Can AI Recognize Unseen/Untrained Objects?

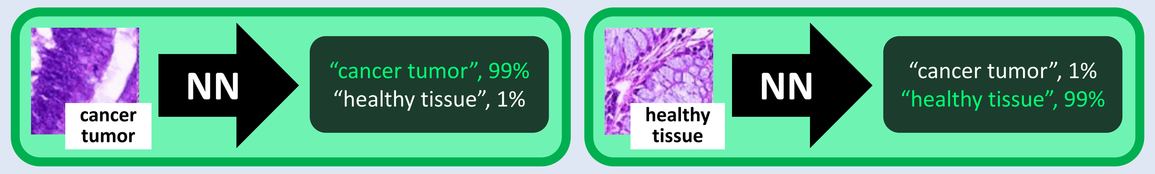
Mitigating the Impact of Out-of-Distribution Objects for Digit & Medical Classification Problems

SCM102

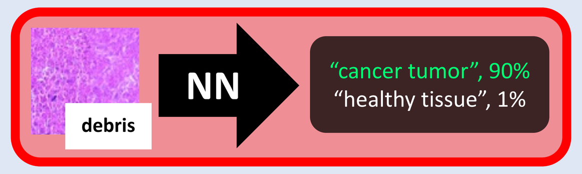
Research Problem

Making NNs Applicable to Real Life

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



- NNs confidently misclassify objects NOT seen in training
  - **Out-of-Distribution (OOD)** objects [1]



- OOD objects are rare & overlooked, but even 1 misclassification can be fatal

OOD Classification Issues

- NN doesn't know what OOD objects it will encounter
  - You can't train a NN on ALL OOD objects to prepare it
    - So new algorithms must be designed & implemented

Background

- > 90% NN research focuses on improving ID classification
- Generalized OOD field includes OOD & OSR
  - OOD – differentiate between ID & OOD datasets
  - OSR – differentiate between ID & OOD classes of single dataset

OOD Literature Review

- Thresholding SoftMax confidences [1]
  - ID objects (in train set) should be classified confidently
  - OOD unseen inputs should be less confidently classified
- Auxiliary OOD class [2]
  - NN learns representation of OOD

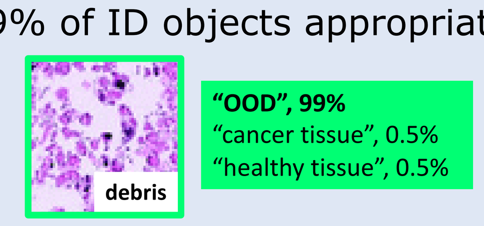
OSR Literature Review

- Generative Adversarial Networks [3]

New Algorithms

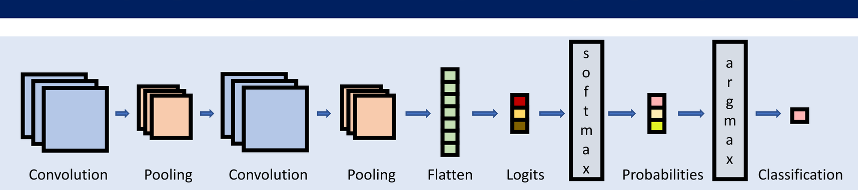
OOD Classification w/ Minimal Training

- Prevent fatal misclassification
  - Classify 85% of OOD objects as "OOD"
  - Classify 99% of ID objects appropriately



- Reduce need for image labeling by user

Basic NN Structure



5-layer LeNet (CNN) [4], 15 epochs, Exponential learning rate of 0.001

References

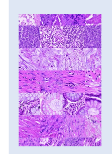
1. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017. https://arxiv.org/pdf/1610.02136.pdf
2. D. Hendrycks, M. Mazeika, T. Dietterich, "Deep Anomaly Detection with Outlier Exposure," in *Proc. Int. Conf. Learn. Representations*, 2019.
3. C. Geng, S. Huang, and S. Chen, "Recent Advances in Open Set Recognition: A Survey", 2015, https://arxiv.org/pdf/2110.11334.pdf
4. 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.
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. 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.

