Out-of-Distribution Learning:

Preventing Fatal Neural Network Mistakes

In Distribution v. Out-of-Distribution

- Traditional Neural Networks (NNs) accurately classify objects seen during training
  - These are **In-Distribution (ID)** objects[1]

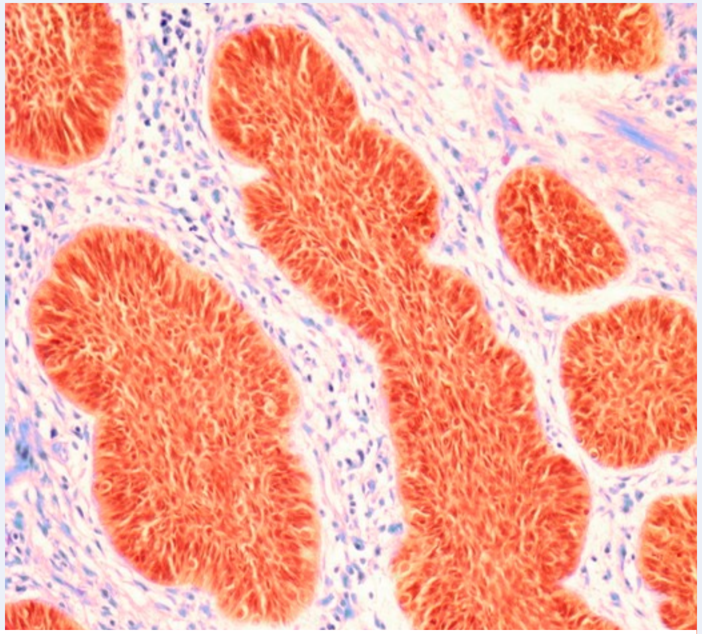
eg. cancer-detecting NN accurately classifies ID tumor

- Traditional NNs confidently misclassify objects NOT seen during training
  - These are **Out-of-Distribution (OOD)** objects[1]

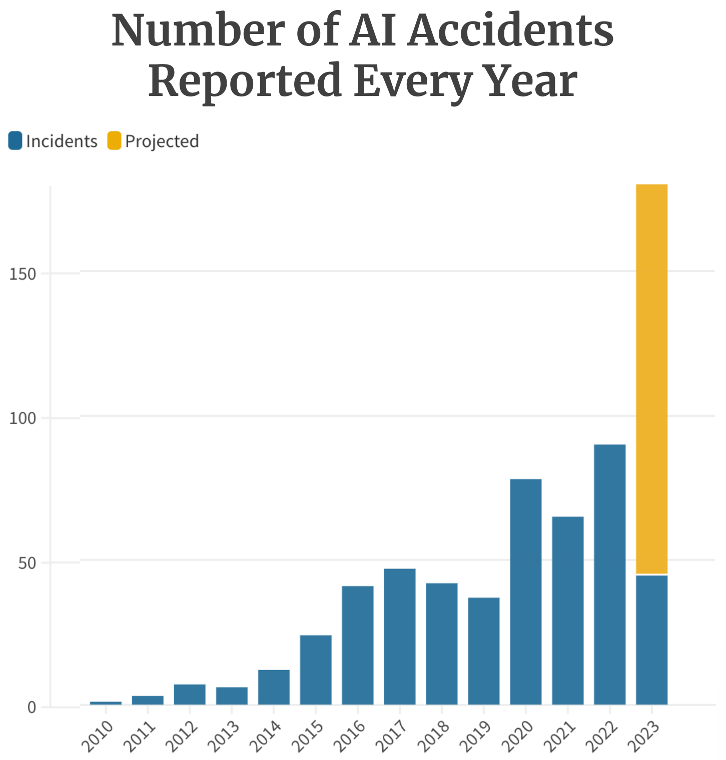
eg. cancer-detecting NN ***misclassifies*** OOD debris as cancer, causing ***misdiagnosis***

Fatal Out-of-Distribution Misclassifications

Experts predict NN accidents will double every year



The Atlantic: “Cancer Detector Misdiagnoses Black Users” due to lack of training & testing with Black samples[2]



Source: AI Incident Database, founded by Sean McGregor Ph.D. in machine learning[3]

