Information Gain Sampling for Active Learning in Medical Image Classification




Raghav Mehta, Changjian Shui, Brennan Nichyporuk, Tal Arbel

(UNSURE 2022: MICCAI 2022)

Medical Imaging and Deep Learning

- **Problem: Acquiring annotations** for data used to train deep learning models is **time-consuming**

Medical Imaging and Deep Learning

- **Problem: Acquiring annotations** for data used to train deep learning models is **time-consuming**
- **Solution: Active Learning** methods to select most useful data for annotation

Active Learning framework

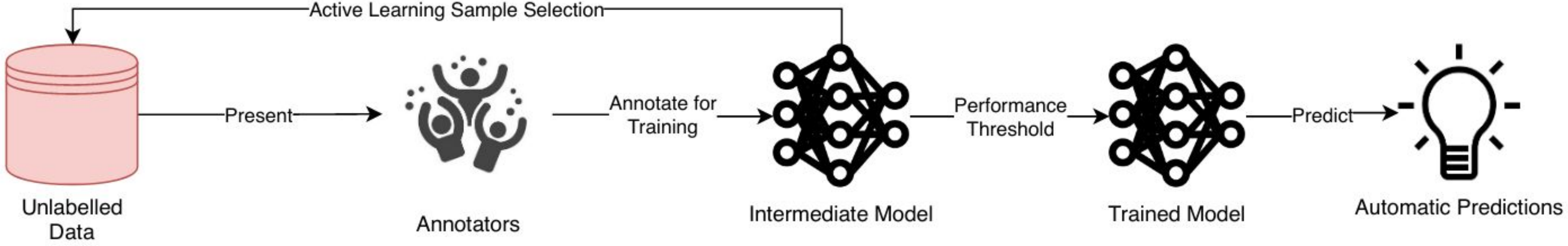


Image Courtesy: [7]

Active Learning framework

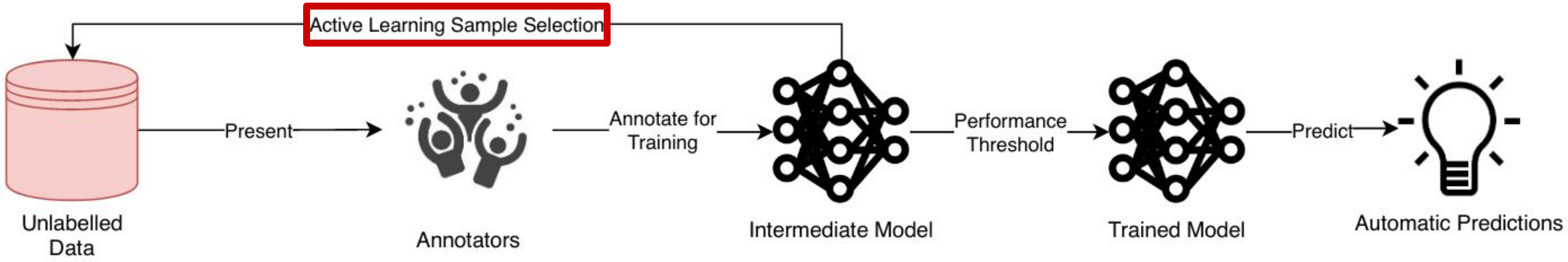


Image Courtesy: [7]

Active Learning Sample Selection

- **Uncertainty Based**
    - Least confidence [8]
    - Maximum entropy [9]
    - Smallest margin [10]
    - Minimum expected generalization loss [11]
    - Deep Bayesian active learning [12]
- Representation Based
    - CoreSet [13]
    - Variational Adversarial Active Learning [14]
    - Reinforcement Learning [15]
- Combination of both Uncertainty and Representation based [16,17]

Uncertainty Based Sample Selection

- **Pros**
  - Indicates **the samples** which are **hardest** for the current model to classify
  - **Useful in medical imaging context** where there is a high class imbalance

- **Cons**
  - Doesn’t convey the **source of uncertainty**
    - Classes that are source of confusion
  - No information about **how** the addition of the **sample’s labels** will **influence** the **downstream performance**

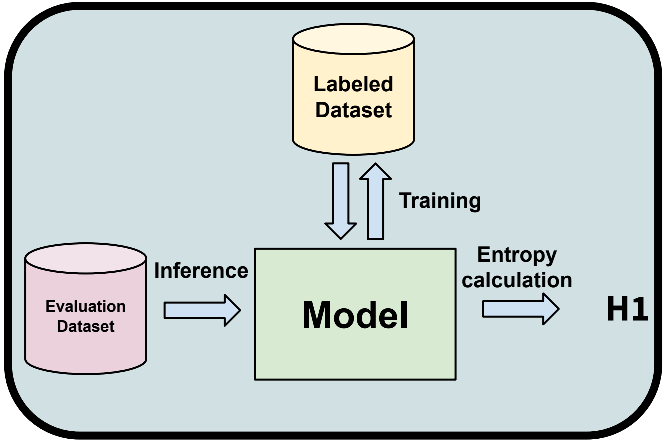
Information Gain Sampling for AL

- **Expected Information Gain (EIG)**
  - **EIG (X; Y)** measures the **reduction in the entropy** H of a random variable, X, by observing the state of another random variable, Y

- **In active learning context**, EIG measure the **reduction in the entropy** of the predicted labels of **the evaluation set**, if we have access to the **true state of an image in the unlabeled set**

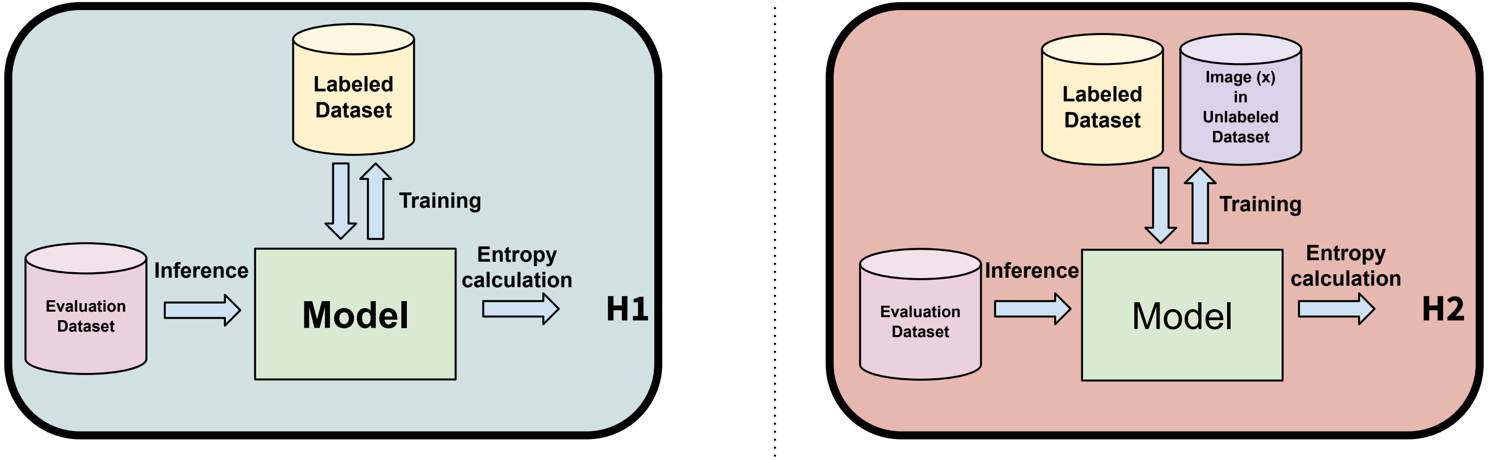
Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**,



