# Deep Curriculum Learning in Task Space for Multi-Class Based Mammography Diagnosis

**Jun Luo**[a], Dooman Arefan[b], Margarita Zuley[b,d],
Jules Sumkin[b,d], and Shandong Wu[a,b,c]

[a]Intelligent Systems Program / [b]Dept. of Radiology / [c]Dept. of Biomedical Informatics and Dept. of Bioengineering,
University of Pittsburgh, Pittsburgh, PA USA
[d]Magee-Womens Hospital, University of Pittsburgh Medical Center

# Background

- Breast cancer is the second leading cause of cancer death in women
  - takes up 12% of new annual cancer cases globally [1]

- Full-Field Digital Mammography (FFDM)
  - Is the primary screening modality
  - Computer-aided detection and diagnosis methods were developed on FFDM

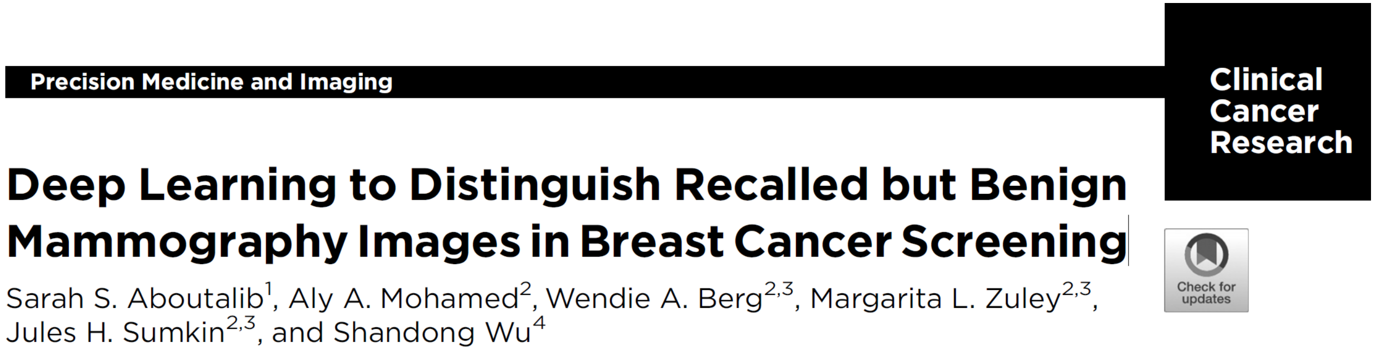
- Deep learning
  - In recent years, many deep-learning-based methods have been developed to enhance computer-aided detection and diagnosis in FFDM

[1] "U.S. Breast Cancer Statistics." American Cancer Society, 4 February 2021, https://www.breastcancer.org/symptoms/understand_bc/statistics (2021).

# Background

- High recall rate (10%)* for additional workup
  - Approx. 5 million women recalled in the US annually
  - Approx. 80% of recall-biopsied are benign findings*
  - Unnecessary psychological stress, medical costs, and clinical workload

- In a previous work, we developed a deep learning-based model to identify false recalls
  - Showed promising effects



Precision Medicine and Imaging

Clinical Cancer Research

**Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening**

Sarah S. Aboutalib[1], Aly A. Mohamed[2], Wendie A. Berg[2,3], Margarita L. Zuley[2,3], Jules H. Sumkin[2,3], and Shandong Wu[4]

Check for updates

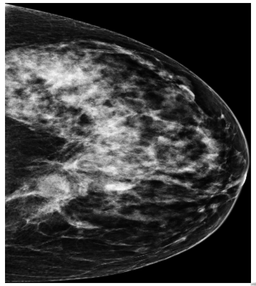
*American College of Radiology
https://www.acr.org/~/media/ACR/Images/Quality%20Safety/Resources/MammographyScreeningFacts.jpg
https://www.npr.org/sections/health-shots/2015/10/15/448888415/called-back-after-a-mammogram-doctors-are-trying-to-make-it-less-scary

# Purpose

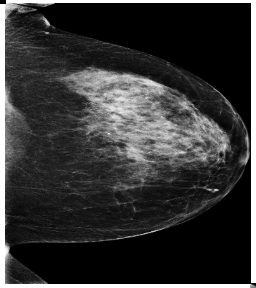
- Investigate a curriculum learning strategy of training a model to classify three classes of mammography images (i.e., malignant, false recall, and negative).