NBC: Self-driving car killed a woman because it was trained to “only classify an object as a pedestrian [if it’s] near a crosswalk”[4]

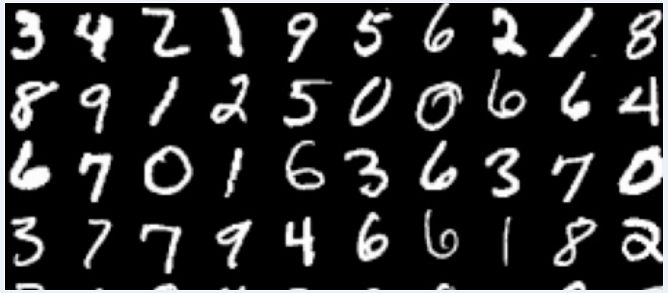
When NNs are implemented in the world, they encounter OOD objects and make fatal misclassifications

OOD Classification Issue

- Unfeasible to predict which OOD objects a NN will encounter
- Unfeasible to train NNs on all OOD objects

New algorithm that can classify OOD objects w/ minimal training must be implemented

Datasets

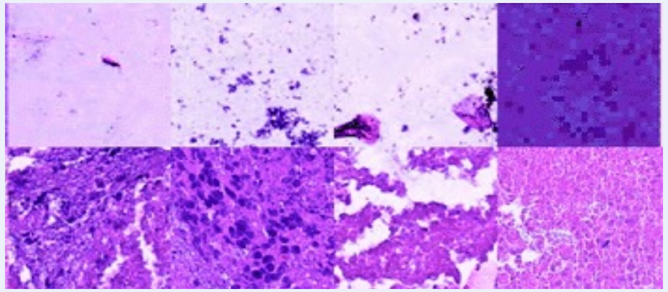


MNIST[5]

- handwritten digits
- 60000 train, 10000 test images

OrganAMNIST[6]

- low-res organ scans
- 34581 train, 6491 test images



PathMNIST[6]

- colorectal tissue
- 89996 train, 10004 test images

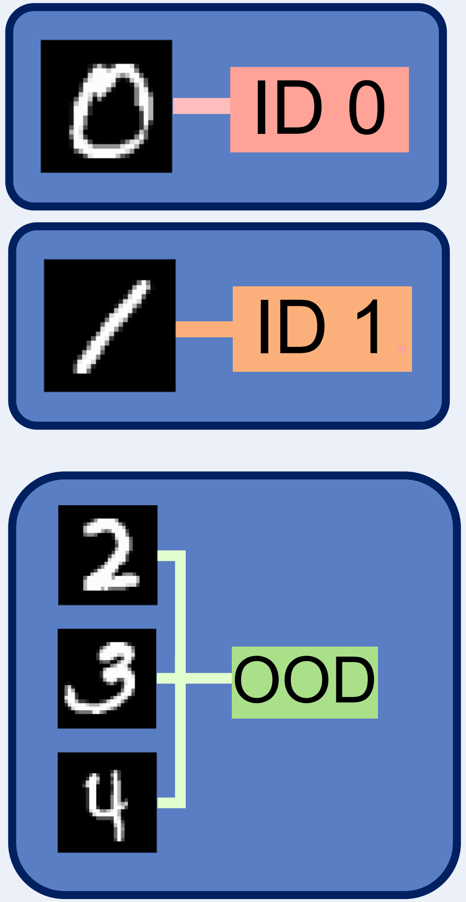
Step 1: Train NN w/ OOD class to Differentiate ID & OOD

Train **10 NNs** to account for chance differences during training

Start w/ traditional NN (**2 ID classes**, 5-layer LeNet CNN, exponential learning rate, 15 epochs)

Add *general* **OOD class** (mix of 3 non-ID objects)