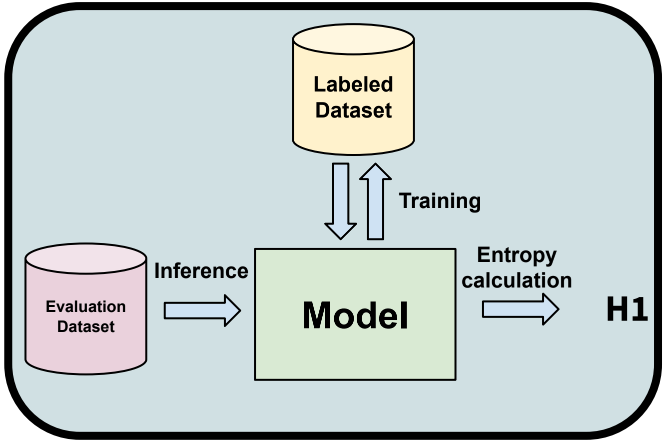
Information Gain Sampling for AL

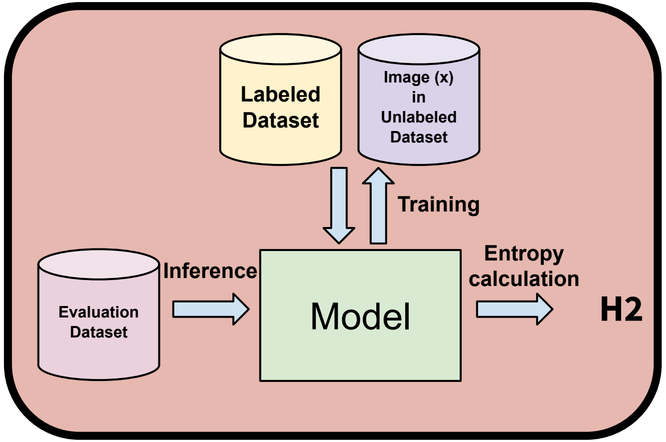
- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**



Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**



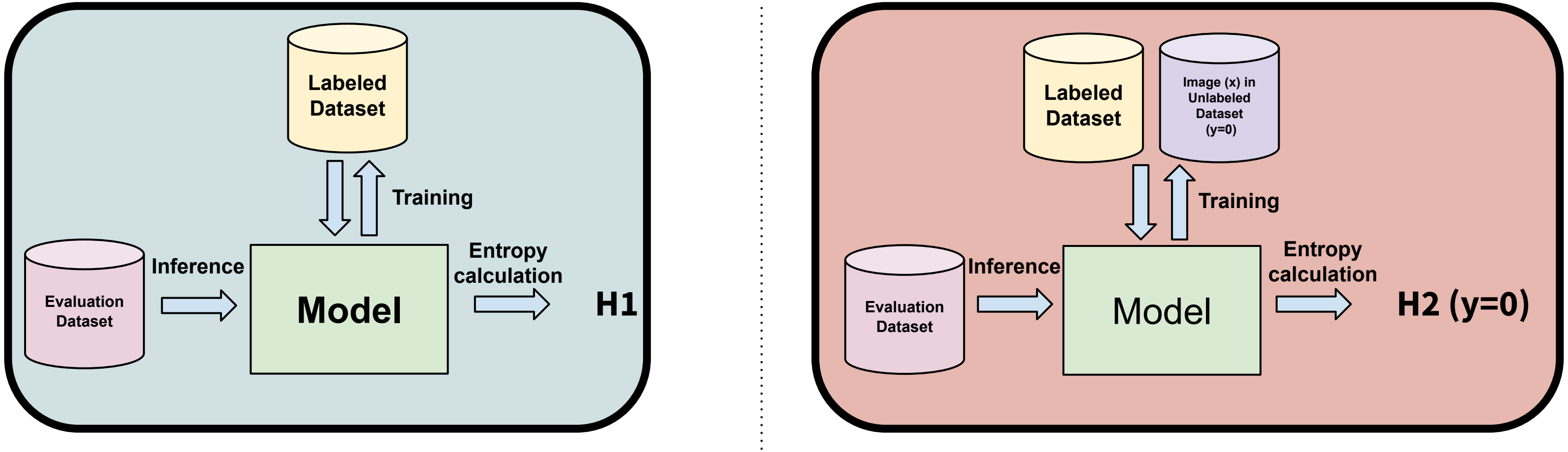


But we don't have actual label for the image in the unlabeled set

Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**

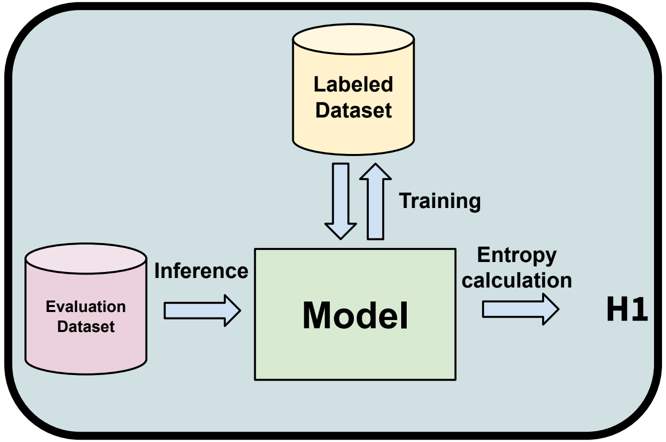
Simulate this by assuming all possible labels for the image