- Focus on the task space and create a sub-task of classifying the false recall cases against the combined group of the negative and the malignant cases. Push the model to first learn the “easier” sub-task prior to learning the “harder” original (3-class classification) task.

- Evaluate our learning strategy on a mammography dataset against a baseline learning strategy.



Malignant



False recall

Negative

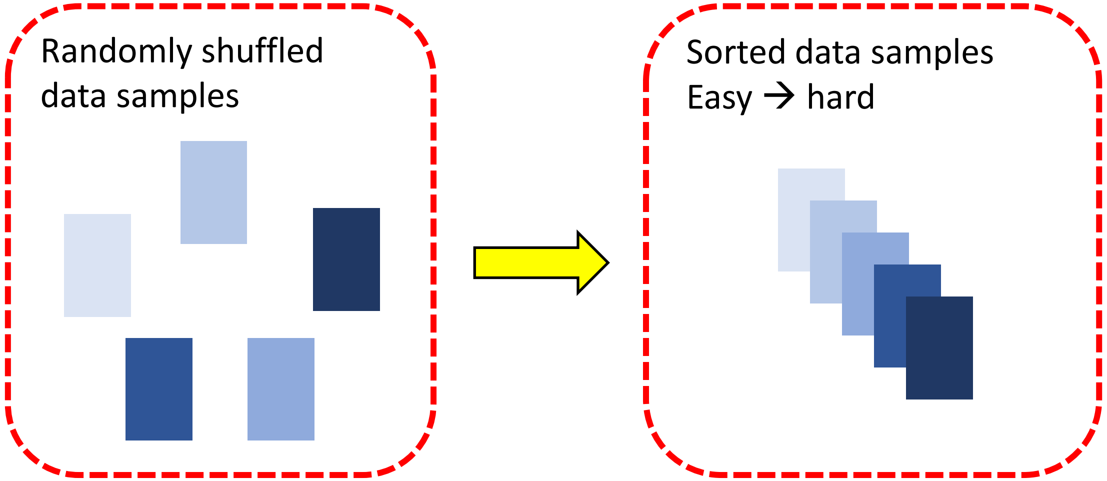
# Curriculum Learning

- Curriculum learning (CL)
  - Let model to learn like human with an “easy-to-hard” curriculum.
  - Medical knowledge can serve as an important factor to the definition of “easy” and “hard”. [2,3]

[2] Luo, J., Kitamura, G., Doganay, E., Arefan, D., and Wu, S., “Medical knowledge-guided deep curriculum learning for elbow fracture diagnosis from x-ray images,” in [Medical Imaging 2021: Computer-Aided Diagnosis], 11597, 1159712, International Society for Optics and Photonics (2021).
[3] Jim´enez-S´anchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M. A. G., and Piella, G., “Medical-based deep curriculum learning for improved fracture classification,” in [International Conference on Medical Image Computing and Computer-Assisted Intervention], 694–702, Springer (2019).

