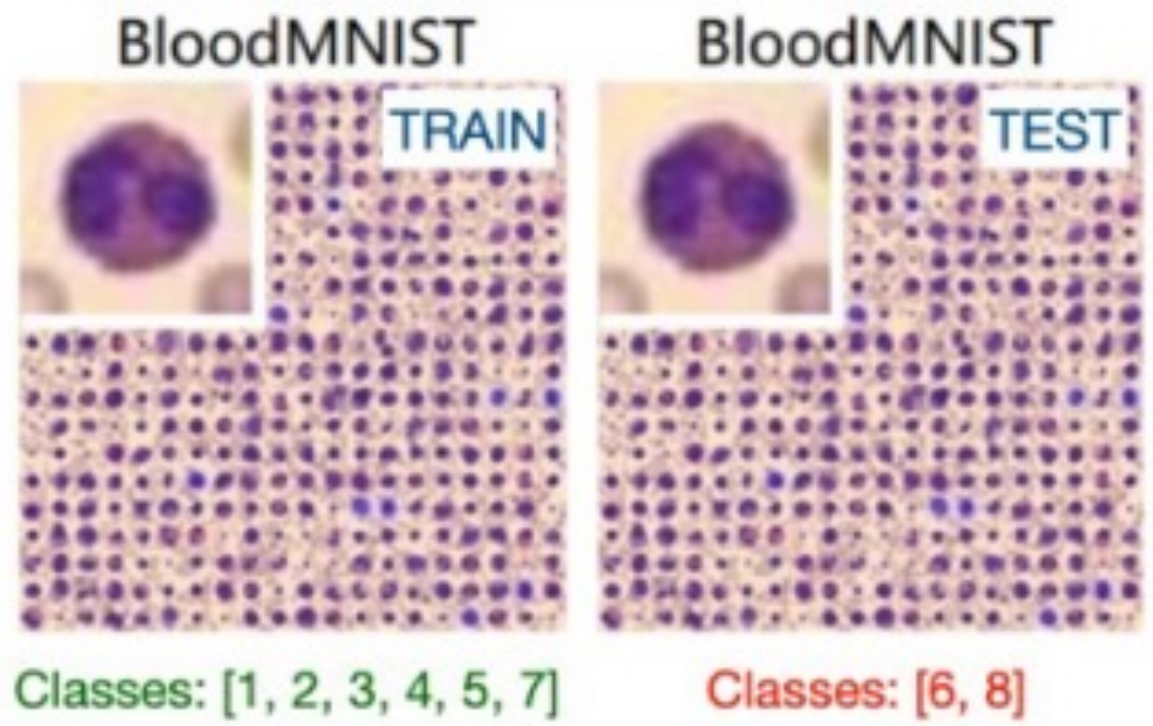
Modality Shift Detection



Balanced Accuracy: 99.3 ✓

OOD Rejection AUROC: 75.4 ✗

Novel Class Detection



Balanced Accuracy: 96.2 ✓

OOD Rejection AUROC: 53.9 ✗

A Potential Fix: Explicitly Calibrate OOD Detectors during Predictive Model Training

In this work, we consider the popular energy-based OOD detectors

$$G(\mathbf{x};\theta,\tau)=\begin{cases}\text{outlier}, & \text{if } -E(\mathbf{x};\theta)\le\tau\\ \text{inlier}, & \text{if } -E(\mathbf{x};\theta)>\tau\end{cases}$$

$$\min_{\theta}\ \mathop{\mathbb{E}}_{(\mathbf{x},y)\in\mathcal{D}}\mathcal{L}_{CE}(\mathcal{F}_\theta(\mathbf{x}),y)+\alpha\underbrace{\mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\in\mathcal{D}_{\text{in}}}}_{\text{Inlier Specification}}\mathcal{L}_{\text{ID}}(\overbrace{E(\tilde{\mathbf{x}})}^{\text{Energy}};\theta)+\beta\underbrace{\mathop{\mathbb{E}}_{\bar{\mathbf{x}}\in\mathcal{D}_{\text{out}}}}_{\text{Outlier Specification}}\mathcal{L}_{\text{OOD}}(E(\bar{\mathbf{x}});\theta)$$

Gibbs Distribution

$$p(y\mid\mathbf{x})=\frac{e^{-E(\mathbf{x},y)/T}}{\int_{y'}e^{-E(\mathbf{x},y')/T}}=\frac{e^{-E(\mathbf{x},y)/T}}{e^{-E(\mathbf{x})/T}}$$

Softmax Distribution

$$p(y\mid\mathbf{x})=\frac{e^{f_y(\mathbf{x})/T}}{\sum_i^K e^{f_i(\mathbf{x})/T}}$$

Energy

$$E(\mathbf{x};f)=-\underbrace{T}_{\text{Temperature}}\cdot\log\sum_i^K e^{\overbrace{f_i(\mathbf{x})}^{\text{Logit of Class i}}/T}.$$

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 2020

A Potential Fix: Explicitly Calibrate OOD Detectors during Predictive Model Training

In this work, we consider the popular energy-based OOD detectors

$$G(\mathbf{x};\theta,\tau) = \begin{cases} \text{outlier}, & \text{if } \; -E(\mathbf{x};\theta) \leq \tau \\ \text{inlier}, & \text{if } \; -E(\mathbf{x};\theta) > \tau \end{cases}$$

Energy

$$\min_{\theta} \; \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} \; \mathcal{L}_{CE}(\mathcal{F}_\theta(\mathbf{x}), y) + \alpha \underbrace{\mathbb{E}_{\tilde{\mathbf{x}}\in\mathcal{D}_{\text{in}}}}_{\text{Inlier Specification}} \mathcal{L}_{\text{ID}}(E(\tilde{\mathbf{x}});\theta) + \beta \underbrace{\mathbb{E}_{\bar{\mathbf{x}}\in\mathcal{D}_{\text{out}}}}_{\text{Outlier Specification}} \mathcal{L}_{\text{OOD}}(E(\bar{\mathbf{x}});\theta)$$

OOD detector calibration is implemented using margin-based loss functions

$$\mathcal{L}_{\text{ID}} = \mathbb{E}_{\mathbf{t}_k\sim\mathcal{T}} \left[\max\left(0, E(h(\mathbf{x})=\mathbf{t}_k) - m_{\text{ID}}\right)\right]^2$$

$$\mathcal{L}_{\text{OOD}} = \mathbb{E}_{\bar{\mathbf{x}}\sim\mathcal{D}_{\text{out}}} \left[\max\left(0, m_{\text{OOD}} - E(\mathbf{x}=\bar{\mathbf{x}})\right)\right]^2.$$

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 2020

A Potential Fix: Explicitly Calibrate OOD Detectors during Predictive Model Training

In this work, we consider the popular energy-based OOD detectors

$$G(\mathbf{x};\theta,\tau) = \begin{cases} \text{outlier}, & \text{if } \; -E(\mathbf{x};\theta) \leq \tau \\ \text{inlier}, & \text{if } \; -E(\mathbf{x};\theta) > \tau \end{cases}$$

$$\min_{\theta} \; \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}} \mathcal{L}_{CE}(\mathcal{F}_\theta(\mathbf{x}),y) + \alpha \underbrace{\mathbb{E}_{\tilde{\mathbf{x}}\in\mathcal{D}_{\text{in}}}}_{\text{Inlier Specification}} \mathcal{L}_{\text{ID}}(\overbrace{E(\tilde{\mathbf{x}})}^{\text{Energy}};\theta) + \beta \underbrace{\mathbb{E}_{\bar{\mathbf{x}}\in\mathcal{D}_{\text{out}}}}_{\text{Outlier Specification}} \mathcal{L}_{\text{OOD}}(E(\bar{\mathbf{x}});\theta)$$

Held-out calibration set from the train data can be used to specify inliers – challenging in small data scenarios

A representative OOD dataset is curated for specifying the outlier regimes – non-trivial in medical imaging

OOD calibration must not compromise the accuracy of the trained detector – avoid over-conservative models

How is this optimization carried out in practice?

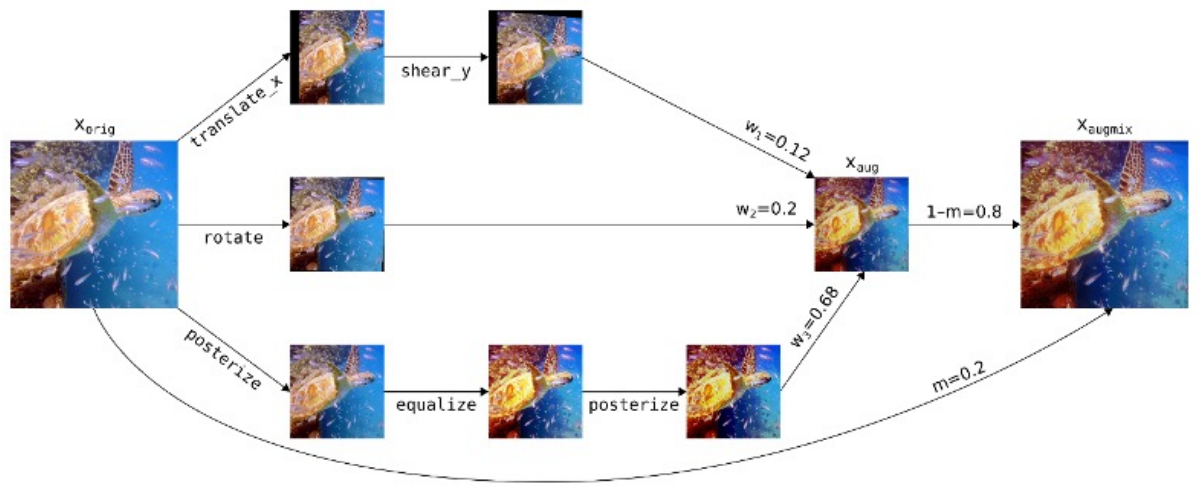
Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 2020

An Alternative Approach: Using Synthetic Data Augmentations to Specify Inliers and Outliers

<u>Pixel-Space</u> Augmentations

Inlier Specification

Geometric Transforms

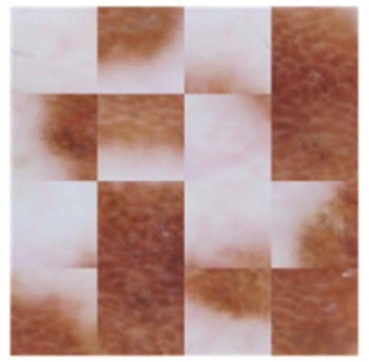


Compositional (e.g., AugMix))