Set-Up

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

Dataset Class Structure

- **Train set** – train NN to differentiate ID & OOD
  - ID: ID class 0, ID class 1
    - Primary classes NN is trained to detect
  - OOD: OOD train
    - Mixed image set of 3 known OOD objects
- **Test set** – evaluate ID & OOD classification
  - ID: ID class 0, ID class 1
  - OOD: OOD unseen
    - Mixed image set of true OOD objects (not in train set)

Original class #	MNIST	OrganAMNIST	PathMNIST	Altered Class
0	"handwritten 0"	"left lung"	"cancerous stroma"	ID class 0
1	"handwritten 1"	"right lung"	"normal colon mucosa"	ID class 1
2	"handwritten 2"	"femur"	"debris"	OOD train
3	"handwritten 3"	"spleen"	"background"	
4	"handwritten 4"	"kidney"	"adipose"	
5	"handwritten 5"	"bladder"	"mucus"	OOD unseen
6	"handwritten 6"	"heart"	"lymphocytes"	
7	"handwritten 7"	"liver"	"smooth muscle"	
8	"handwritten 8"	"pancreas"	"colorectal adenocarcinoma"	
9	"handwritten 9"			

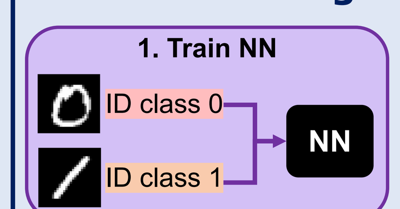
Procedure Phase 1:

Preventing Fatal Misclassification

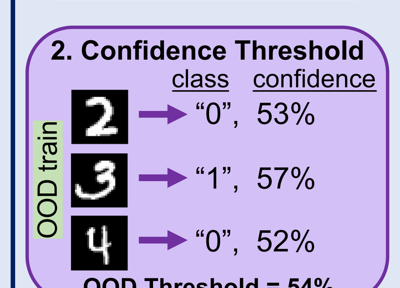
Algorithm 1: Thresholded NN

Procedure

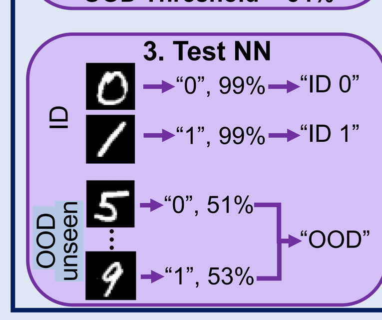
Confidence level threshold for OOD detection using Traditional 2-class NN