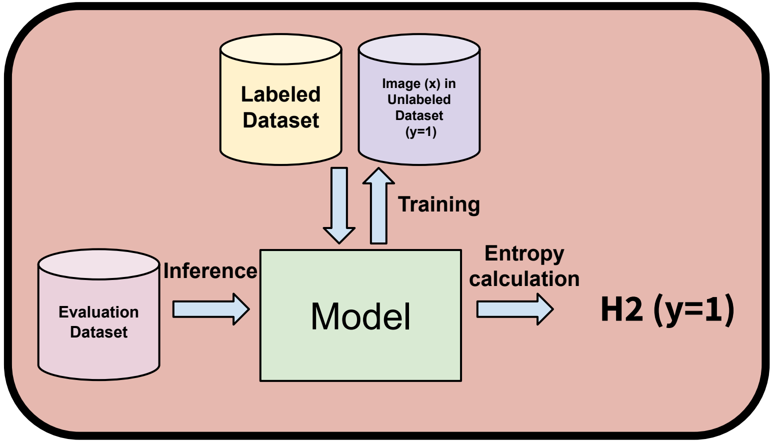


Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**

Simulate this by assuming all possible labels for the image

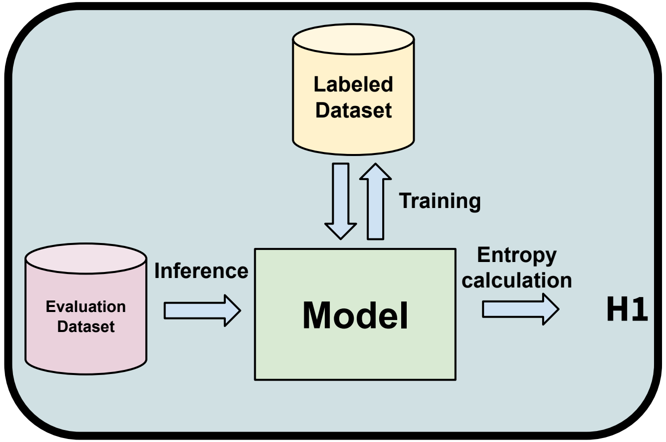


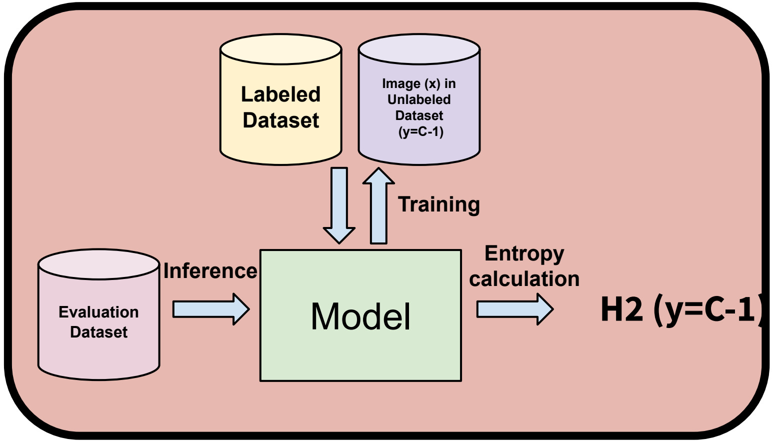


Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**

Simulate this by assuming all possible labels for the image

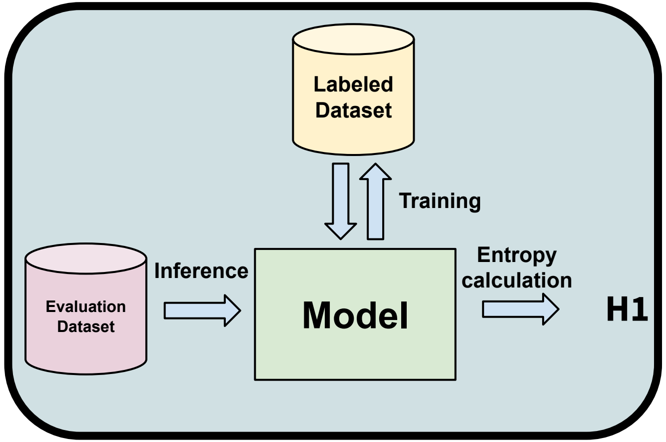


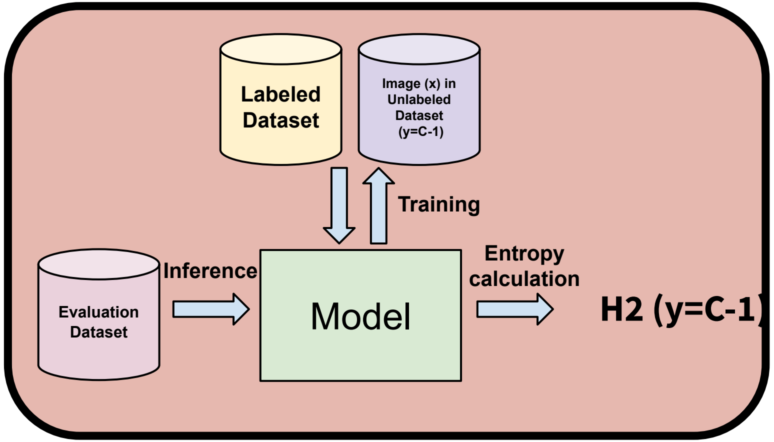


Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**

Calculate EIG for image x





EIG (x) = **H1 - Σ_y p(Y=y) H2(Y=y)**

Information Gain Sampling for AL

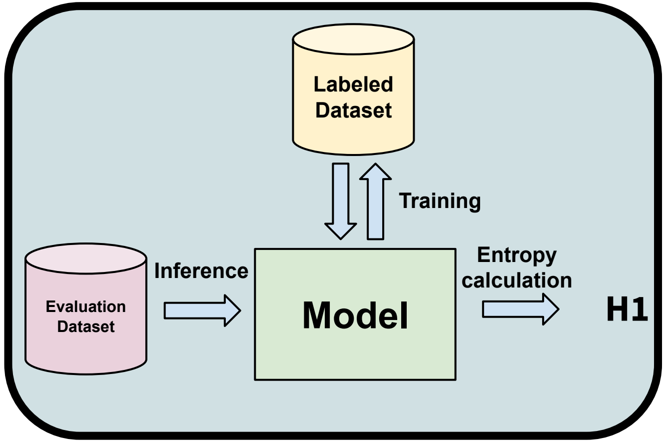
- **Practical Consideration**
  - EIG calculation requires model update for **each image (N) in the unlabeled set** and for **each possible labels (C)**
  - Total **N*C model** updates
  - Calculation of evaluation set entropy after each of these updates
  - **Too much computational overhead**

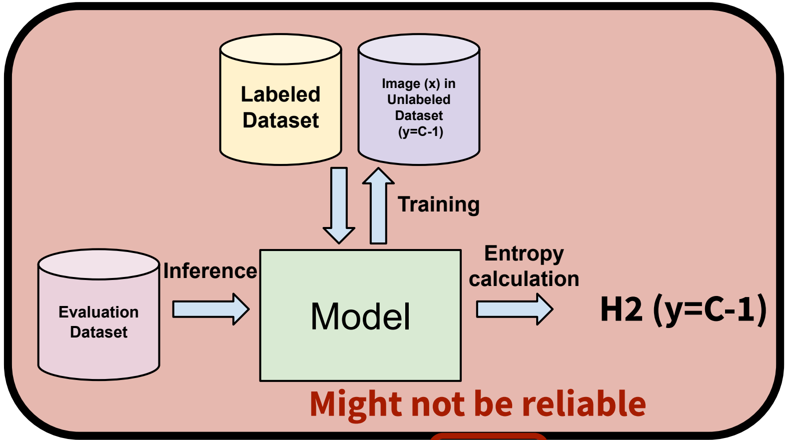
Information Gain Sampling for AL

- **Practical Consideration**
  - EIG calculation requires model update for **each image (N) in the unlabeled set** and for **each possible labels (C)**
  - Total **N*C model** updates
  - Calculation of evaluation set entropy after each of these updates
  - **Too much computational overhead**
- **Design choices**
  - Model update using **only a single gradient step**
  - Deep Learning models have two parts
    - Convolutional Layer: feature extraction
    - Multi-Layer Perceptron: classification
  - **Only update classification layer** parameters during EIG calculation

Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**

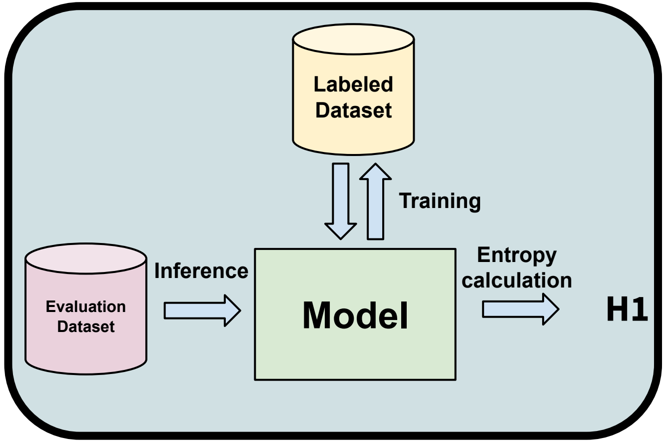


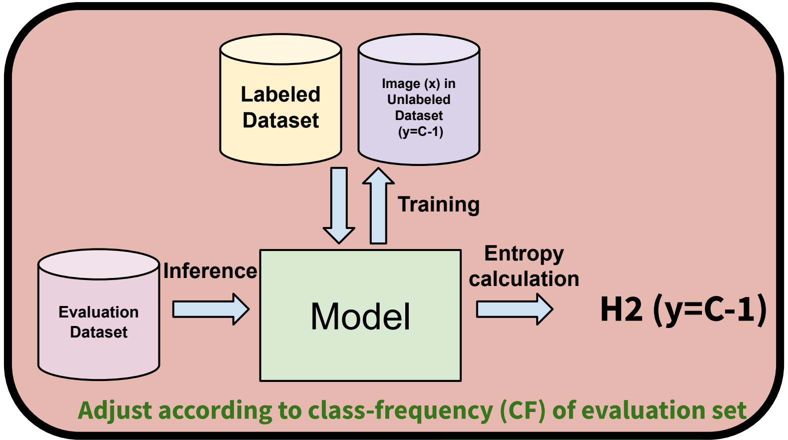


$$\text{EIG}(x) = \text{H1} - \Sigma_y \boxed{p(Y=y)}\, \text{H2}(Y=y)$$

Information Gain Sampling for AL

- In short, EIG **measures difference** in the entropy for two models. (i) **H1:** the entropy for a **model trained on the labeled set**, and (ii) **H2**: the conditional entropy of for a **model trained on the labeled set and an image in the unlabeled set**