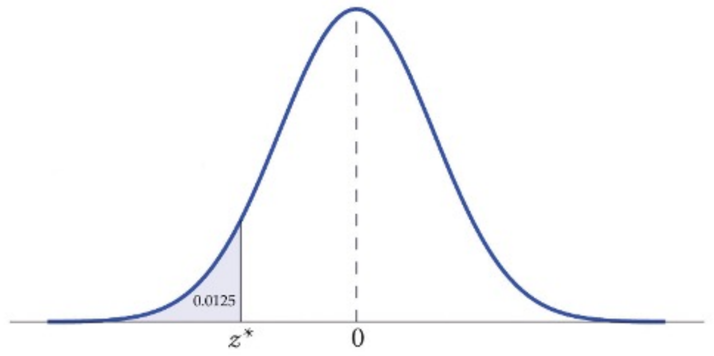
Outlier Specification

~~**Outlier Exposure with Representative Datasets**~~

Hard to Define for Medical Imaging Models



<u>Pixel-Space</u> Augmentation
(Negative Data Augmentation)



<u>Latent-Space</u> Augmentation
(Virtual Outlier Synthesis)

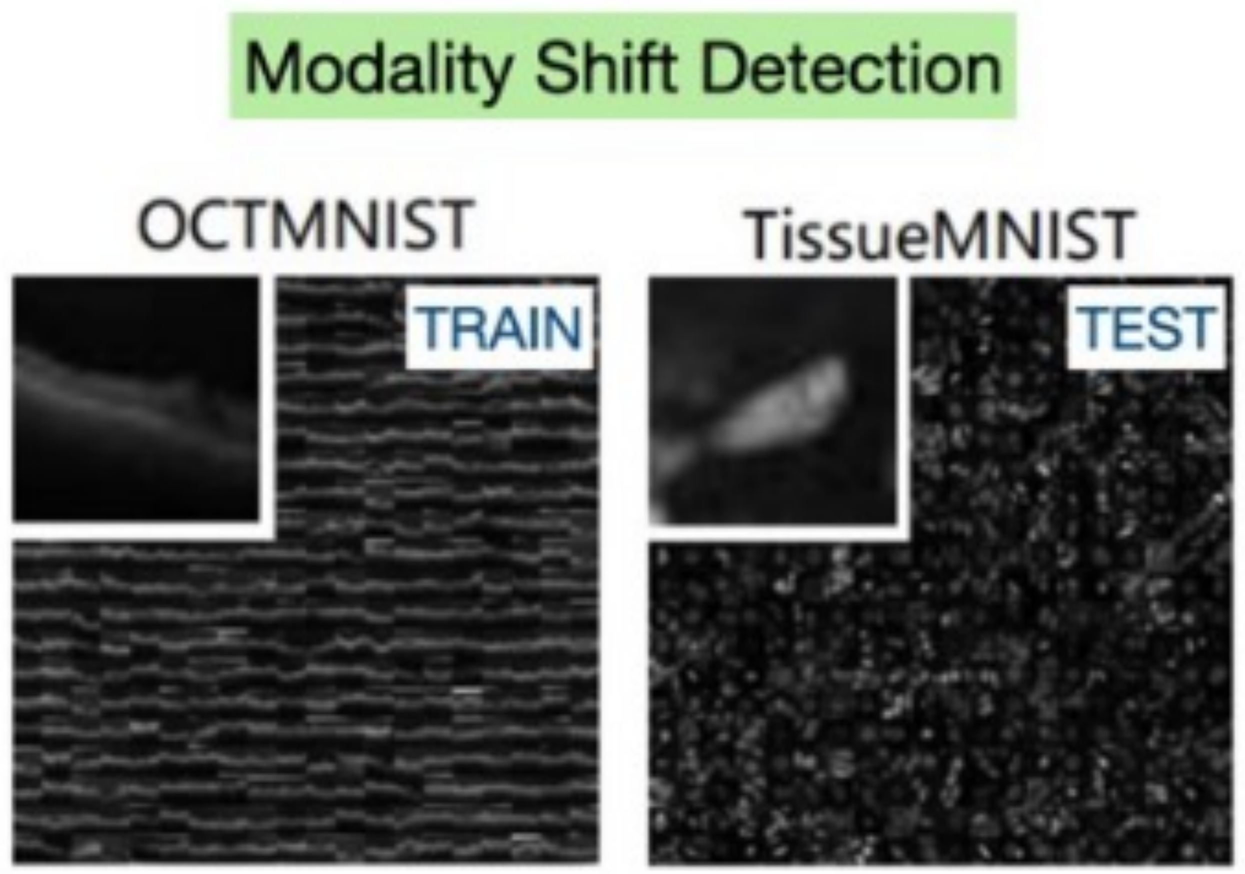
Hendrycks, Dan, et al. "Augmix: A simple data processing method to improve robustness and uncertainty." *arXiv preprint arXiv:1912.02781* (2019).

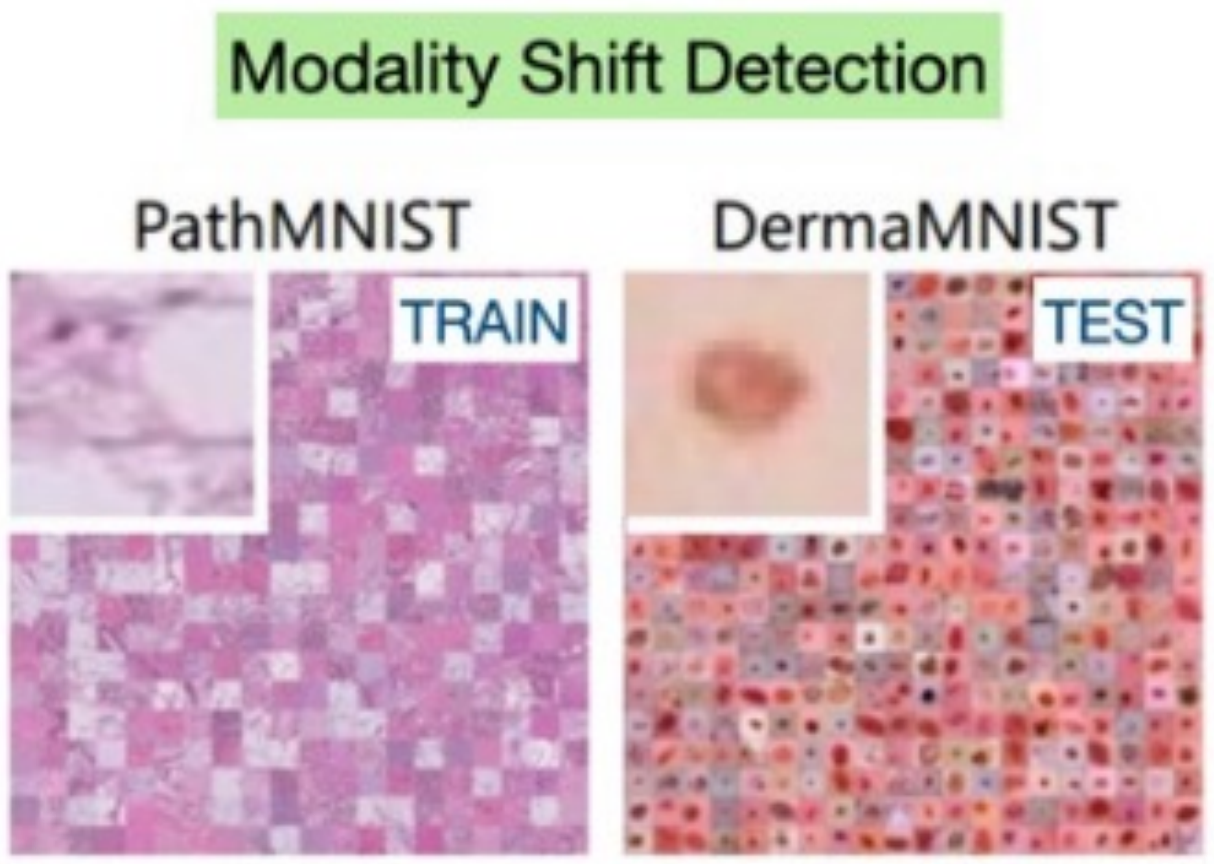
Hendrycks, Dan, Mantas Mazeika, and Thomas Dietterich. "Deep anomaly detection with outlier exposure." arXiv preprint arXiv:1812.04606 (2018).

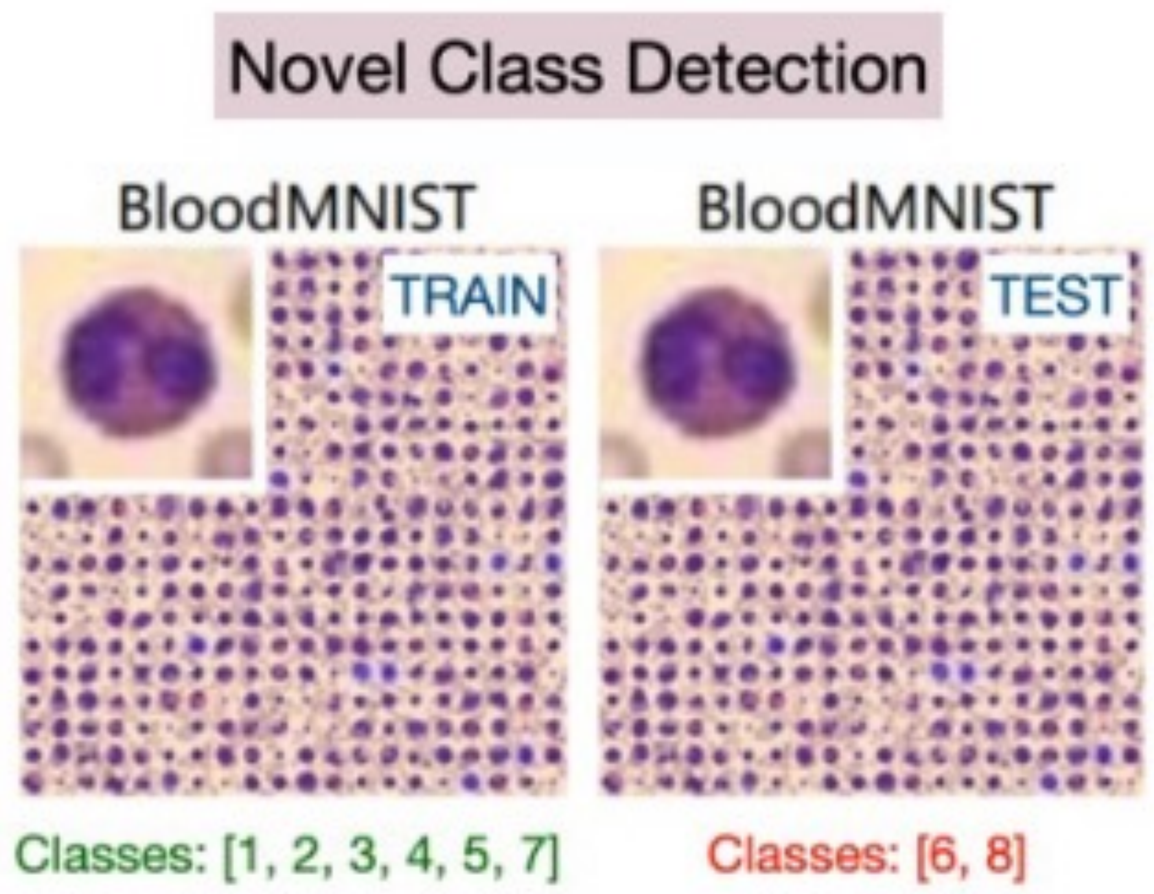
Sinha, A et al. Negative data augmentation. In International Conference on Learning Representations, 2021

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. arXiv preprint arXiv:2202.01197, 2022

How Effective are the Calibrated OOD Detectors?