# Curriculum Learning

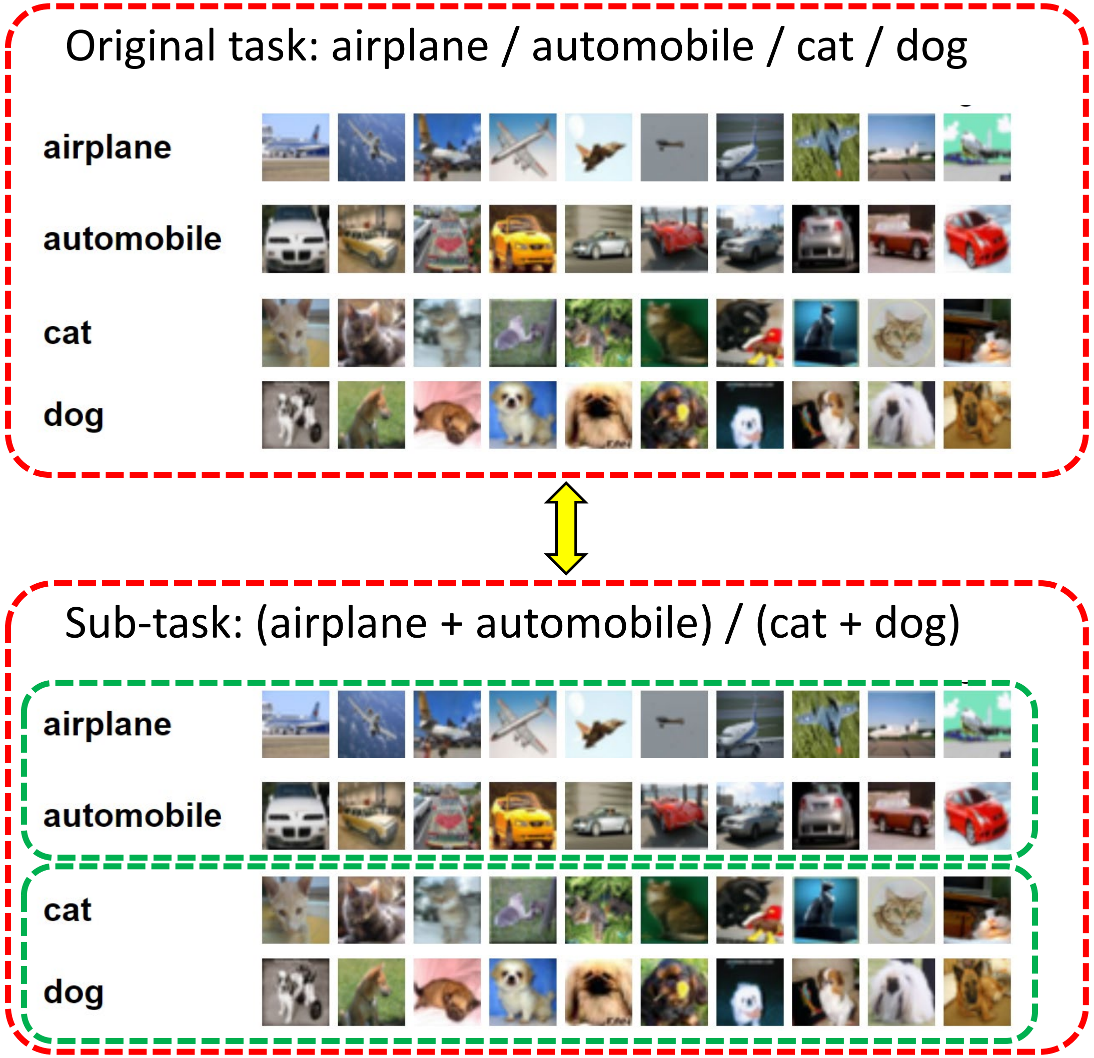
- Curriculum learning (CL)
  - Input space CL: design of the curriculum mainly focuses on the order of the training samples. [2,3]




[2] Luo, J., Kitamura, G., Doganay, E., Arefan, D., and Wu, S., "Medical knowledge-guided deep curriculum learning for elbow fracture diagnosis from x-ray images," in [Medical Imaging 2021: Computer-Aided Diagnosis], 11597, 1159712, International Society for Optics and Photonics (2021).
[3] Jim´enez-S´anchez, A., Mateus, D., Kirchhoff, S., Kirchhoff, C., Biberthaler, P., Navab, N., Ballester, M. A. G., and Piella, G., "Medical-based deep curriculum learning for improved fracture classification," in [International Conference on Medical Image Computing and Computer-Assisted Intervention], 694–702, Springer (2019).

# Curriculum Learning

- Curriculum learning (CL)
  - Input space CL: design of the curriculum mainly focuses on the order of the training samples
  - Output space (task space): focus more on the learning task itself. [4, 5]