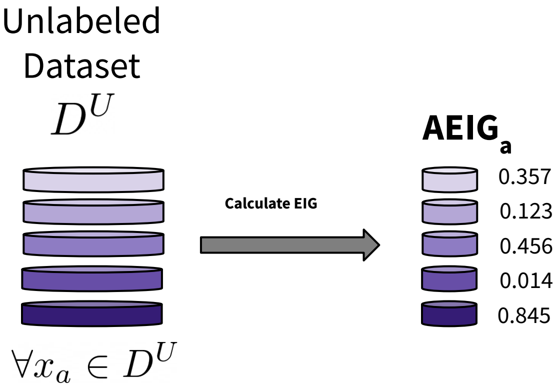




$$\text{AEIG }(x) = \text{H1} - \Sigma_y \boxed{p(Y=y) * \text{CF\_eval}}\ \text{H2}(Y=y)$$

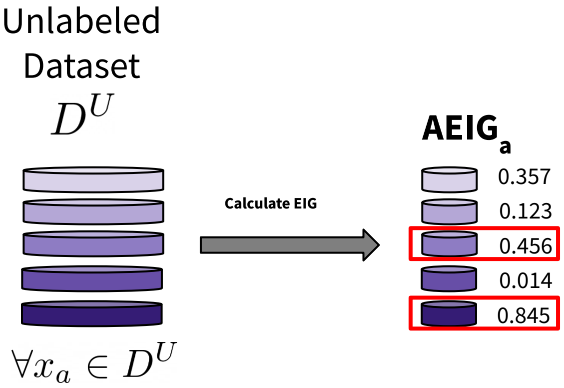
Information Gain Sampling for AL

- **Calculate AEIG for all images in the unlabeled dataset**



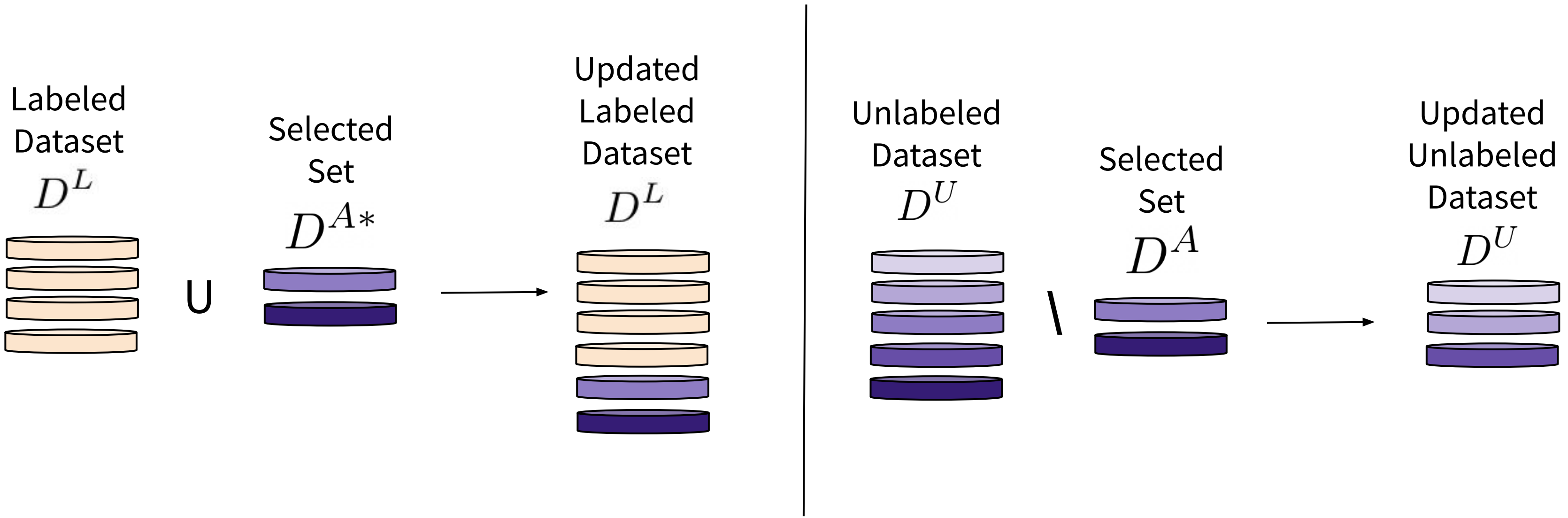
Information Gain Sampling for AL

- **Select top-B images from the unlabeled dataset:** $D^A : \{x_a\}_{a=0}^{B}$



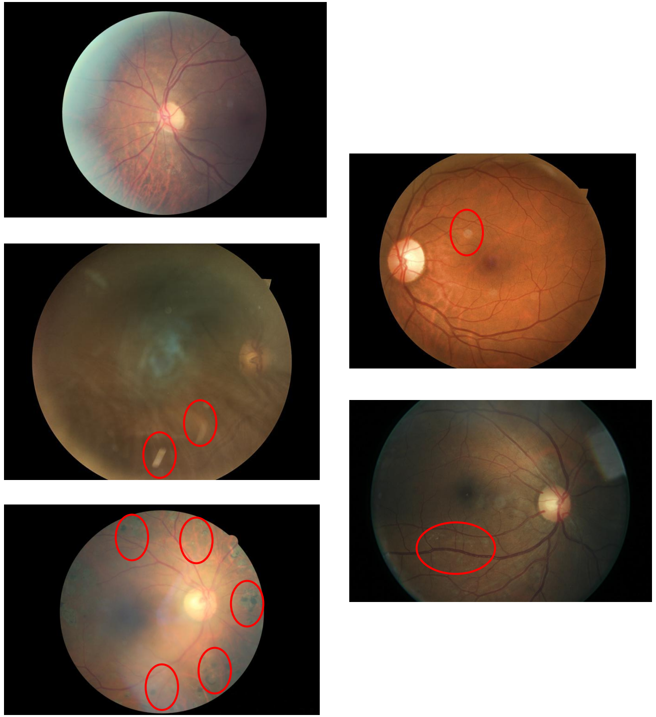
Information Gain Sampling for AL

- **Update both the labeled and unlabeled datasets**



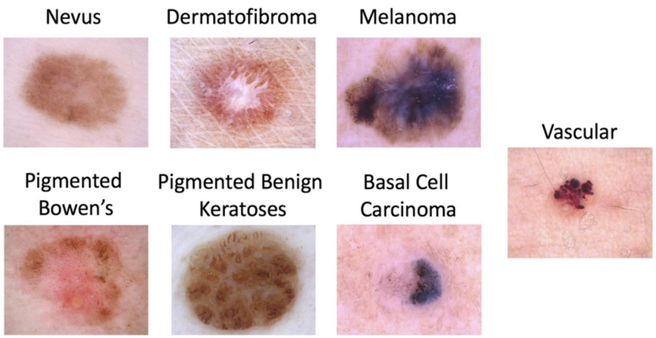
Experiments and Results

- **<u>Datasets:</u>**
  - Multi-class Diabetic Retinopathy (DR) disease classification