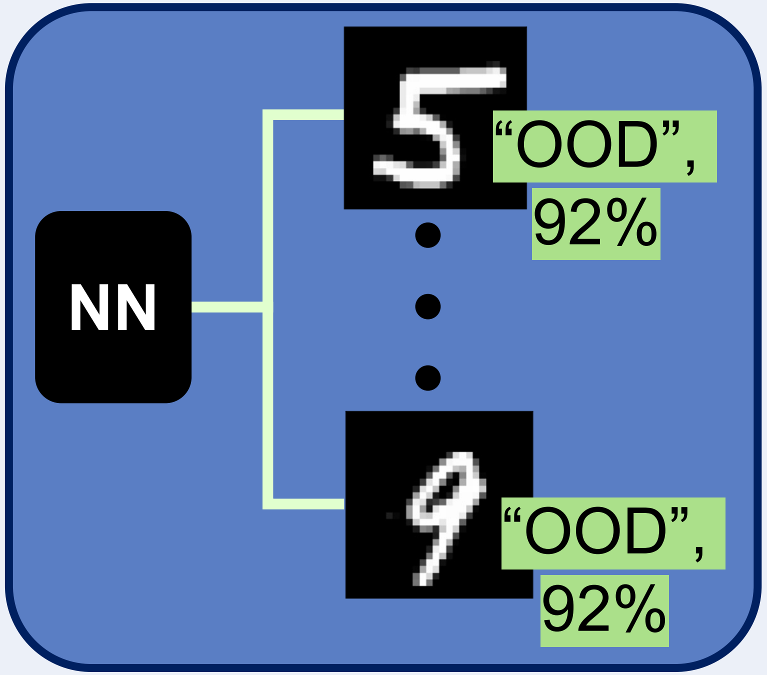
Train NN



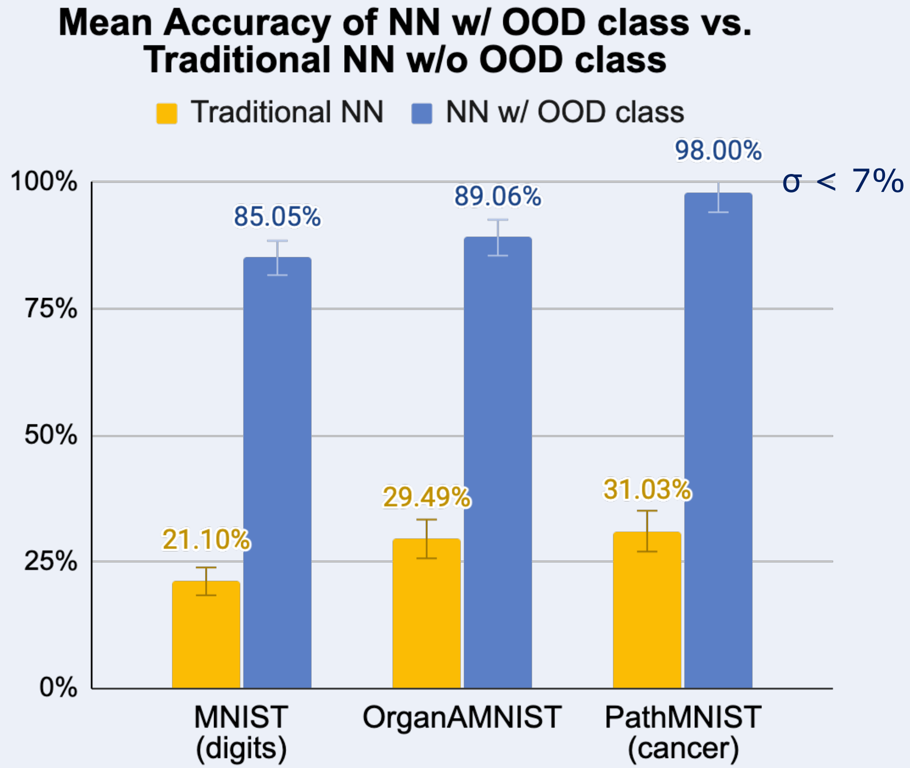
NN trained to classify all non-ID objects as OOD

Test NN's Ability to Differentiate ID & OOD

Test new NN w/ OOD class on OOD objects not seen during training



NN w/ OOD class classifies digits 5-9 as "OOD" even though it wasn't trained using these digits



NN w/ OOD class ~60% more accurate than Traditional NN w/o OOD class

NN generalizes trained concept of OOD to objects not seen during training

Discovering OOD Objects Not Seen During Training

There are 2 types of OOD objects: **OOD seen** (seen during training) & **OOD unseen** (not seen during training) both stored in NN’s “OOD” class