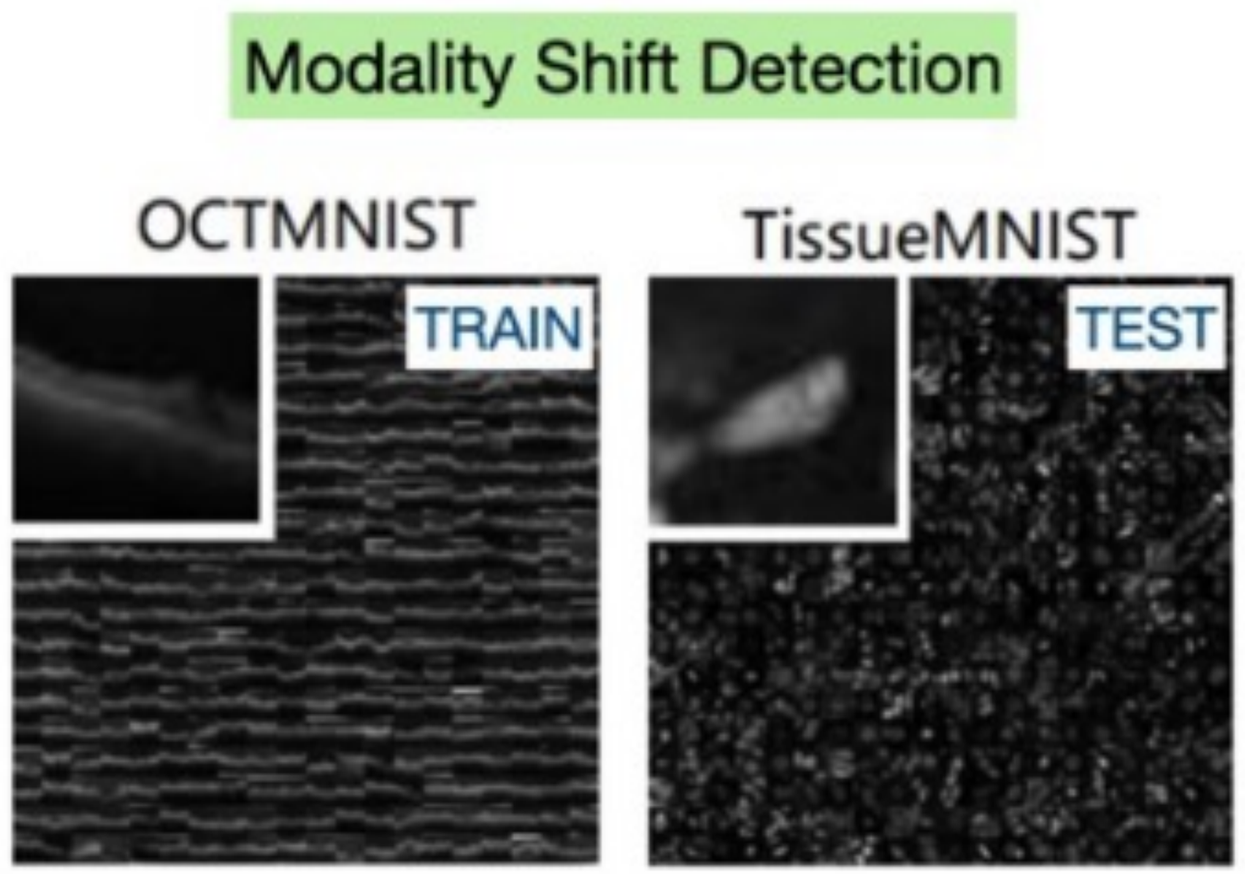


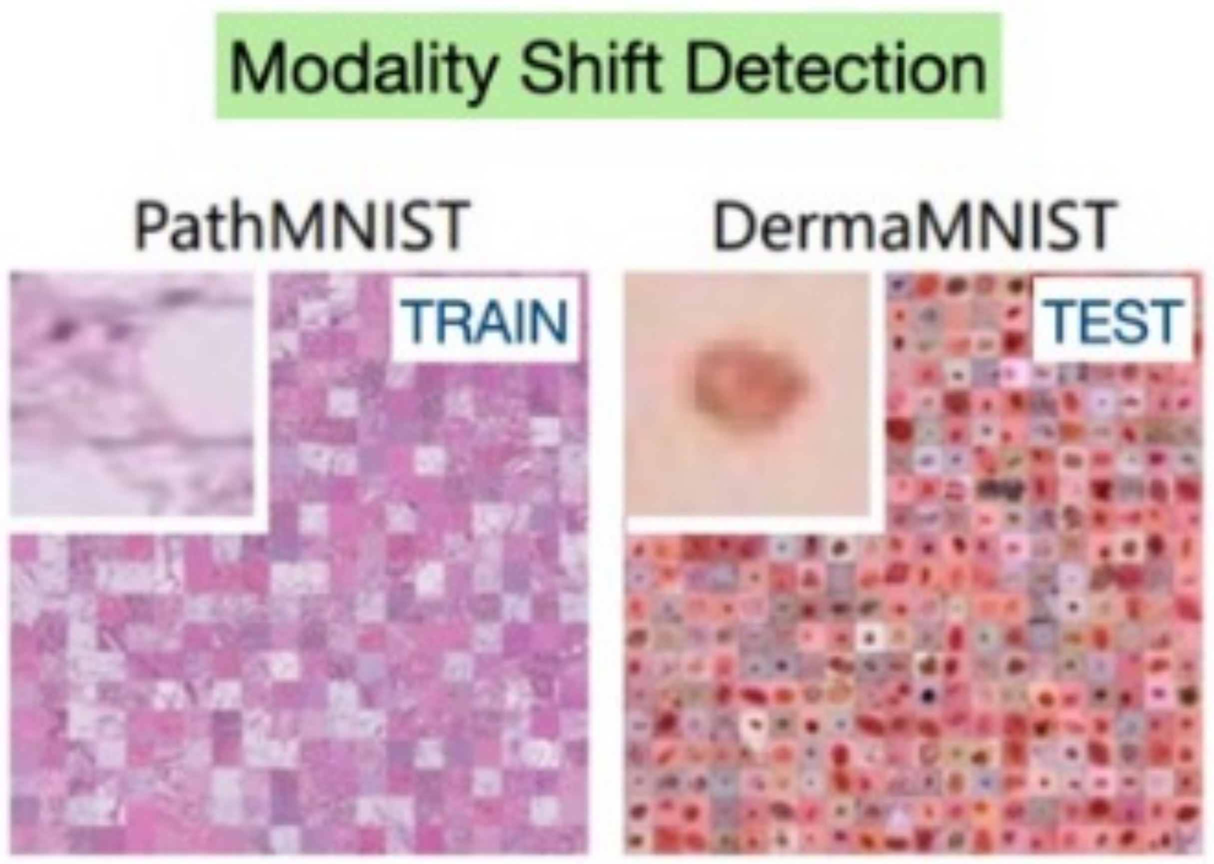


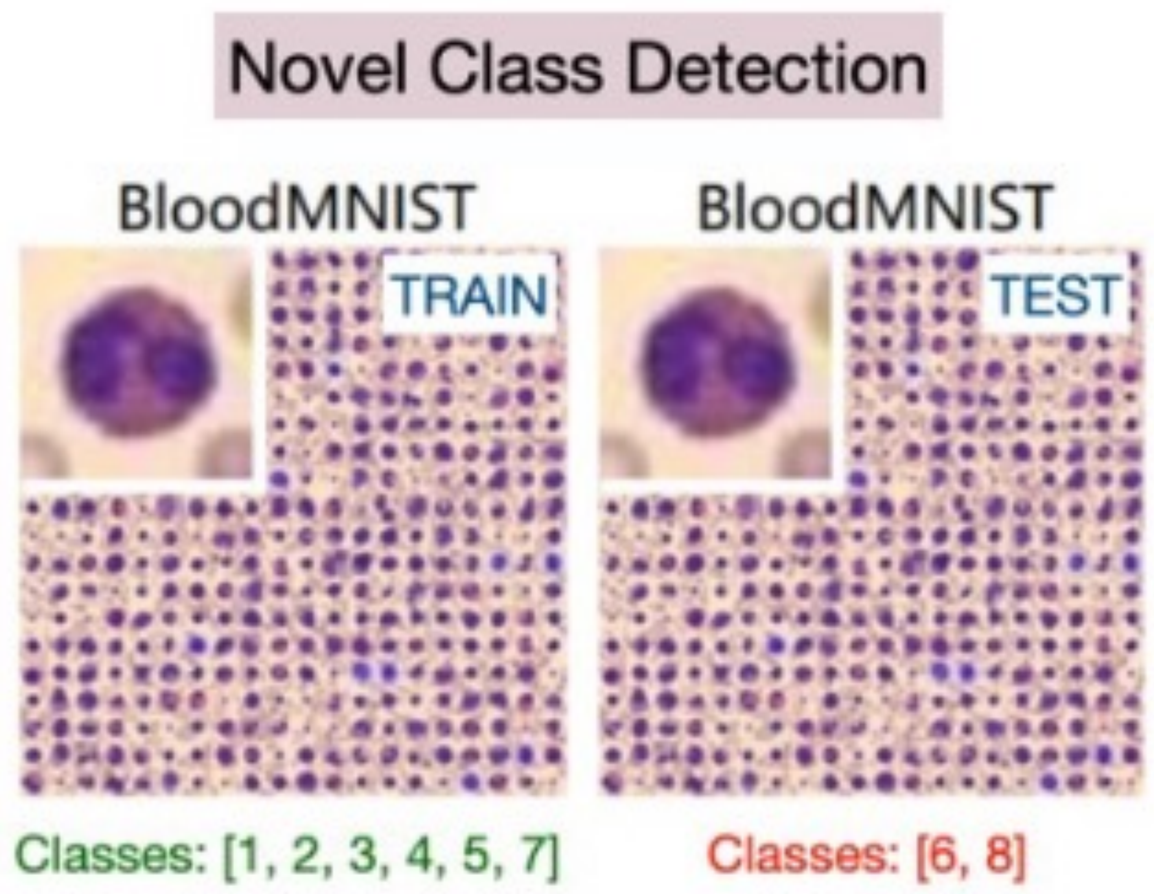


	OCTMNIST → TissueMNIST	PathMNIST → DermaMNIST	BloodMNIST → BloodMNIST
G-ODIN	Balanced Accuracy: 95.9 OOD Rejection AUROC: 49.7 ✗	Balanced Accuracy: 99.3 OOD Rejection AUROC: 75.4 ✗	Balanced Accuracy: 96.2 OOD Rejection AUROC: 53.9 ✗
Augmix + VOS	Balanced Accuracy: 95.3 OOD Rejection AUROC: 62.0 ✗	Balanced Accuracy: 99.2 OOD Rejection AUROC: 39.8 ✗	Balanced Accuracy: 96.5 OOD Rejection AUROC: 38.2 ✗
Augmix + NDA	Balanced Accuracy: 95.5 OOD Rejection AUROC: 90.9 ✓	Balanced Accuracy: 99.2 OOD Rejection AUROC: 43.5 ✗	Balanced Accuracy: 96.9 OOD Rejection AUROC: 53.5 ✗

How Effective are the Calibrated OOD Detectors?