Experiments and Results

- **Datasets:**
  - Multi-class Diabetic Retinopathy (DR) disease classification
  - Multi-class skin lesion classification (ISIC)



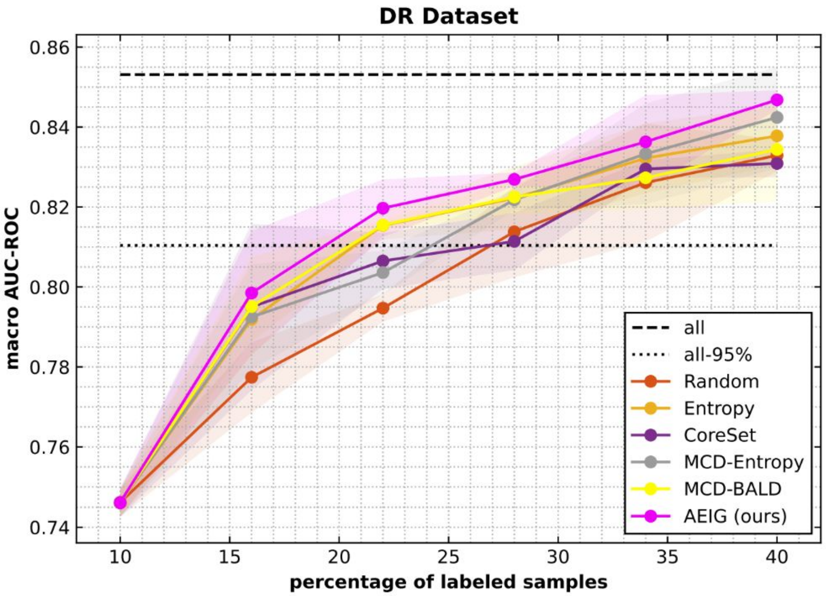
Experiments and Results

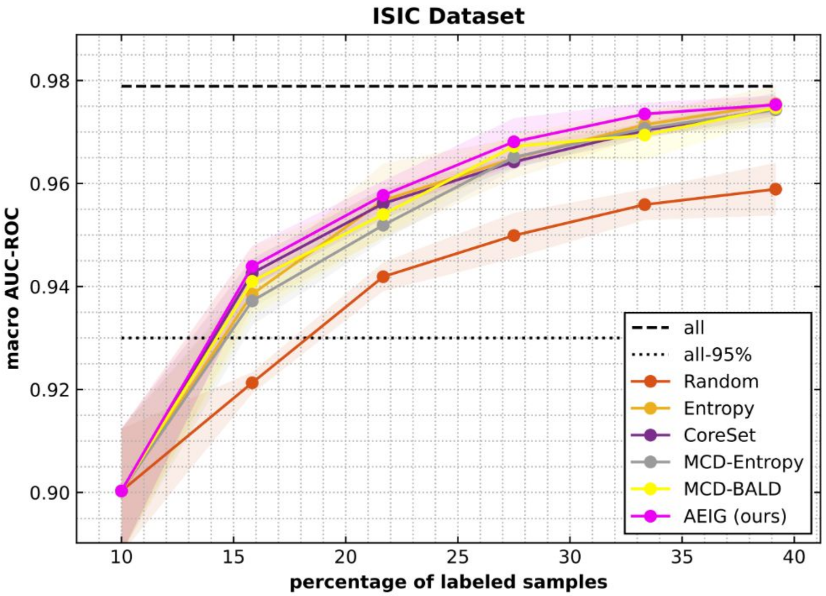
- **Datasets:**
  - Multi-class Diabetic Retinopathy (DR) disease classification
  - Multi-class skin lesion classification (ISIC)

- **Evaluation Metric:**
  - ‘macro’ Area Under the Receiver Operating Characteristic Curve (ROC AUC)