1. Train NN w/ ID class 0 & ID 1
   - NN expected to
     - confidently classify ID
     - less confidently classify OOD

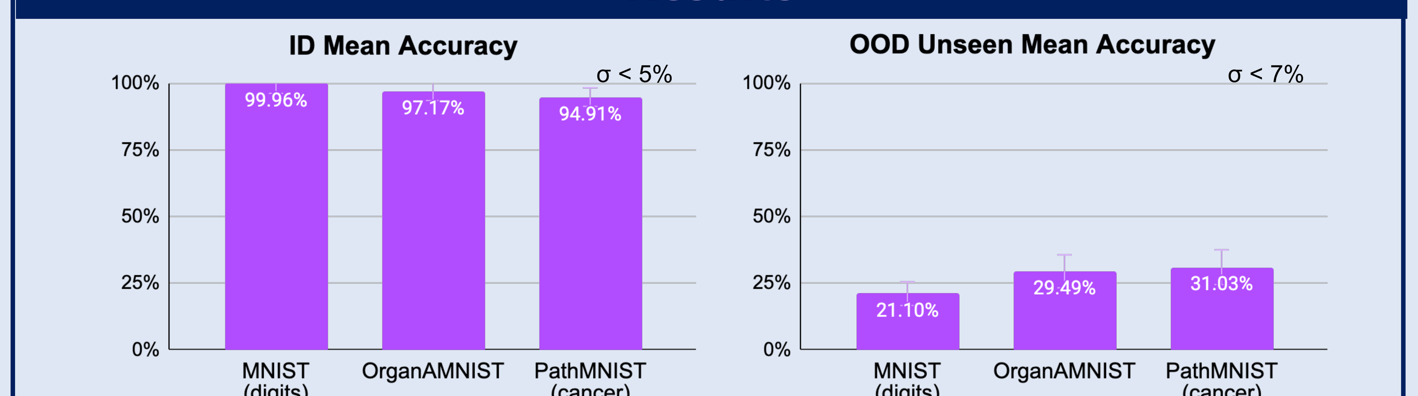


2. Use confidence threshold to differentiate ID & OOD
   - Threshold = average of OOD train confidences



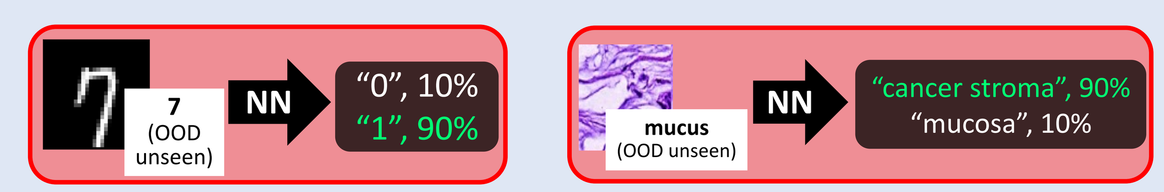
3. Test NN w/ ID & OOD unseen
   - Confidence < threshold → image is OOD unseen
   - Confidence > threshold → image is ID

Results



Thresholded NN accurately classifies ID, inaccurately classifies OOD unseen

Analysis of Results: Inaccurate OOD

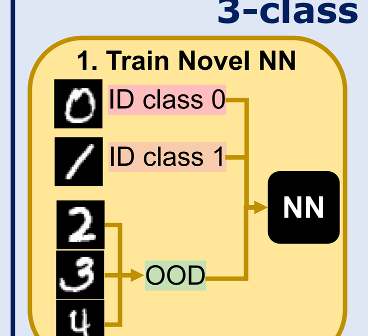


Single threshold value can't represent the entire general OOD class → OOD unseen misclassified as ID