G-ODIN

Modality Shift (OCTMNIST → TissueMNIST)	Modality Shift (PathMNIST → DermaMNIST)	Novel Class (BloodMNIST)
Balanced Accuracy: 95.9	Balanced Accuracy: 99.3	Balanced Accuracy: 96.2
OOD Rejection AUROC: 49.7 ✗	OOD Rejection AUROC: 75.4 ✗	OOD Rejection AUROC: 53.9 ✗

Surprisingly, OOD detectors calibrated using state-of-the-art approaches from vision literature do not perform consistently on both modality shifts and novel class scenarios

Hypothesis: Space in which inliers and Outliers are Specified Plays a Critical Role in Calibrating Medical OOD Detectors

With no outlier exposure, feature updates in a deep network are concentrated in the subspaces pertinent to the ID data

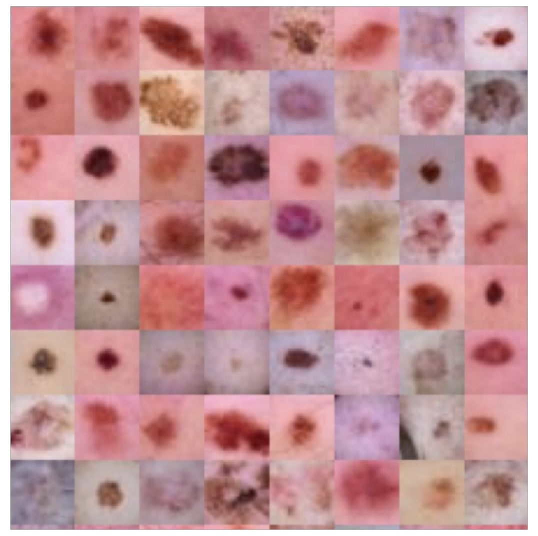
Inlier specification is used to expand model generalization and identify the optimal subspace for ID data – <u>Protect against shortcut learning</u>

Outlier specification is required to ensure that the subspaces for outlier data do not overlap with the ID subspaces – <u>Avoid over-generalization</u>

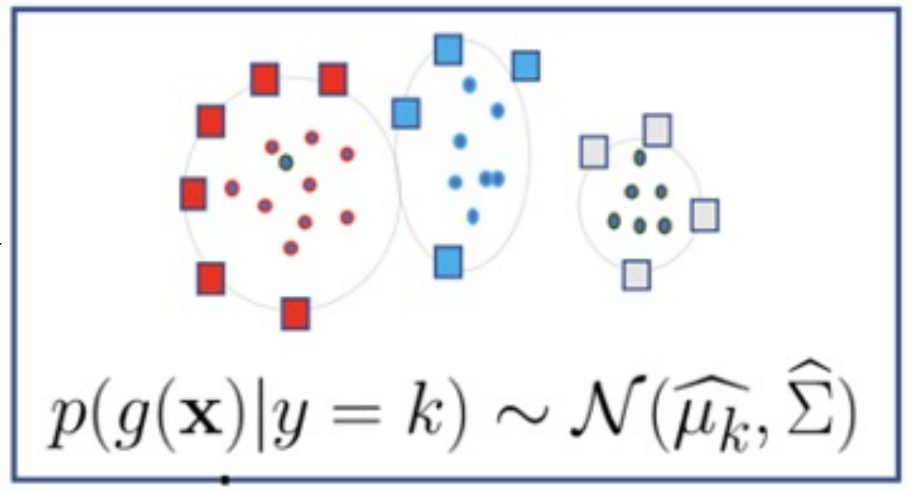
Contrary to existing works, we advocate for the use of the <u>latent-space for inlier specification</u> and <u>pixel-space for outlier specification</u> to perform OOD calibration, while not compromising on the test accuracy

Proposed Approach for Calibrating Energy-based OOD Detectors in Medical Imaging Models

Training Data



Sample low-likelihood regions from class-specific feature distributions



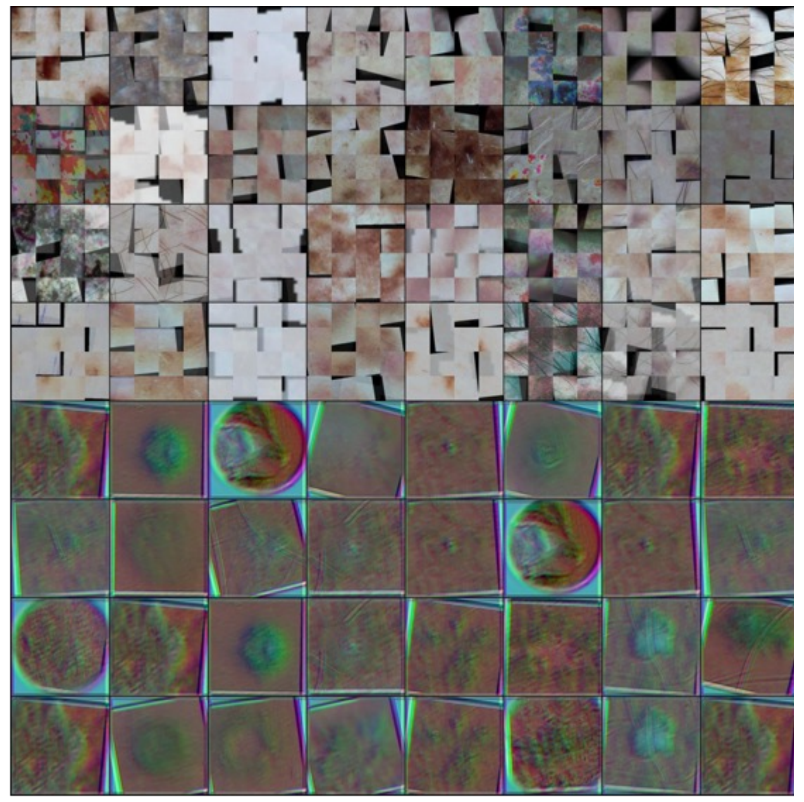
Lower bound on energy for the virtual inliers

$$E(h(\mathbf{x}) = \mathbf{t}_k)$$

$$-\log \sum_{k=1}^{K} \exp\left(- E(h(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_k, k) + \frac{1}{2}(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{t}_k + \widehat{\boldsymbol{\mu}}_k) \right)$$

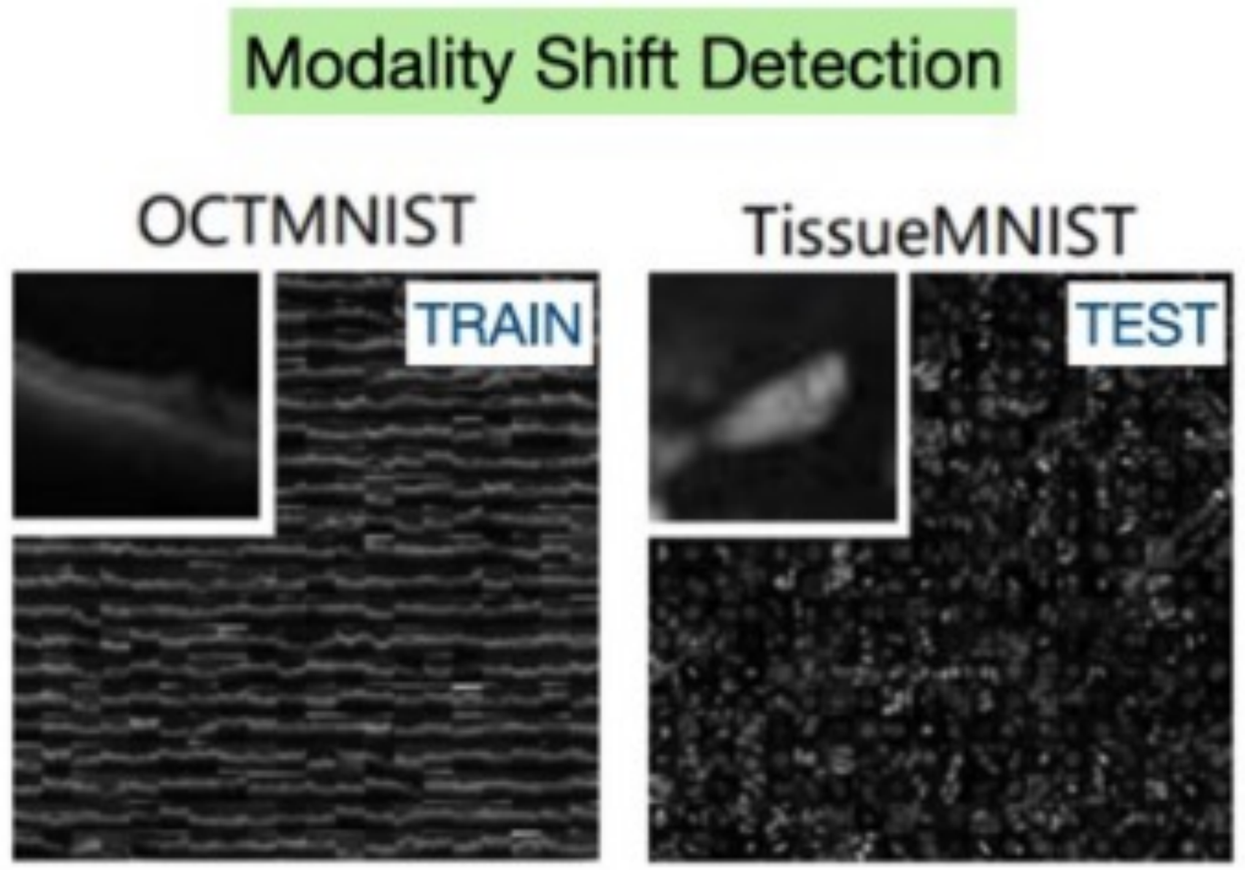
push the tail samples closer to the class-specific prototypes