Experiments and Results

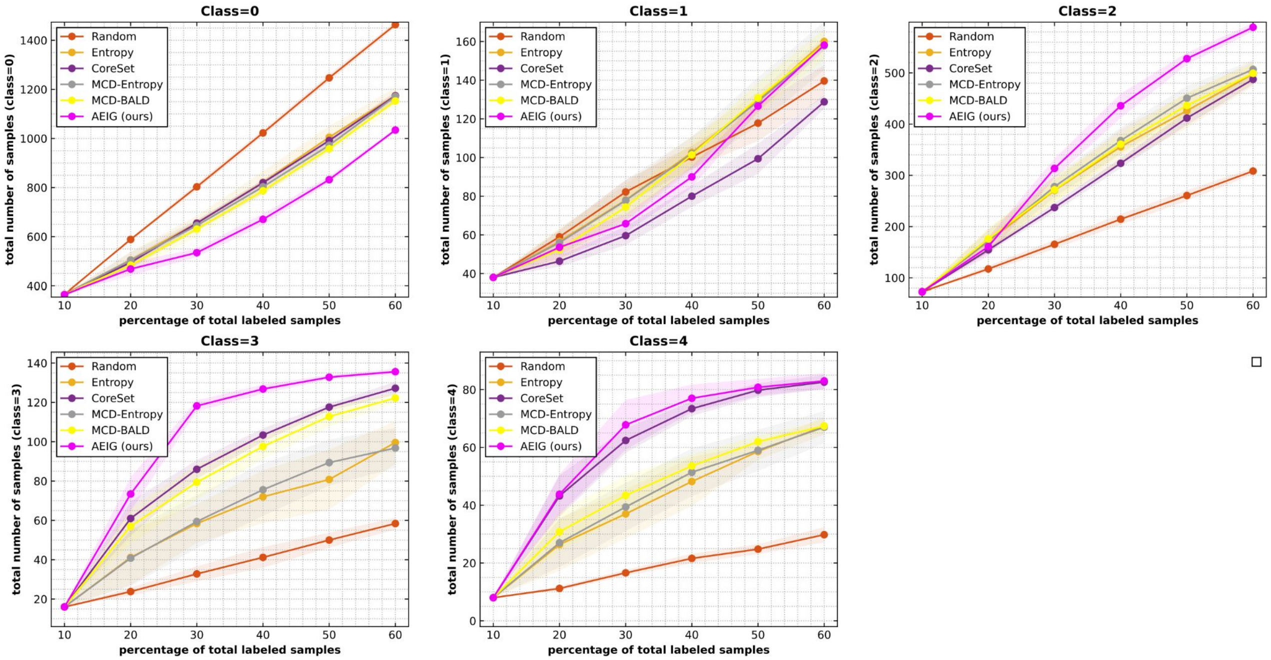
- **Results**





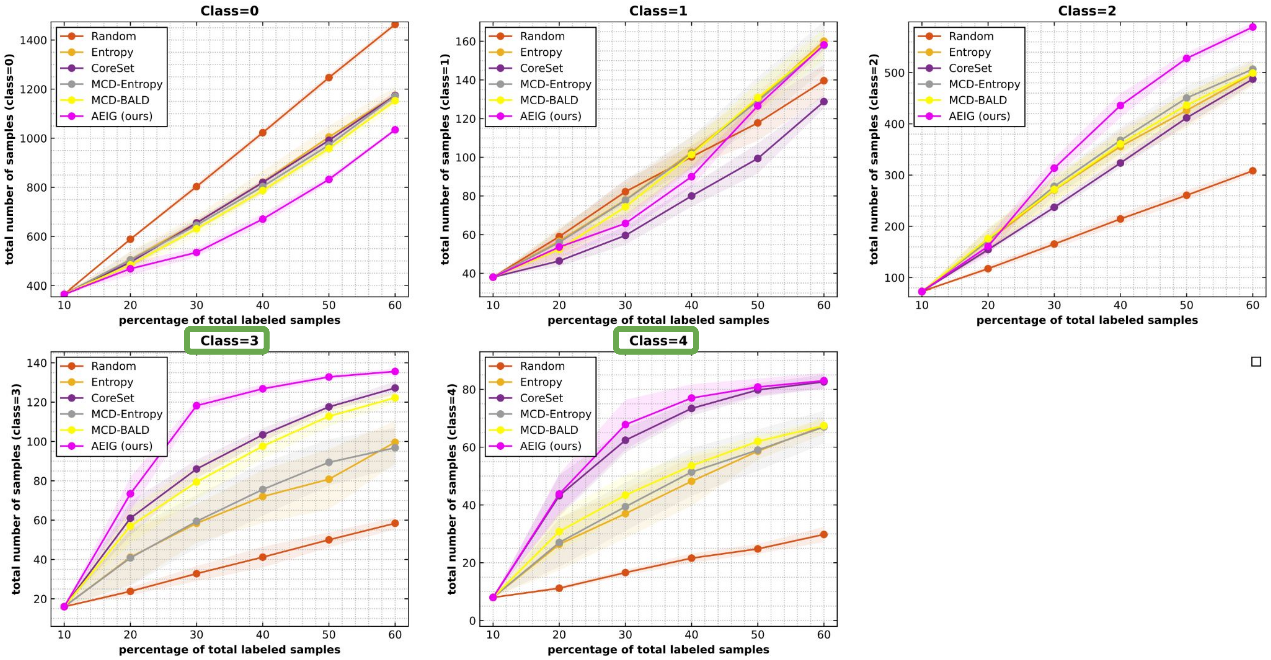
Experiments and Results

- **why AEIG works better? - DR**



Experiments and Results

- **why AEIG works better? - DR**



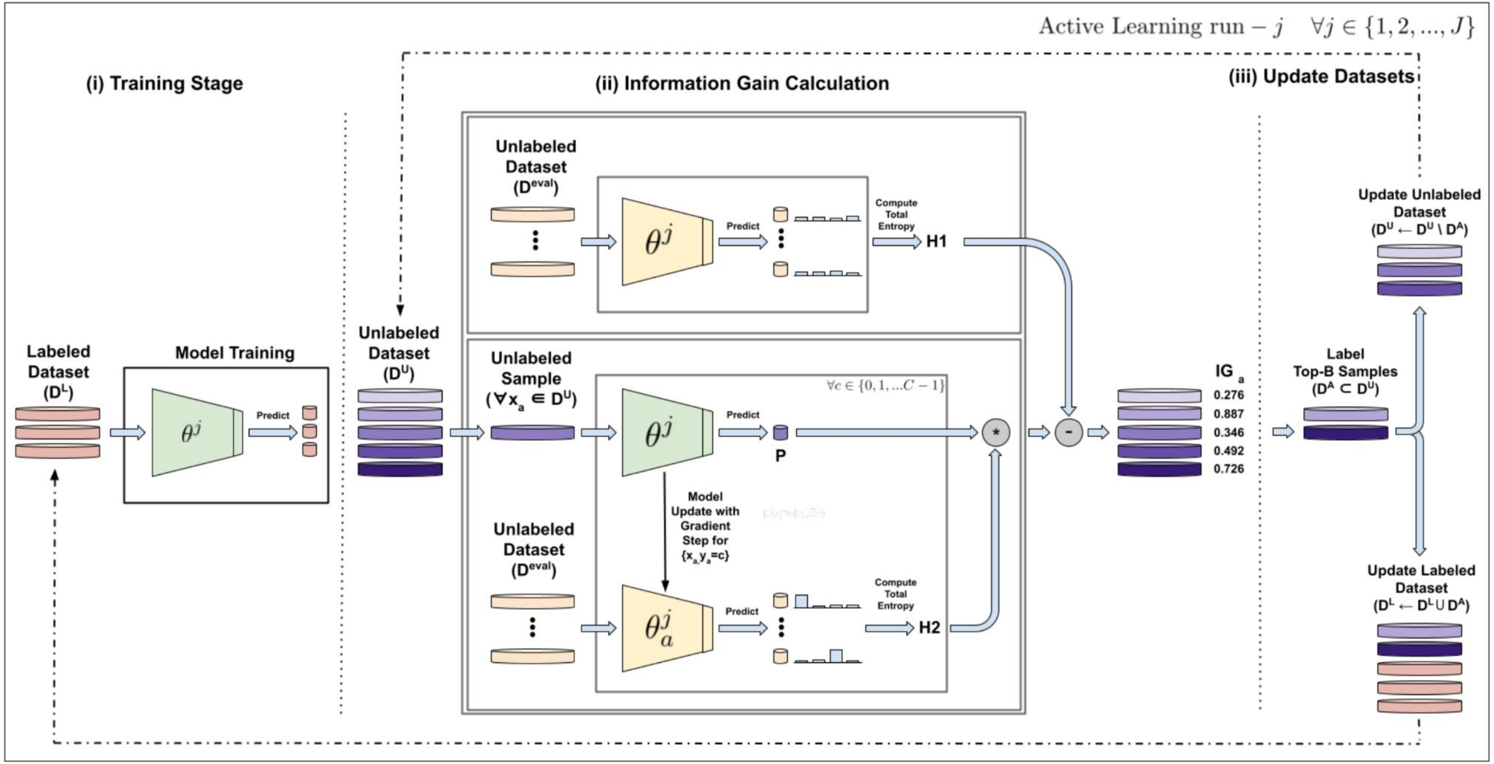
CIM CENTRE FOR INTELLIGENT MACHINES

Conclusions

- Proposed **an information theoretic** active learning **samples selection approach**
- With **careful design choices**, method can be easily integrated into **existing deep learning classifiers**
- The proposed method achieves 95% of overall performance with **only 19%** of the training data
  - Random: 34%
  - Maximum entropy: 23%
  - CoreSet: 22%
- The proposed method **selects more samples from the least representative classes**
  - Useful for medical imaging context with high class imbalance

Thank You

Information Gain Sampling for AL



Information Gain Sampling for AL

- **Expected Information Gain (EIG)**

$$\mathrm{EIG}(Y^{\mathrm{eval}}; y_a | x_a, X^{\mathrm{eval}}, D^L)$$

$$= \mathbf{H}[Y^{\mathrm{eval}} | X^{\mathrm{eval}}, D^L] - \mathbf{H}[Y^{\mathrm{eval}} | y_a, x_a, X^{\mathrm{eval}}, D^L]$$

$$= \underbrace{\mathbf{H}[Y^{\mathrm{eval}} | X^{\mathrm{eval}}, D^L]}_{\mathbf{H1}} - \sum_{c=0}^{C-1} p(y_a = c | x_a, D^L) \underbrace{\mathbf{H}[Y^{\mathrm{eval}} | y_a = c, x_a, X^{\mathrm{eval}}, D^L]}_{\mathbf{H2}}$$

$$= \underbrace{\sum_{j=0}^{K} \mathbf{H}[y_j^{\mathrm{eval}} | x_j^{\mathrm{eval}}, D^L]}_{\mathbf{H1}} - \sum_{c=0}^{C-1} p(y_a = c | x_a, D^L) \underbrace{\left(\sum_{j=0}^{K} \mathbf{H}[y_j^{\mathrm{eval}} | y_a = c, x_a, x_j^{\mathrm{eval}}, D^L] \right)}_{\mathbf{H2}}$$