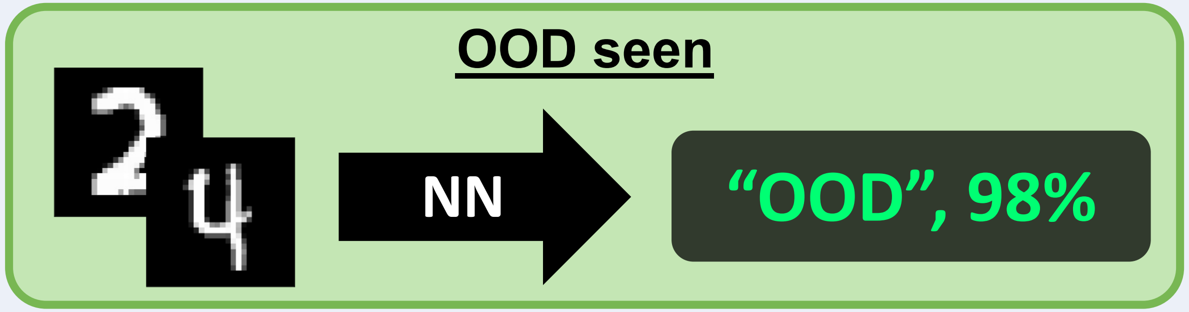
Fig. PathMNIST NN’s ID, OOD seen, OOD unseen classes

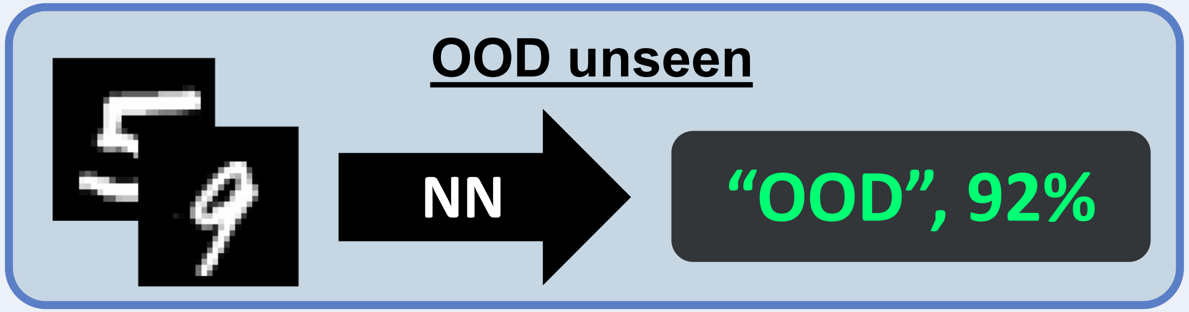
Algorithm should create new class for undiscovered **OOD unseen** objects like carcinoma to alert users to existence of important undiscovered objects