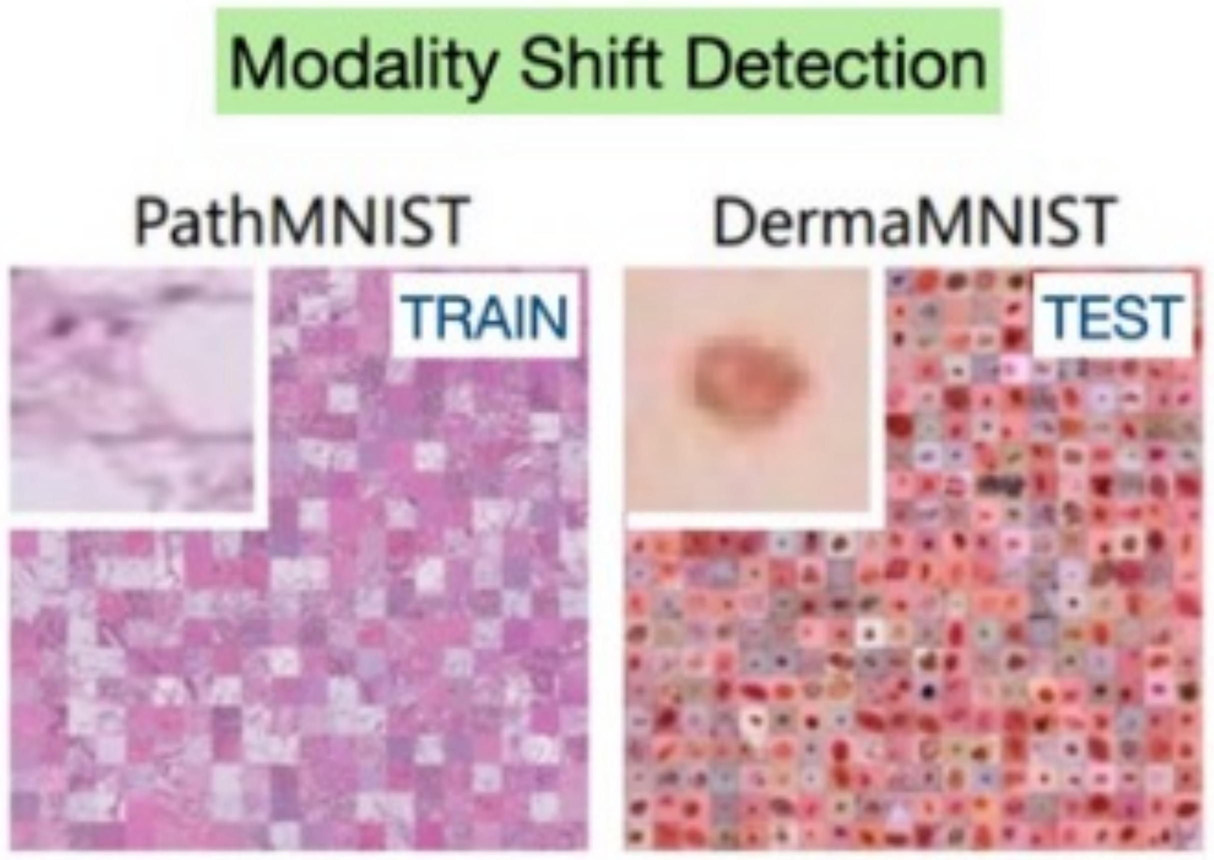
Synthesize high-severity compositional image manipulations (e.g., Augmix, RandConv)

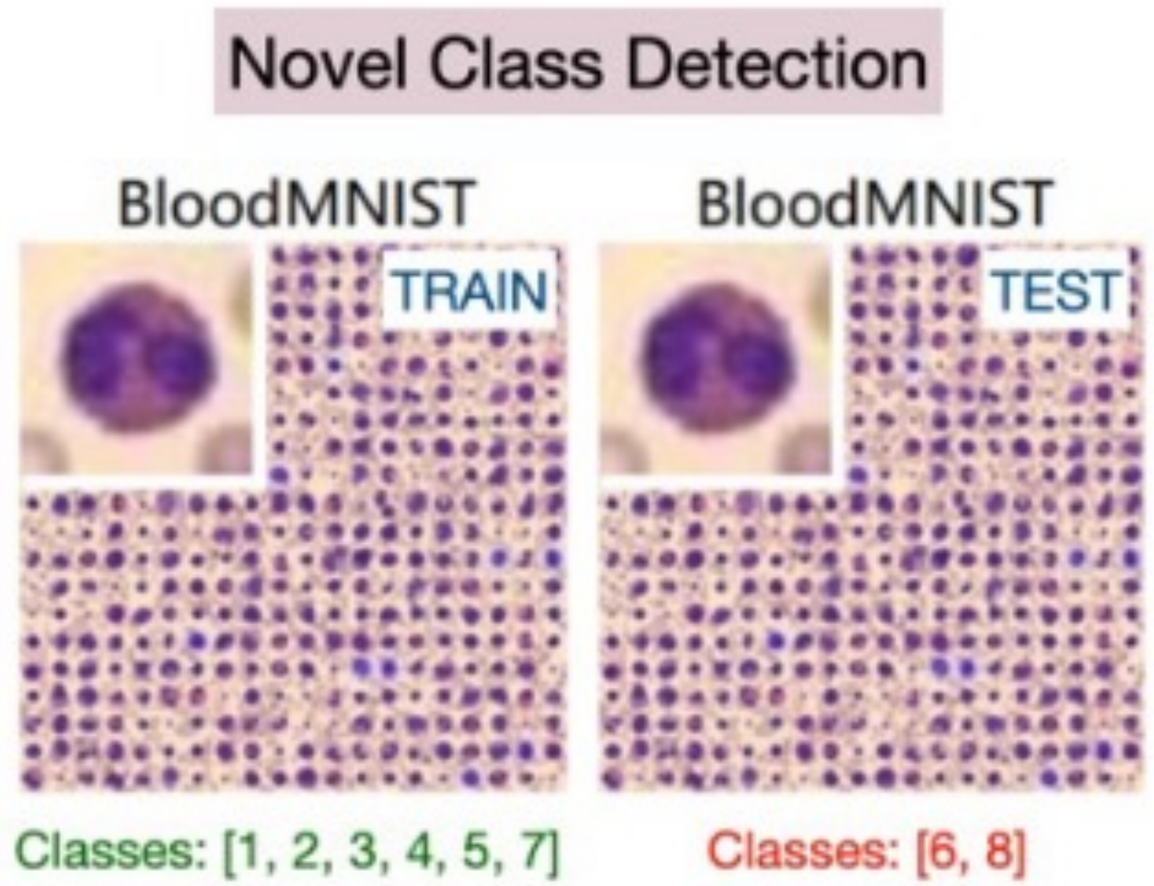


Highly diverse set of pixel-space outliers to ensure the OOD subspace does not overlap with the ID subspace in the feature space of the deep network

Using the Combination of Latent-Space Inliers and Synthetic Pixel-Space Outliers Leads to Powerful OOD Detectors