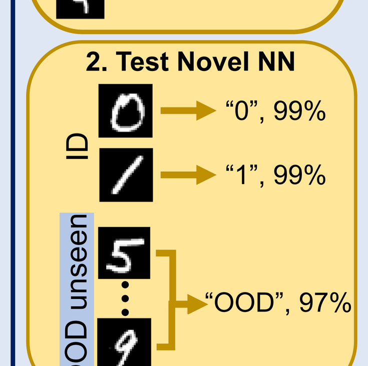
Algorithm 2: Novel NN

Procedure

3-class NN w/ Novel OOD class

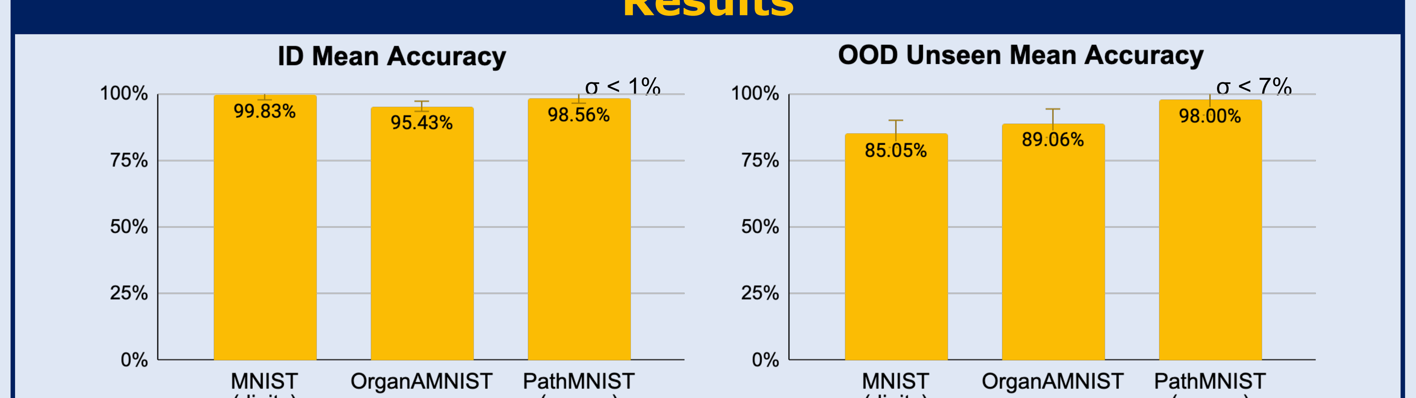


1. Train NN w/ ID class 0, ID class 1, **OOD train**
   - NN expected to
     - confidently classify ID
     - confidently classify OOD



2. Test NN w/ ID & OOD unseen

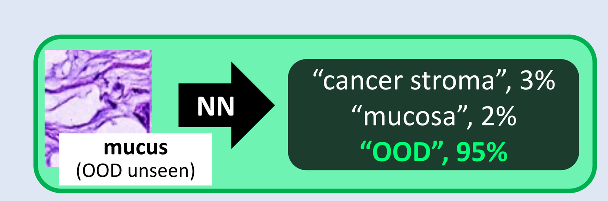
Results



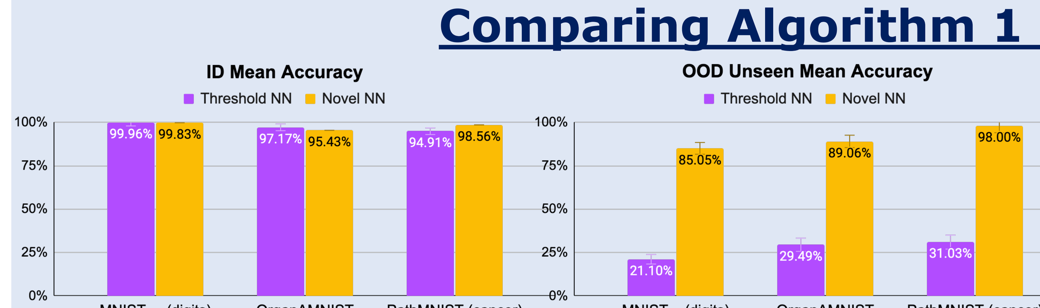
Novel NN accurately classifies both ID & OOD unseen

Analysis of Results: Accurate OOD

- OOD class in train set → NN tunes many weights & biases for accurate, general representation of entire OOD class
  - Accurate representation → preventing misclassification



Comparing Algorithm 1 & 2:



Novel NN outperforms Thresholded NN by…

- **63.95%** for MNIST (digits)
- **59.57%** for OrganAMNIST
- **66.97%** for PathMNIST (cancer)

- Both NNs achieve > 90% ID accuracies
- Only Novel NN achieves > 85% OOD unseen accuracy

T-Test Analysis

Thresholded v. Novel OOD Unseen Accuracies

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 < α → Novel NN's OOD unseen accuracy is *significantly* greater than the Thresholded NN's

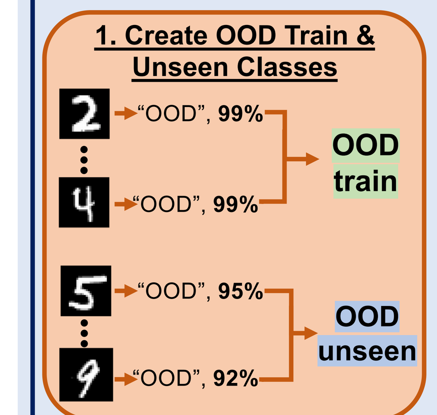
Procedure Phase 2:

Reducing Image Labeling & Discovering new OOD objects

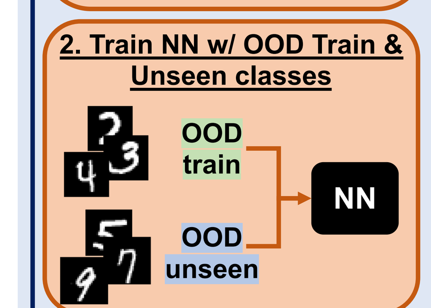
Algorithm 3: Semi-Supervised Iterative

Procedure

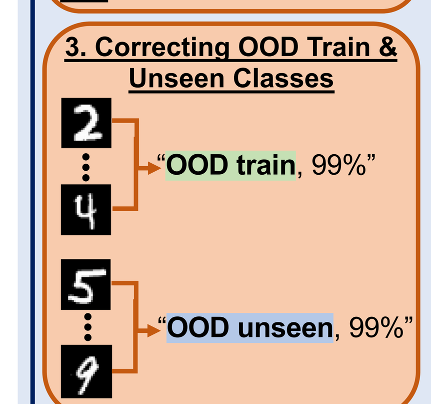
Algorithm self-labels OOD unseen images & trains 2-class NN to differentiate OOD unseen & OOD train



1. Create OOD train & unseen classes
   - Use confidence threshold (97%) to differentiate OOD train & unseen
     - Confidence < 97% → add image to new OOD unseen class
     - Confidence > 97% → add image to new OOD train class



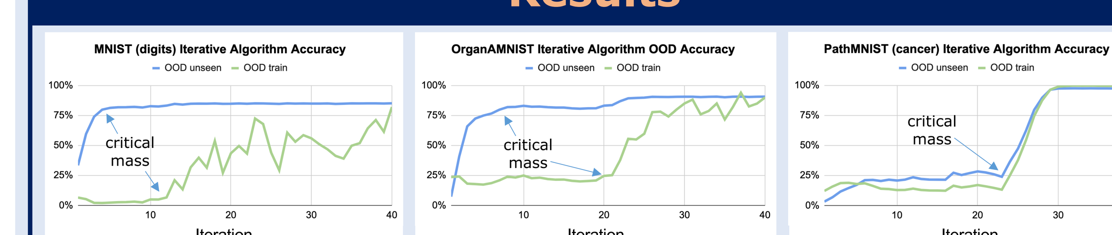
2. Train 2-class NN w/ OOD train & unseen classes (created in Step 1)
   - Improves NN's understanding of OOD unseen & OOD train objects



3. Correcting OOD train & unseen classes
   - After its understanding improved, NN reclassifies images into OOD train & unseen
     - Corrects misclassifications

4. Iterate Steps 2 & 3 40 times
   - Step 3 corrects misclassifications made in previous iterations
   - Step 2 retrains NN w/ corrections → better understanding
   - More iterations → more corrections → better understanding → higher accuracy

Results



Accuracy increases quickly after critical mass is reached (~25%)

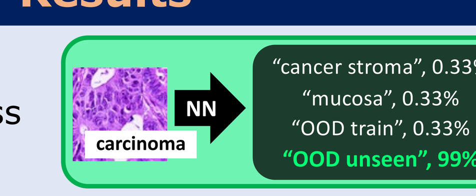
Iterative Algorithm's Accuracies After 40 Iterations

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

Accuracy increases over iterations → > 80% accuracies

Analysis of Results

- Saves on user image labeling
  - Self-labels an entire new class
    - Trains w/ ~3000 images
    - Self-labels ~9000 images
- Stores objects unseen by trainers
  - Likely to make discoveries



NN discovers carcinoma images, which were not in train set, & stores images in OOD unseen

Conclusion

Algorithm developed to enhance the capabilities of NNs for real life applications beyond controlled environments

With Iterative Algorithm NNs…

- are freed from limitation of training and testing on similar images
- classify objects completely different from those seen in training
- can discover new objects before humans

Limitations

- NNs only successful if OOD objects look notably different from ID

Future Work

- Implement image augmentation & JSD loss [7] for low-res datasets & datasets with very similar classes
  - Test limits of computer vision