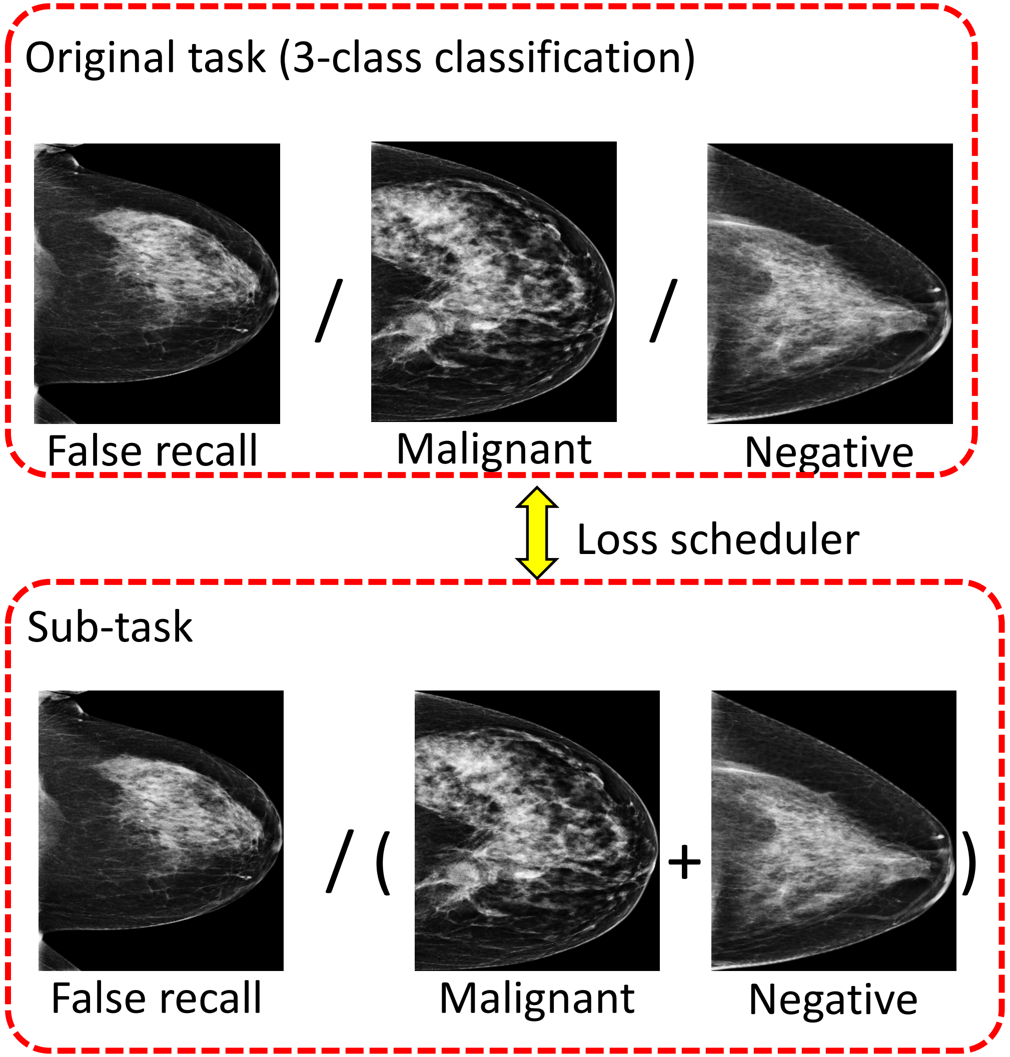


[4] Stretcu, O., Platanios, E. A., Mitchell, T., and P´oczos, B., "Coarse-to-Fine Curriculum Learning for Classification," in [International Conference on Learning Representations (ICLR) Workshop on Bridging AI and Cognitive Science (BAICS)], (2020).
[5] Saxena, S., Tuzel, O., and DeCoste, D., "Data parameters: A new family of parameters for learning a differentiable curriculum," (2019).

# Method

- Curriculum learning in output (task) space
  - Original task: false recall / malignant / negative
  - Created sub-task: false recall / (malignant + negative)

  - Transition between the two tasks: loss scheduler
    - A scheduled curriculum for the loss
    - dynamically weight the contribution of the two tasks to the loss



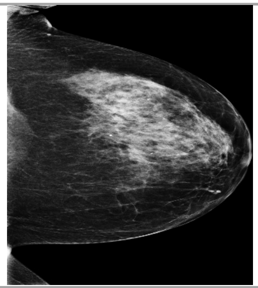
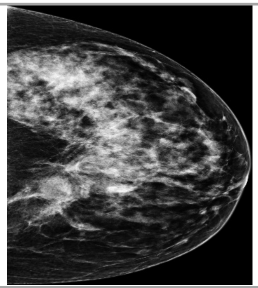
# Method