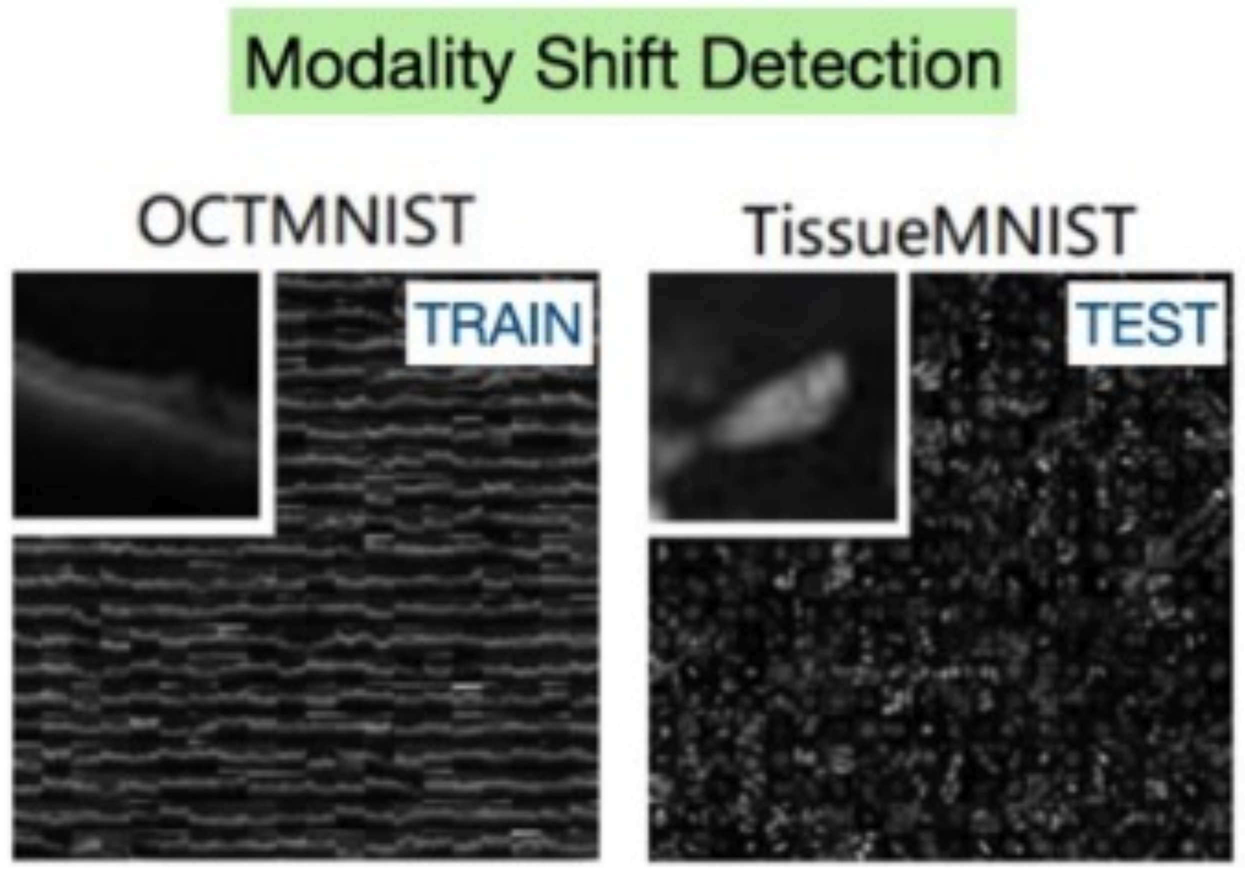


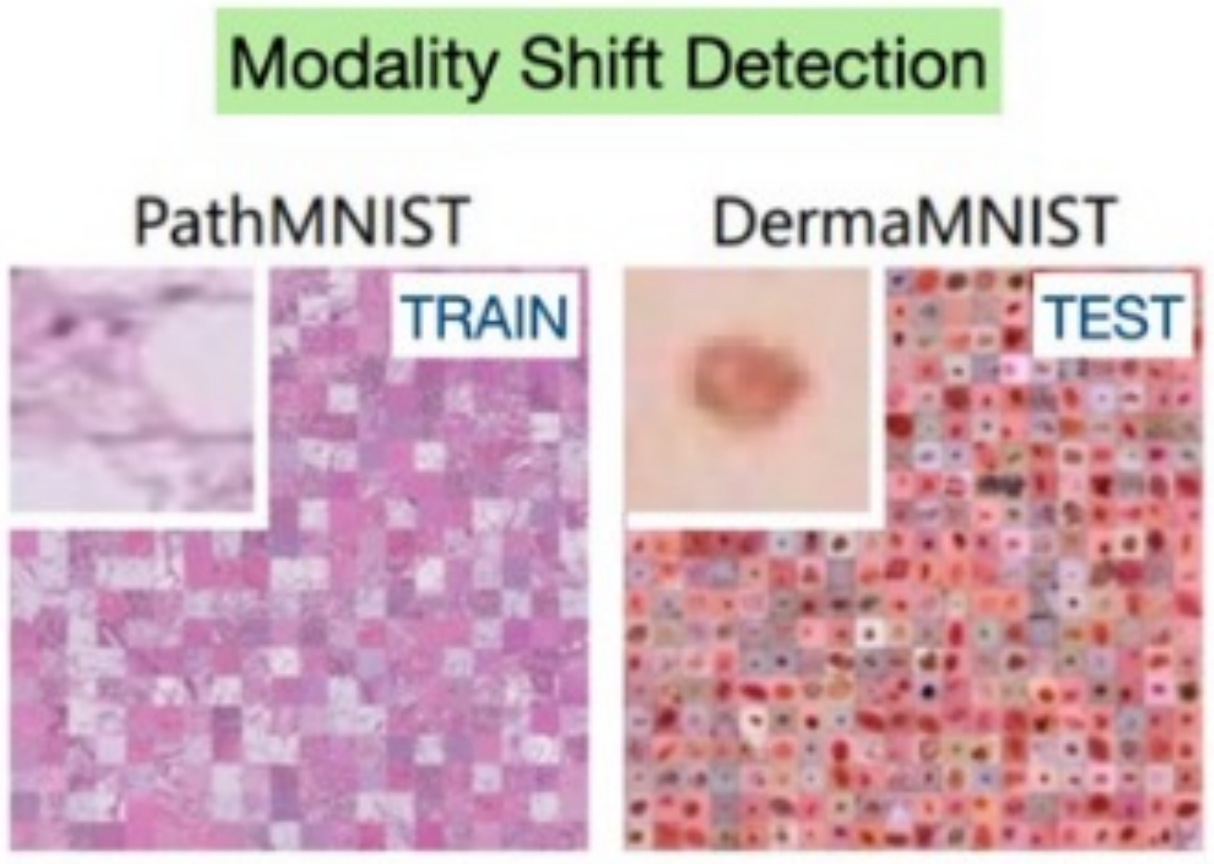


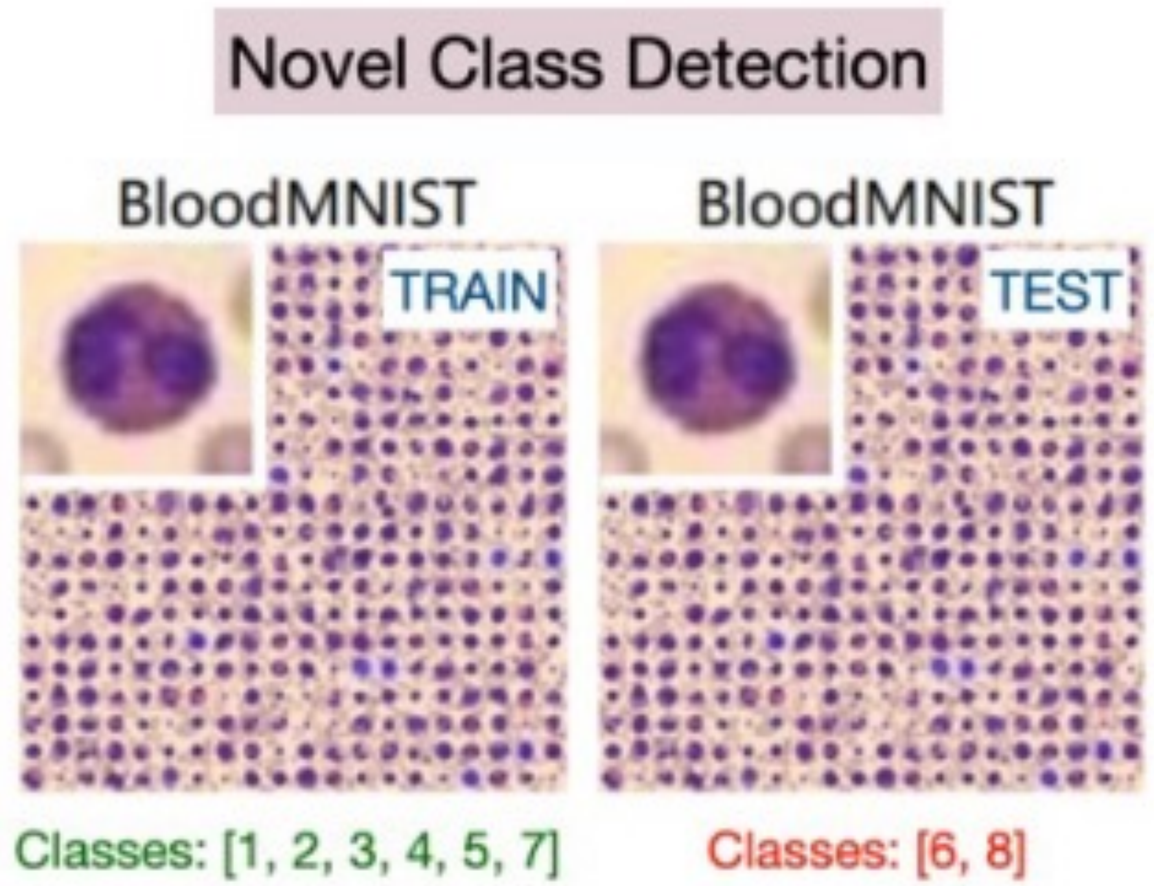


G-ODIN	OOD Rejection AUROC: 49.7 ✗	OOD Rejection AUROC: 75.4 ✗	OOD Rejection AUROC: 53.9 ✗
Augmix + VOS	OOD Rejection AUROC: 62.0 ✗	OOD Rejection AUROC: 39.8 ✗	OOD Rejection AUROC: 38.2 ✗
Augmix + NDA	OOD Rejection AUROC: 90.9 ✓	OOD Rejection AUROC: 43.5 ✗	OOD Rejection AUROC: 53.5 ✗
Ours	OOD Rejection AUROC: 97.5 ✓	OOD Rejection AUROC: 98.1 ✓	OOD Rejection AUROC: 89.1 ✓

Using the Combination of Latent-Space Inliers and Synthetic Pixel-Space Outliers Leads to Powerful OOD Detectors

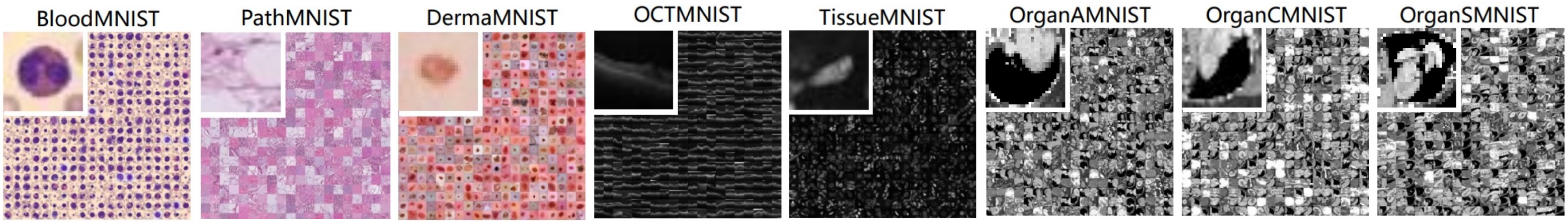






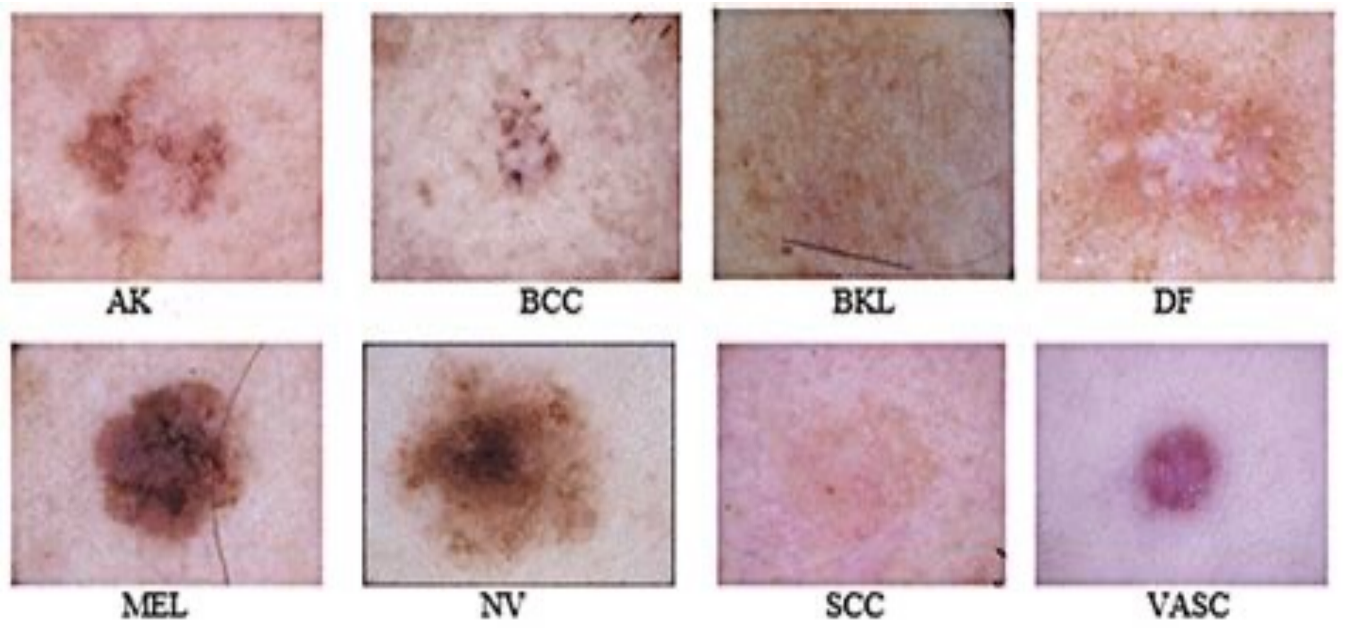
G-ODIN	OOD Rejection AUROC: 49.7 ✗	OOD Rejection AUROC: 75.4 ✗	OOD Rejection AUROC: 53.9 ✗
Augmix + VOS	OOD Rejection AUROC: 62.0 ✗	OOD Rejection AUROC: 39.8 ✗	OOD Rejection AUROC: 38.2 ✗
Augmix + NDA	OOD Rejection AUROC: 90.9 ✓	OOD Rejection AUROC: 43.5 ✗	OOD Rejection AUROC: 53.5 ✗
Ours	OOD Rejection AUROC: 97.5 ✓	OOD Rejection AUROC: 98.1 ✓	OOD Rejection AUROC: 89.1 ✓