Information Gain Sampling for AL

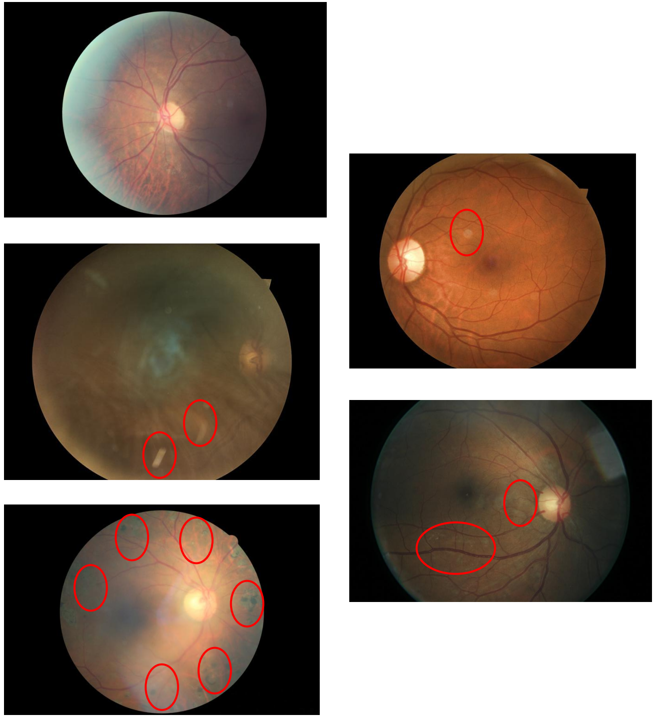
- **Adjusted Expected Information Gain (AEIG)**

$$\mathrm{AEIG}(Y^{\mathrm{eval}}; y_a | x_a, X^{\mathrm{eval}}, D^L) = \mathbf{H1} - \underbrace{p(y_a = c | x_a, D^L) \frac{|y_{\mathrm{eval}} = c|}{\sum_{j=0}^{C-1} |y_{\mathrm{eval}} = j|}}_{\mathbf{P}} \mathbf{H2}$$

- The predicted softmax probability (P) of the training model is adjusted with the class frequencies of the evaluation set

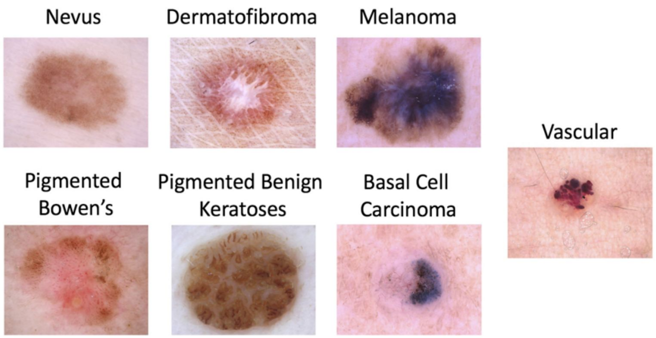
Medical Image Disease Classification

- **Diabetic Retinopathy disease classification**
  - Multi-class classification dataset
  - Classify Colour fundus images into five stages
    - 0 - No DR
    - 1 - Mild
    - 2 - Moderate
    - 3 - Severe
    - 4 - Proliferative
  - Dataset
    - Kaggle challenge dataset
    - a subset of 8408 retinal fundus images
      - randomly divide the whole dataset into 5000/1000/2408 images for training/validation/testing sets



Medical Image Disease Classification

- **ISIC skin lesion classification**
  - Multi-class classification dataset
  - Classify dermoscopic images into seven types
  - Dataset
    - ISIC 2018 dataset
    - a subset of 10015 dermoscopic images
      - randomly divide the whole dataset into 6000/1500/2515 images for training/validation/testing sets



Implementation details

- **AL framework**
  - Total Active Learning runs (J): 6
  - Labeled Set (D^L): 600 for ISIC, 500 for DR
  - Unlabeled Set (D^U): 5400 for ISIC, 4500 for DR
  - Selected Set (D^A): 350 for ISIC, 300 for DR
  - 5 repetition for both dataset
- **Evaluation Metric**
  - 'macro' Area Under the Receiver Operating Characteristic Curve (ROC AUC)
    - For multi-class DR classification, macro average (unweighted) one-vs-rest (ovr) classifier ROC AUC

Experiments and Results

- **Comparison of EIG, AEIG, and other variants**

$$\mathrm{EIG}(Y^{\mathrm{eval}}; y_a | x_a, X^{\mathrm{eval}}, D^L) = \underbrace{\mathbf{H}[Y^{\mathrm{eval}} | X^{\mathrm{eval}}, D^L]}_{\mathbf{H1}} - \sum_{c=0}^{C-1} p(y_a = c | x_a, D^L)\, \underbrace{\mathbf{H}[Y^{\mathrm{eval}} | y_a = c, x_a, X^{\mathrm{eval}}, D^L]}_{\mathbf{H2}}$$

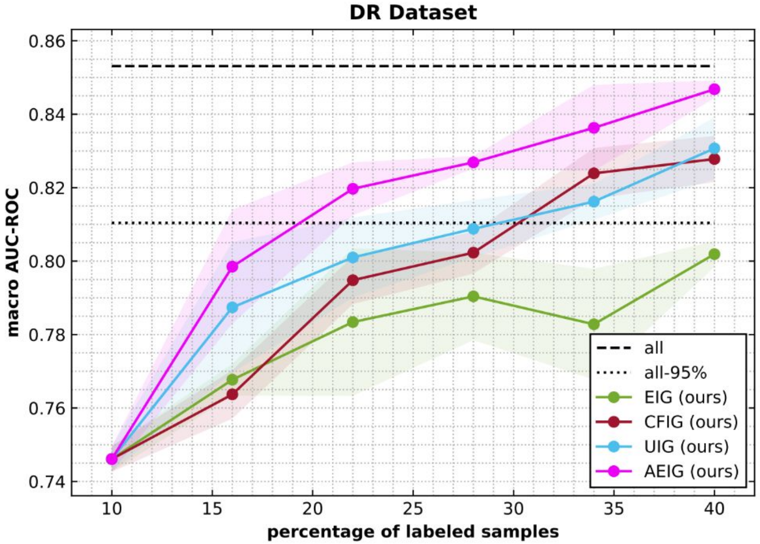
$$\mathrm{UIG}(Y^{\mathrm{eval}}; y_a | x_a, X^{\mathrm{eval}}, D^L) = \underbrace{\mathbf{H}[Y^{\mathrm{eval}} | X^{\mathrm{eval}}, D^L]}_{\mathbf{H1}} - \sum_{c=0}^{C-1} \frac{1}{C}\, \underbrace{\mathbf{H}[Y^{\mathrm{eval}} | y_a = c, x_a, X^{\mathrm{eval}}, D^L]}_{\mathbf{H2}}$$