- Curriculum learning in output (task) space

  - Consider an image-label pair $(x_i, y_i)$
    - $x_i$ - image
    - $y_i \in \{0,1,2\}$ – label for the original task

  - Model $f(\cdot)$
    - $f(x_i)^{(c)}$ - the probability that the predicted label is $c \in \{0,1,2\}$

  - Original task (hard task) cross-entropy loss:
    - $\mathcal{L}_{hard}(x_i, y_i) = -\sum_c (y_{i,c} \cdot \log(f(x_i)^{(c)}))$

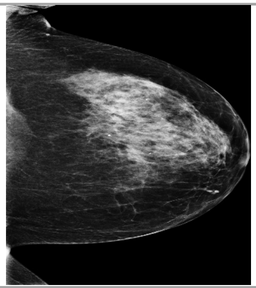
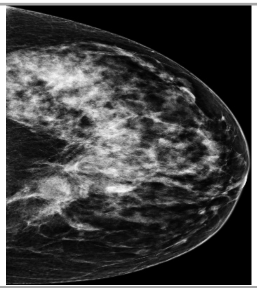
| | | | |
|---|---|---|---|
| $x_i$ |  |  |  |
| $y_i$ | 0 (false recall) | 1 (negative) | 2 (malignant) |

# Method

- Curriculum learning in output (task) space

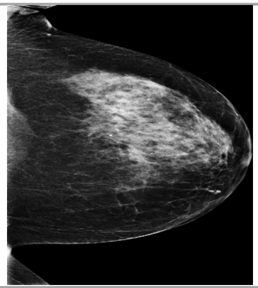
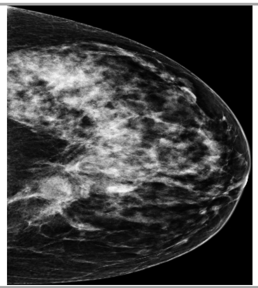
  - Consider an image-label pair $(x_i, y_i)$
    - $x_i$ - image
    - $y_i \in \{0,1,2\}$ – label for the original task

  - For the created “easier” sub-task
    - $z_i = [\![ y_i \neq 0 ]\!]$ - label for the “easier” task

| | | | |
|---|---|---|---|
| $x_i$ |  |  |  |
| $y_i$ | 0 (false recall) | 1 (negative) | 2 (malignant) |
| $z_i$ | 0 | 1 | 1 |

# Method

- Curriculum learning in output (task) space
  - The probabilities
    - $P(z_i = 0)$ still as $f(x_i)^{(0)}$
    - $P(z_i \neq 0)$ as $f(x_i)^{(1)} + f(x_i)^{(2)}$
      or $1 - f(x_i)^{(0)}$

| | | | |
|---|---|---|---|
| $x_i$ |  |  |  |
| $y_i$ | 0 (false recall) | 1 (negative) | 2 (malignant) |
| $z_i$ | 0 | 1 | 1 |

  - "Easier" task cross-entropy loss:
    - $\mathcal{L}_{easy}(x_i, z_i) = -(z_i \cdot \log\left(f(x_i)^{(0)}\right) + (1 - z_i) \cdot \log(1 - f(x_i)^{(0)}))$