With a Large Suite of Medical Imaging Benchmarks, We Systematically Evaluate our Proposed Approach



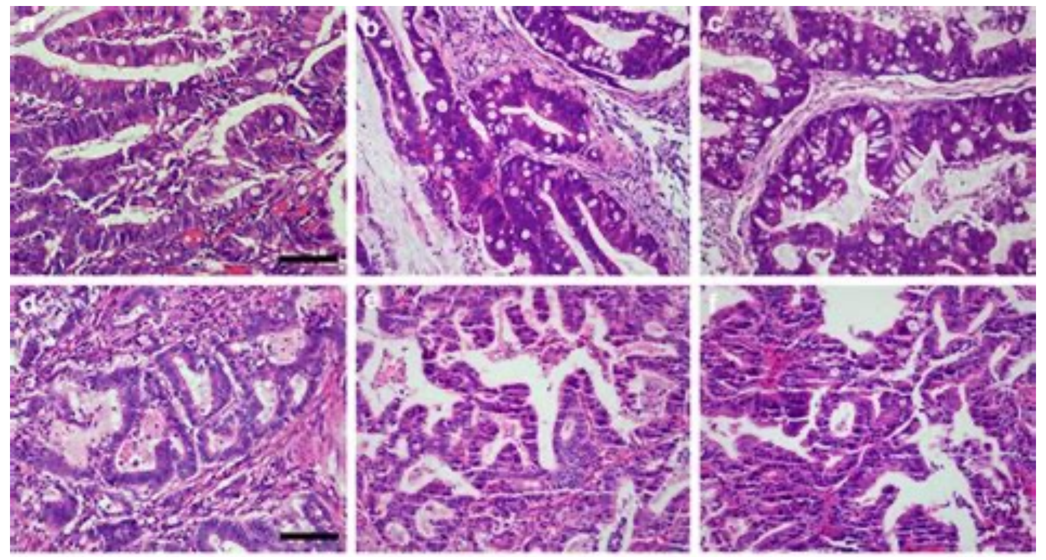
- **Modality shifts**: For each case, samples from all other datasets
- **Semantic shifts**: Held-out classes from the training dataset
- **Architecture**: WideResNet 40-2

ISIC 2019 Skin Lesion



- **Modality shifts:** CXR, WILDS, Retina images
- **Semantic shifts:** Held-out classes from train dataset, Clin Skin (control group), Derm Skin
- **Architecture:** ResNet 50

Colorectal Cancer



- **Modality shifts:** Clin Skin, Derm Skin, CXR, WILDS, Retina images
  Semantic shifts: Held-out classes from train dataset
- **Architecture:** ResNet 50

Evaluation

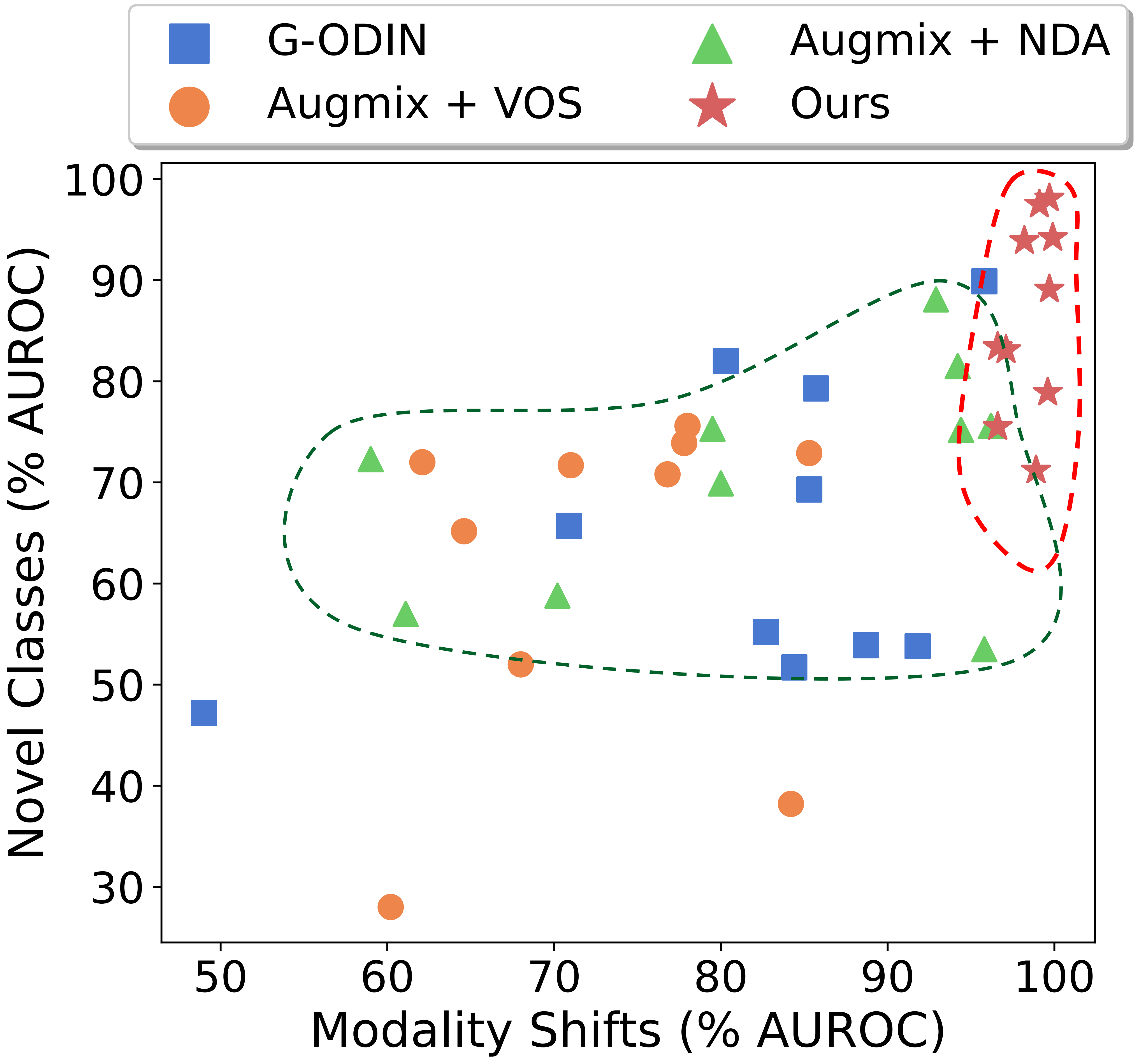
Balanced Accuracy

Modality Shift Detection

Novel Class Detection

AUROC

Strikingly, our Calibration Approach Leads to Significantly Superior Detection Performance in all Cases



Across all benchmarks, our approach achieves large gains over existing baselines in both modality shifts and novel classes

	G-ODIN	Augmix + VOS	Augmix + NDA	Ours
Modality Shifts	81.5	72.8	82.3	98.5
Novel Classes	64.8	62.1	70.7	86.49

Existing baselines tend to produce large variances in AUROC scores across datasets

Visualization of the Energy Scores for ID and OOD Data Clearly Reveals the Benefits of the Proposed Calibration Protocol