Step 2: Unsupervised Algorithm to Discover OOD Unseen Objects

NN classifies **OOD seen** objects with ***higher confidence***

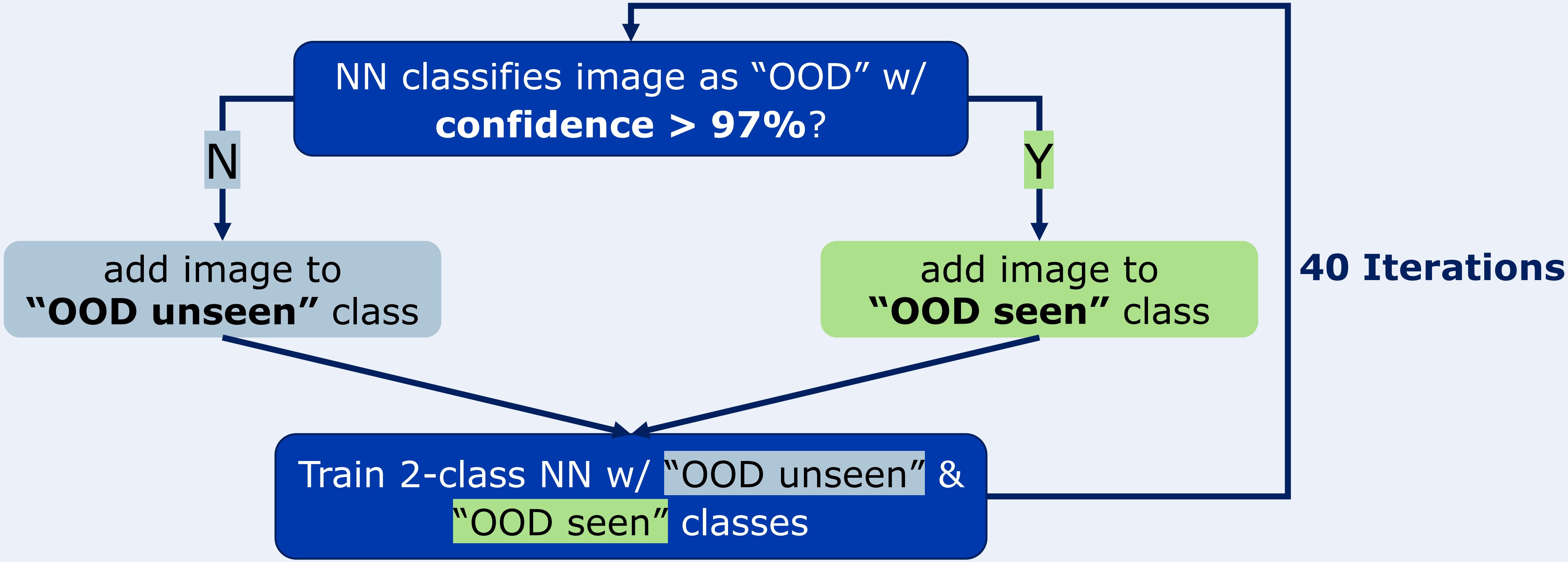
than **OOD unseen** objects





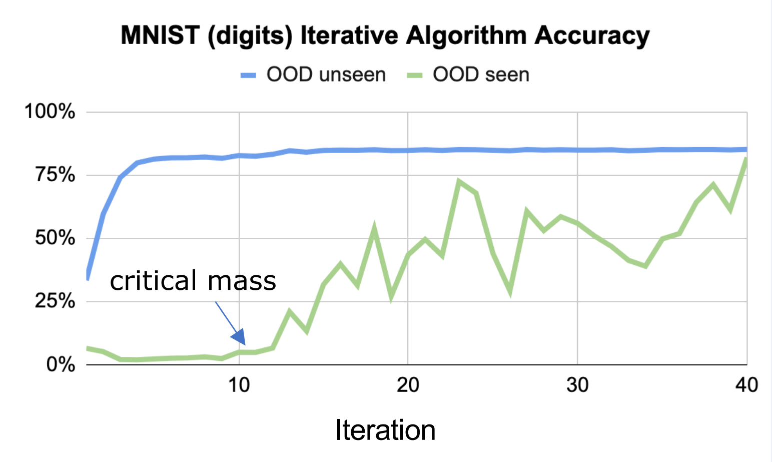
Step 2: Unsupervised Algorithm to Discover OOD Unseen Objects

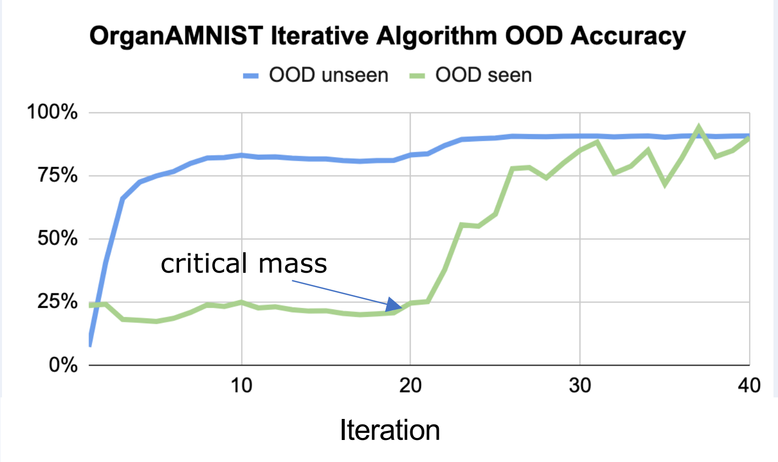
Novel unsupervised algorithm uses confidence scores & NN training to differentiate **OOD seen** & **unseen** objects w/o user labeling:

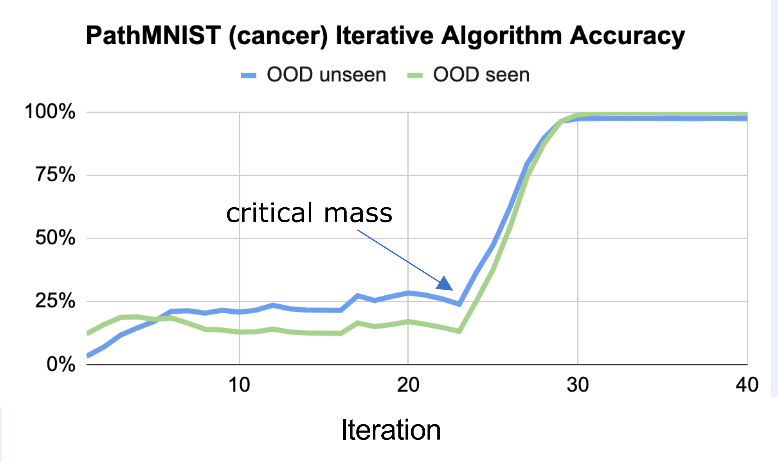


Algorithm **iterates 40 times**, correcting misclassifications & improving understanding of differences between **OOD unseen** & **seen** classes

Test Unsupervised Algorithm's Ability to Differentiate OOD Seen & OOD Unseen







Accuracies increase quickly after critical mass is reached (~25%)

Unsupervised Algorithm's Accuracies After 40 Iterations

	OOD seen	OOD unseen
MNIST (digits)	82.05%	85.14%
OrganAMNIST	90.02%	90.70%
PathMNIST (cancer)	99.25%	97.37%

Unsupervised Algorithm attains above 80% accuracies, some reaching 99%

Conclusion

- Addition of OOD class improved accuracy by ~60%
- Algorithm can now classify objects not seen in training
- Algorithm discovers new objects before humans
- Saves on user image labeling
  - Trained w/ ~3000 images
  - Self-labels ~9000 images

Cancer, Organ, Digit classification success → algorithm has wide range of applicability

Limitations

- Neural Networks only successful if Out-of-Distribution objects look notably different from In-Distribution

Future Work

- Implement image augmentation & JSD loss for low-res datasets & datasets with very similar classes

Bibliography

My GitHub Repo: https://github.com/olimu/OOD

1. D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017. https://arxiv.org/pdf/1610.02136.pdf
2. A. Lashbrook, “Ai-driven dermatology could leave dark-skinned patients behind,” The Atlantic, https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/.
3. R. Minto, “Ai accidents are set to skyrocket this year,” Newsweek, https://www.newsweek.com/ai-accidents-set-skyrocket-this-year-1795928.
4. P. McCausland, “Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk”, NBC News, https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281.
5. Y. LeCun, C. Cortes, and J. C. Burges, “The MNIST Database of Handwritten Digits”, 1999.
6. J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, “MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification,” *Sci Data 10*, 41 (2023). https://doi.org/10.1038/s41597-022-01721-8.
7. D. Hendrycks, M. Mazeika, T. Dietterich, “Deep Anomaly Detection with Outlier Exposure,” in *Proc. Int. Conf. Learn. Representations*, 2019.
8. C. Geng, S. Huang, and S. Chen, “Recent Advances in Open Set Recognition: A Survey”, 2015, https://arxiv.org/pdf/2110.11334.pdf
9. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient-based learning applied to document recognition”. *Proceedings of the IEEE*. 86 (11): 2278–2324. doi:10.1109/5.726791.
10. N. B. Erichson, S. H. Lim, W. Xu, F. Utera, Z. Cao, and M. W. Mahoney, “NoisyMix: Boosting Model Robustness to Common Corruptions”. https://arxiv.org/abs/2202.01263.

Comparing NN w/ OOD class to Traditional NN

T-Test Comparing Accuracies of NN w/ OOD class to Accuracies of Traditional NN

Dataset	p-value ($\alpha = 0.05$)
MNIST (digits)	5.68×10^{-12}
OrganAMNIST	3.01×10^{-9}
PathMNIST (cancer)	3.41×10^{-11}

P-values $< \alpha$
so NN w/ OOD class
is more accurate
than Traditional NN
w/o OOD class