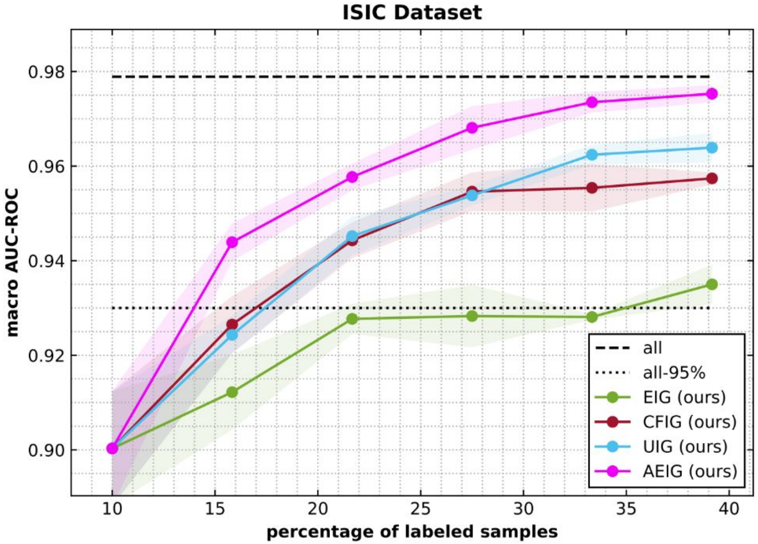
$$\mathrm{AEIG}(Y^{\mathrm{eval}}; y_a | x_a, X^{\mathrm{eval}}, D^L) = \mathbf{H1} - p(y_a = c | x_a, D^L) \frac{|y_{\mathrm{eval}} = c|}{\sum_{j=0}^{C-1} |y_{\mathrm{eval}} = j|} \mathbf{H2}$$

$$\mathrm{CFIG}(Y^{\mathrm{eval}}; y_a | x_a, X^{\mathrm{eval}}, D^L) = \mathbf{H1} - \frac{|y_{\mathrm{eval}} = c|}{\sum_{j=0}^{C-1} |y_{\mathrm{eval}} = j|} \mathbf{H2}$$

Experiments and Results

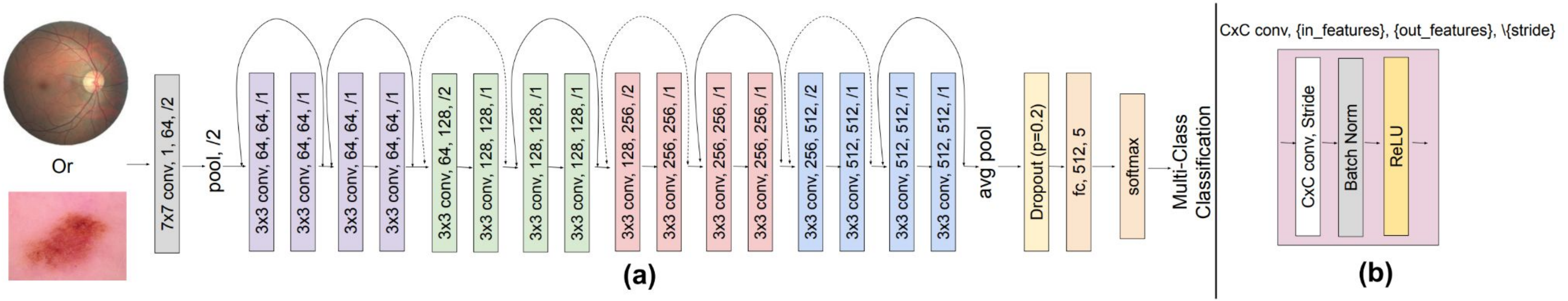
- **Comparison of EIG, AEIG, and other variants**





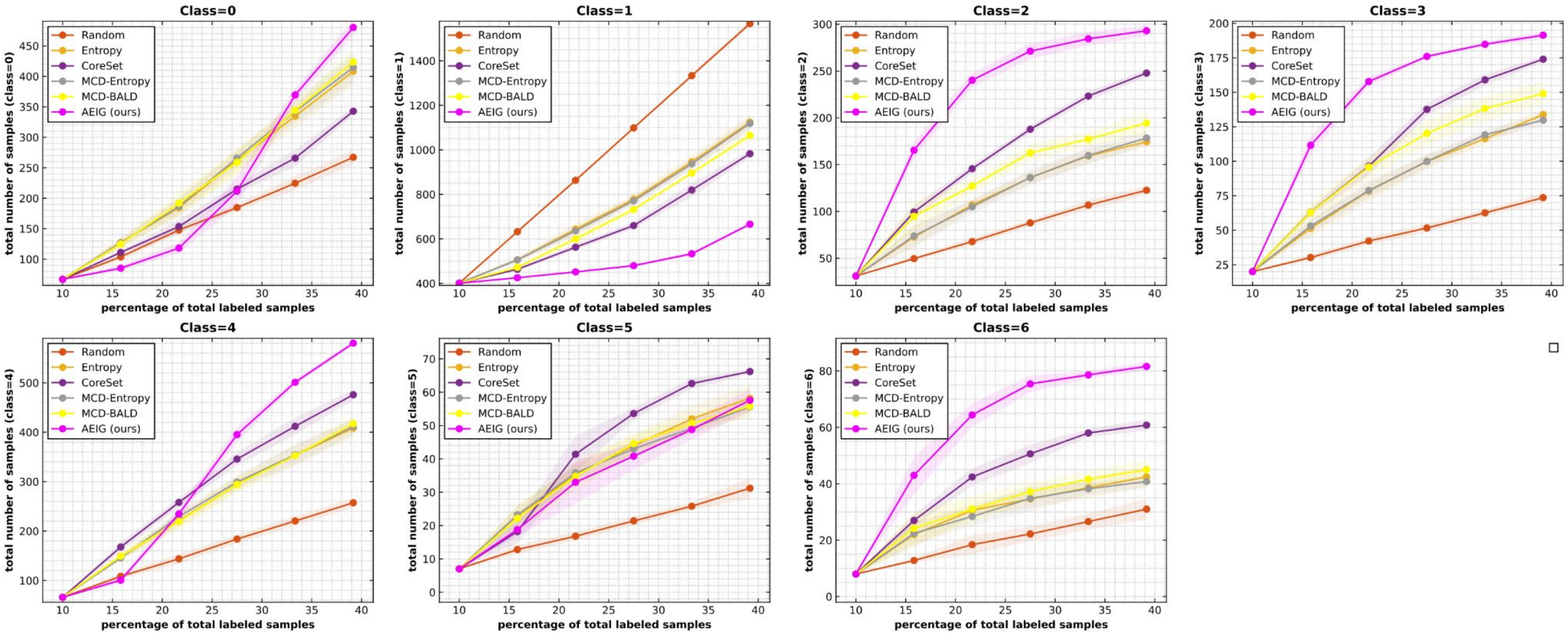
Implementation details

- **Diabetic Retinopathy disease classification**



Experiments and Results

- **Comparisons Against Active Learning Baselines**
  - **Insights: why AEIG works better? - ISIC**



Algorithm

Algorithm 1 Expected Information Gain Based Active Learning

Input: Labeled training dataset D^L : $\{(x_i, y_i^{c\in\{0,1,..,C-1\}})\}_{i=1}^{M}$, an unlabeled dataset D^U : $\{(x_i)\}_{i=1}^{N}$, and an evaluation (validation) dataset D^{valid}

Require: initial machine model (with parameters θ^0) trained on labeled dataset D^L, total active learning iterations J, and active learning batch acquisition size B

```
j ← 1
while active learning iteration j < J do

    Calculate H[Y^valid | X^valid, D^L] based on the model parameters θ^(j−1)

    for each image x_a ∈ D^U do
        Calculate p(y_a = c | x_a, D^L) based on the model parameters θ^(j−1)
        θ_i^(j−1) ← θ^(j−1)

        for each possible class label c ∈ {0, 1, .., C} do
            Using a single gradient step update model parameters (θ_a^(j−1)) with x_a and y_a = c
            Calculate H[Y^valid | X^valid, x_a, y_a = c, D^L]
        end for

        Compute Score S based on EIG according to Equation [1]
    end for

    Select subset of top-B images (D^A) from D^U according to their score S
    Acquire ground-truth labels for D^A ((D^A*))
    Update Unlabeled dataset D^U ← D^U \ D^A
    Update Labeled dataset D^L ← D^L ∪ D^A*
    Retrain the model (θ^j) with the updated labeled training dataset D^L
    j ← j + 1

end while
```