# Method

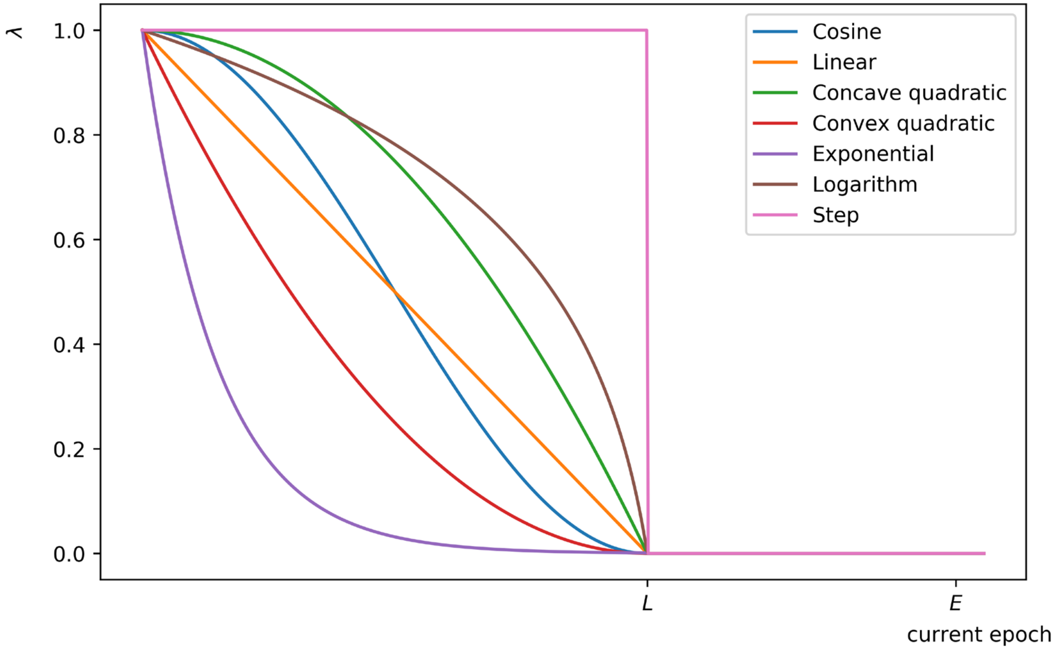
- Curriculum learning in output (task) space
  - Loss of the hard (original) task:
    - $\mathcal{L}_{hard}(x_i, y_i) = -\sum_c (y_{i,c} \cdot \log(f(x_i)^{(c)}))$
  - Loss of the easy task (created sub-task) loss:
    - $\mathcal{L}_{easy}(x_i, z_i) = -(z_i \cdot \log\left(f(x_i)^{(0)}\right) + (1 - z_i) \cdot \log(1 - f(x_i)^{(0)}))$
  - Final loss:
    - $\mathcal{L}(x_i, y_i, z_i) = \lambda \cdot \mathcal{L}_{easy}(x_i, z_i) + (1 - \lambda) \cdot \mathcal{L}_{hard}(x_i, y_i)$

# Method

- Loss scheduler $\lambda$
  - Overall loss:
    - $\mathcal{L}(x_i, y_i, z_i) = \lambda \cdot \mathcal{L}_{easy}(x_i, z_i) + (1 - \lambda) \cdot \mathcal{L}_{hard}(x_i, y_i)$
  - $\lambda$ is a function with respect to the epoch number
    - $\lambda = 1$ – focus completely on the easy task
    - $\lambda = 0$ – focus completely on the hard task
  - Desired $\lambda$ should transition from 1 to 0 as the training progresses.

# Method

- Loss scheduler $\lambda$
  - Here we investigate the following forms of the loss scheduler $\lambda$




| Loss scheduler type | Function (with $0 \leq e < L$) |
| --- | --- |
| Cosine | $\lambda(e) = (\cos(e\pi/L) + 1)/2$ |
| Linear | $\lambda(e) = 1 - e/L$ |
| Concave quadratic | $\lambda(e) = -(e/L)^2 + 1$ |
| Convex quadratic | $\lambda(e) = L^{-2} \cdot (e - L)^2$ |
| Exponential | $\lambda(e) = \epsilon^{e/L}, \ \epsilon = 10^{-3}$ |
| Logarithm | $\lambda(e) = \log(1 + L - e)/\log(1 + L)$ |
| Step | $\lambda(e) = 1$ |

SPIE. MEDICAL IMAGING

University of Pittsburgh

# Study cohort and dataset

- IRB-approved study
- Experiment over 1,709 FFDM images (349 Malignant cases, 653 Negative cases and 707 False recall cases)

- 5-fold cross-validation to make the results robust
  - 4 folds: 80% for training and 20% for validation
  - 1 fold: testing
- Table below are #images for each category.

| | Malignant | Negative | False recall | Total |
|---|---|---|---|---|
| Training set | 226 | 425 | 460 | 1,111 |
| Validation set | 53 | 98 | 106 | 280 |
| Test set | 45 | 130 | 141 | 341 |
| Total | 349 | 653 | 707 | 1,709 |

SPIE. MEDICAL IMAGING

University of Pittsburgh

# Experiments and results

- Experiment details
  - Model backbone: VGG16 [6]
  - Compared baseline: Aboutalib et al.'s [7] – a vanilla training strategy without using CL
  - Metrics:
    - Accuracy and AUC (averaged) for the original task
    - Balanced accuracy: averaging the three accuracies with respect to the three classes, reducing the effect induced by data imbalance
    - Accuracy and AUC for the created sub-task (binary)

  - Implementation: PyTorch
  - Hardware: Nvidia TESLA V100 from Pittsburgh Supercomputing Center

[6] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).
[7] Aboutalib, S. S., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., and Wu, S., "Deep learning to distinguish recalled but benign mammography images in breast cancer screening," Clinical Cancer Research 24(23), 5902–5909 (2018).