Augmix + VOS | Augmix + NDA | Ours

ID: BloodMNIST

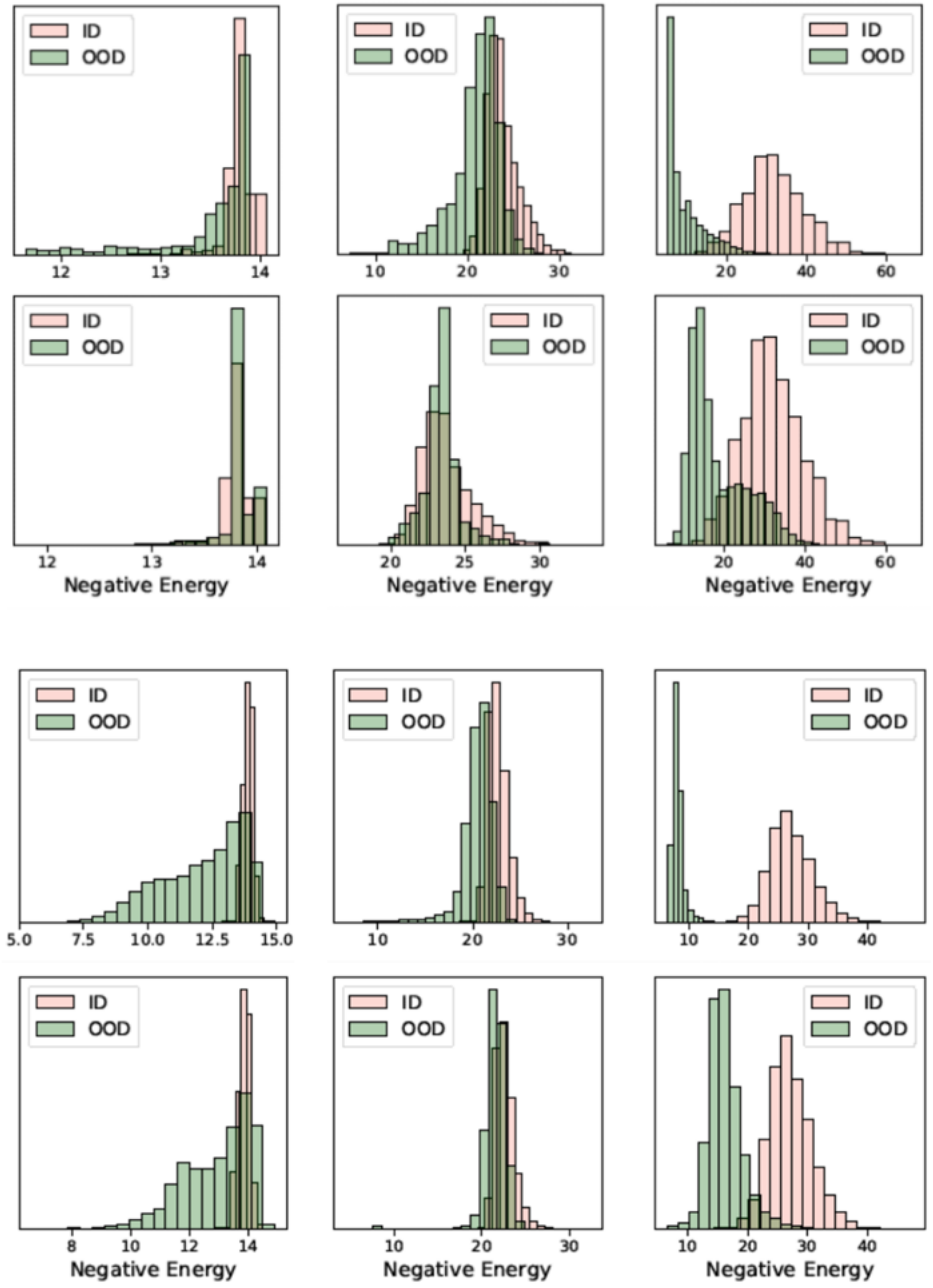
Modality Shift: Derma MNIST

Novel Classes

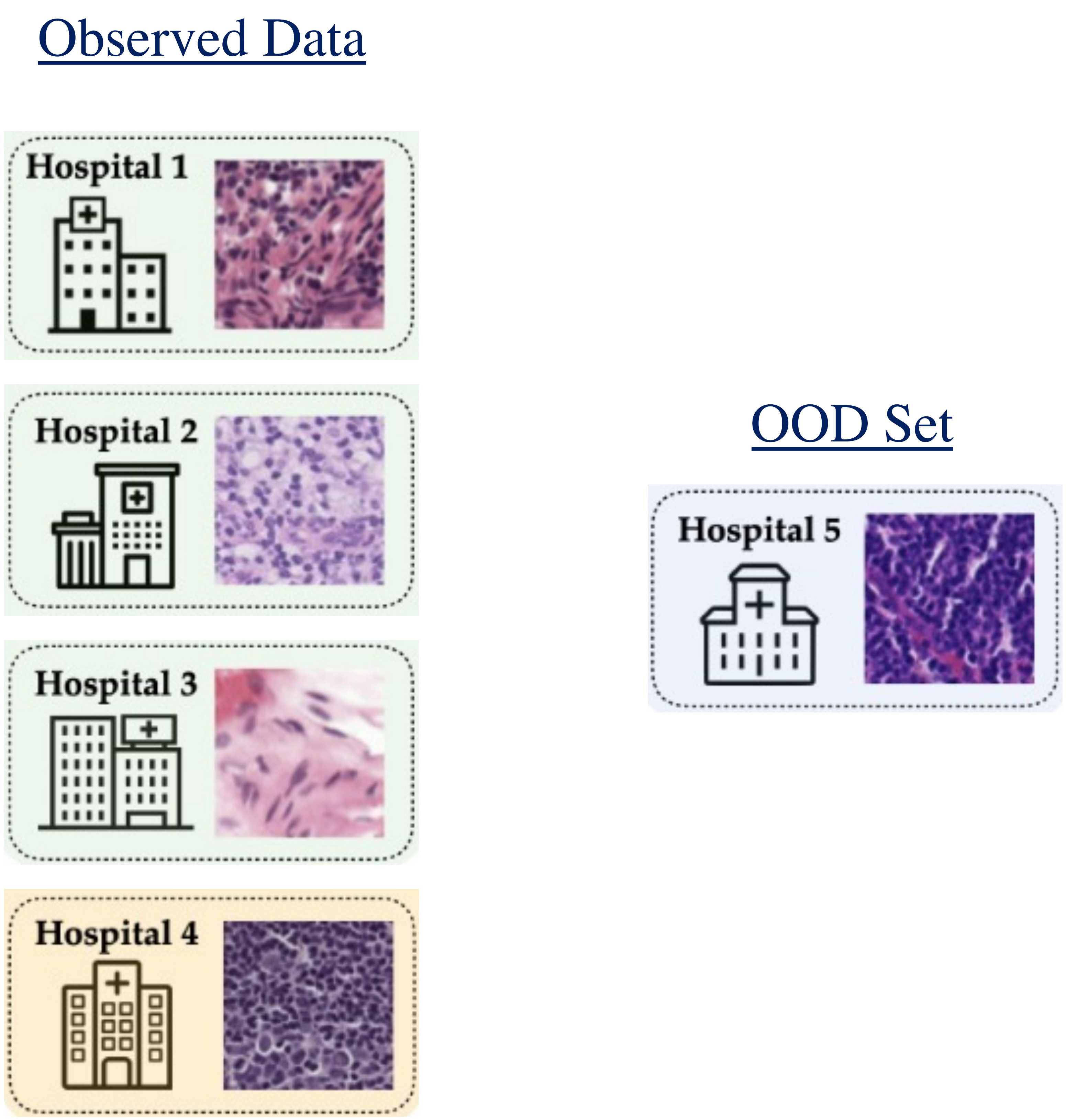
ID: OrganAMNIST

Modality Shift: Tissue MNIST

Novel Classes



Interestingly, our Approach can Even Detect Nuanced Covariate Shifts Arising from Different Hospitals



WILDS Chamelyon Benchmark

	AUROC (%)
Augmix + VOS	79.5
Augmix + NDA	36.5
Ours	87.4

While latent-space outliers are superior to pixel-space outliers synthesized via negative data augmentation, our approach performs the best

Summary: A New State-of-the-Art Baseline for Medical OOD Detection that can be used with any Imaging Modality or Model Architecture

Calibrating OOD detectors is significantly challenging with medical imaging data, and existing solutions from the vision literature do not work effectively!

We find that the choice of space for synthesizing augmentations is critical when calibrating OOD detectors for open-set data

We advocate for the use of virtual inliers from the classifier's latent-space and diverse pixel-space outliers with energy-based training

Using a large suite of medical imaging benchmarks, we show state-of-the-art open-set recognition performance (both modality shifts and novel classes), as well as in detecting covariate shifts arising from different hospital data.

Know Your Space - Inlier and Outlier Construction for Calibrating Medical OOD Detectors

Vivek Narayanaswamy * [1] Yamen Mubarka * [1] Rushil Anirudh [1] Deepta Rajan [2] Andreas Spanias [3]
Jayaraman J. Thiagarajan [1]

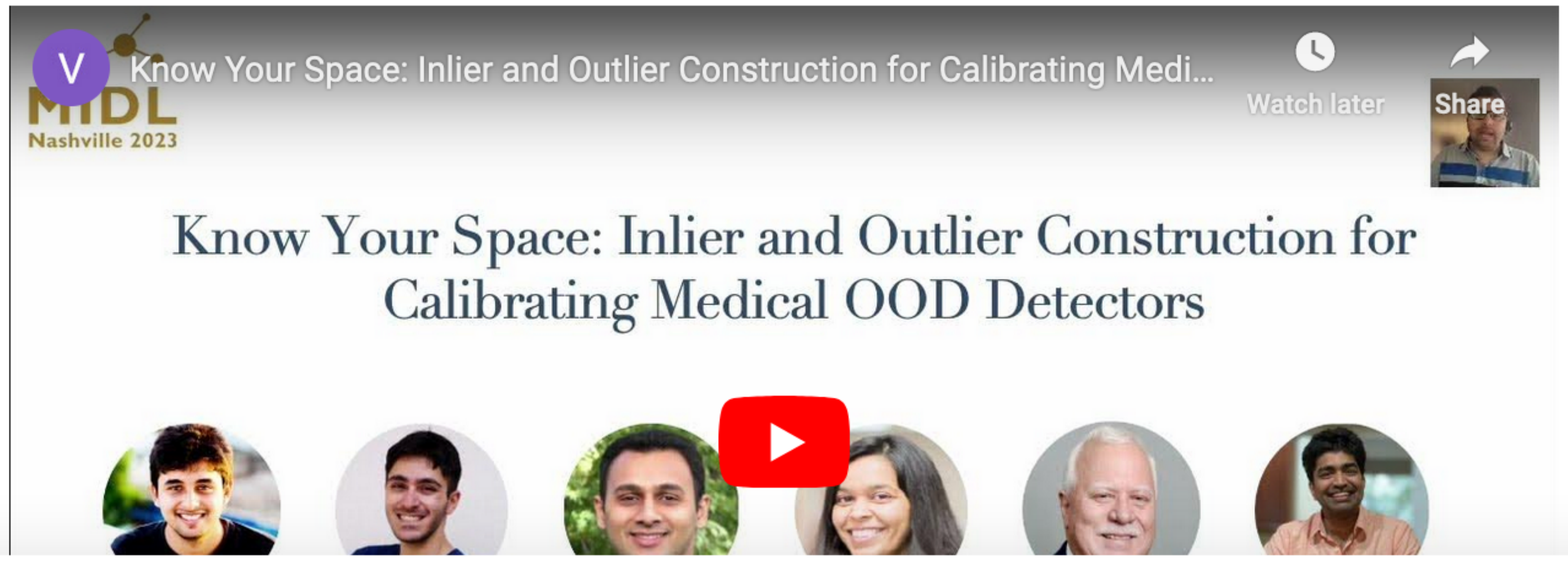
[1]Lawrence Livermore National Laboratory, CA, USA [2]Microsoft, WA, USA [3]Arizona State University, AZ, USA

Paper Slides Poster Video Github

Summary

Our primary focus is on developing well-calibrated out-of-distribution (OOD) detectors to ensure the safe deployment of medical image classifiers. The use of synthetic augmentations has become common for specifying regimes of data inliers and outliers. However, our research findings highlight the substantial influence of both the synthesis space and the type of augmentation on the performance of OOD detectors. After conducting an extensive study using medical imaging benchmarks and open-set recognition settings, we recommend employing a combination of virtual inliers in the classifier's latent space and diverse synthetic outliers in the pixel space. This approach proves highly effective in producing OOD detectors with superior performance.

Video



Codes

SCAN ME

Website

SCAN ME

For questions contact
narayanaswam1@llnl.gov