# Experiments and results

- Results
  - "LS: X" stands for a certain type of loss scheduler.
  - Results here are the means over five cross-validation partitions.

| | Accuracy | Balanced accuracy | Average AUC | Binary task accuracy | Binary task AUC |
|---|---|---|---|---|---|
| Aboutalib et al.[7] (baseline) | 0.489 | 0.458 | 0.658 | 0.598 | 0.633 |
| LS: exponential | 0.491 | 0.468 | 0.658 | 0.623 | 0.641 |
| LS: convex quadratic | 0.510 | 0.474 | **0.672** | 0.607 | 0.648 |
| LS: linear | 0.480 | 0.476 | 0.655 | 0.614 | 0.635 |
| LS: cosine | 0.501 | 0.464 | 0.653 | 0.611 | 0.645 |
| LS: concave quadratic | 0.508 | 0.492 | 0.671 | **0.633** | 0.647 |
| LS: logarithm | 0.511 | **0.494** | 0.668 | 0.617 | 0.644 |
| LS: step | **0.515** | 0.483 | 0.669 | 0.622 | **0.655** |

[7] Aboutalib, S. S., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., and Wu, S., "Deep learning to distinguish recalled but benign mammography images in breast cancer screening," Clinical Cancer Research 24(23), 5902–5909 (2018).

# Discussions

- We design a novel deep curriculum learning in task space strategy for training a model on the task to classify False Recall, Malignant, and Negative FFDM images.
  - We treat the original task (false recall/malignant/negative) as a “hard” task in terms of CL
  - An “easier” sub-task is created by grouping Malignant, and Negative against the False Recall. By focusing on the difficulty of the tasks, our training strategy implicitly utilizes the medical knowledge as a supplement for the data-driven deep learning.
  - We introduce a loss scheduler as a curriculum to transition from the “easier” task to the “harder” (original) task.

# Discussions

- Results show that our CL in task space strategy outperforms the baseline without the CL strategy.
  - loss schedulers make it possible to focus more on the “easier” task and transition to the original “harder” task of the 3-class classification .
  - In the task of classifying False Recall, Malignant, and Negative FFDM images, properly focusing on classifying False Recall against the combination of Malignant and Negative images boosted the results of the original task.

# Discussions

- Limitations and future works
  - The dataset
    - Single-center study. Need a multi-center and larger dataset for further evaluation of our method.
  - The loss scheduler
    - enables the transition from the “easy” task to “hard” task, but introduced two hyperparameters (function $\lambda(\cdot)$ and the epoch number $L$)
  - Creating the sub-task
    - We examined one option of grouping the classes into a binary task. We plan to try other groupings and CL strategies in the future.

# Conclusions

- Our work is a novel study that develops a learning strategy about curriculum learning in task space on FFDM image classification.
- Our method shows that by scheduling the learning from an “easier” sub-task to the original “harder” task, the model’s performance to classify the FFDM images can be boosted.
- Further evaluation and experiments with different grouping of sub-tasks are needed, and will further examine the effects of our proposed curriculum learning methods.

# Acknowledgements:

## **I**ntelligent **C**omputing for **C**linical **I**maging (**ICCI**) Lab, University of Pittsburgh


- ❖ NIH/NCI #1R01CA218405, #1R01EB032896 as part of the NSF/NIH Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science Program
- ❖ Pilot Research Project of the Pittsburgh Center for AI Innovation in Medical Imaging and the associated Pitt Momentum Funds of a Scaling grant from the University of Pittsburgh (2020)
- ❖ Amazon Machine Learning Research Award.
- ❖ Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF #ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award #ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

# Thank you!

Jun Luo
jul117@pitt.edu