Based on the Expanded Version of our Paper which received the Best Paper Award at IEEE International Conference on Machine Learning and Applications, December 2018

SCREEN-DR: O-MedAL: Online Active Deep Learning for Diabetic Retinopathy Detection

Asim Smailagic, Aurelio Campilho, Alex Gaudio, Pedro Costa,
Kartik Khandelwal, Mostafa Mirshekari, Jonathan Fagert, Devesh Walawalkar, Susu, Xu, Adrian Galdran, Pei Zhang, Hae Young Noh

Objective

Deep Learning models typically require labeled data, but labeling data is often costly.

The goal is to develop an advanced machine learning framework including online active deep learning for diabetic retinopathy detection.

Contributions:

- ***Novelty***: we present a novel AL sampling method that queries the unlabeled examples maximally distant to the centroid of all training set examples in a learned feature space.
- ***Accurate***: we introduce an online training technique that is more accurate than the original baseline model trained on a fully labeled dataset.
- ***Efficient***: our method achieves better results with fewer labeled examples than competing methods.
- ***Computationally Efficient***: the online training technique processes fewer examples than the original baseline model trained on the fully labeled training set.
- ***Robust***: we test our active learning method on binary and multi-class classification problems using balanced and unbalanced datasets.

Active Learning

Assumption:

- It is costly and time consuming to ask a physician to label a retinal fundus image.
- Therefore, we use a fully labeled dataset and pretend it is unlabeled.

Algorithm:

1) Start with an unlabeled dataset
2) Selectively label data
3) Pass the labeled data to a model training method

** Our method builds on uncertainty sampling and outputs a stream of newly labeled data.

Online Learning

Assumption:

- Data becomes available sequentially
- Incrementally train the model with incoming data

Algorithm:

While a stream of labeled data exists:

1) Receive newly labeled training data
2) Incrementally train model using newly labeled data and a subset of previously labeled data

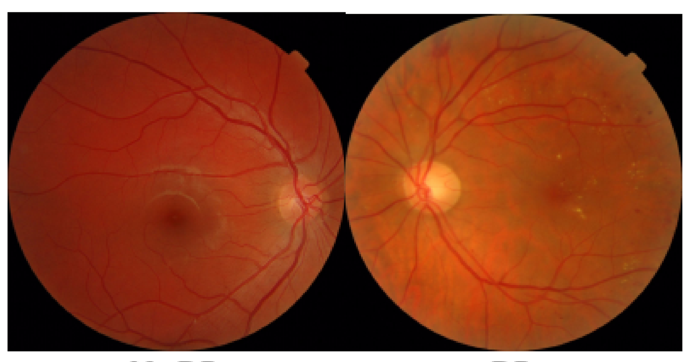
Medical Image Data

Assumption:

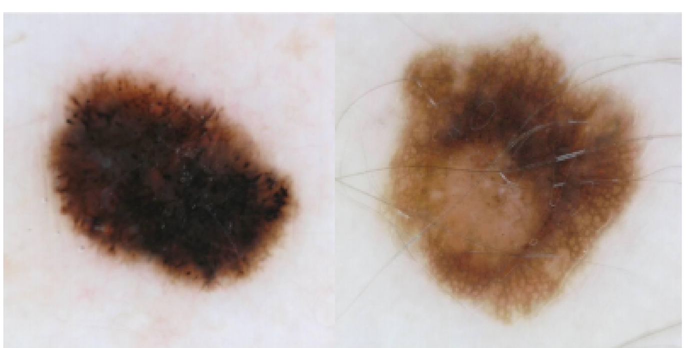
- Small labeled datasets and high resolution images
- Expensive to acquire and label

Datasets:

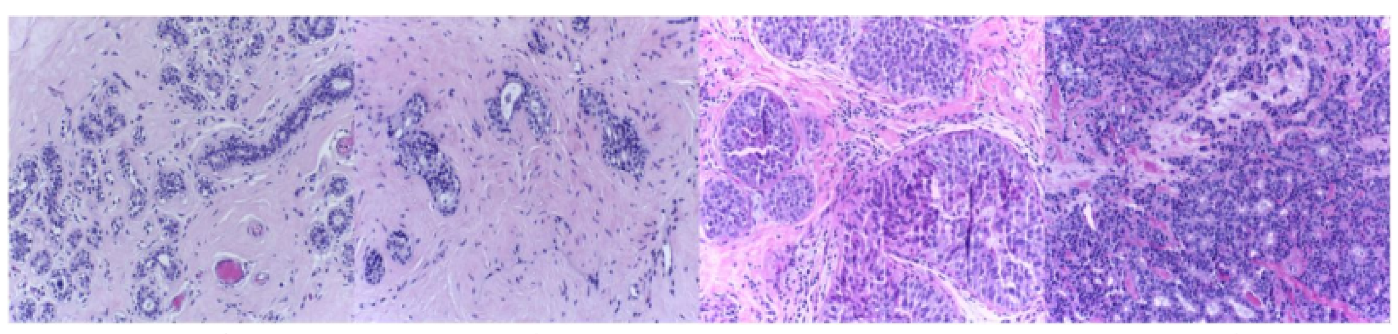
- **Messidor**: Balanced binary classification task. 1187 retinal fundus images to detect *Diabetic Retinopathy* in human eyes.
- **Skin Cancer**: Unbalanced binary classification task with 900 images.
- **Breast Cancer Diagnosis**: Balanced multi-class classification task. 400 images of breast tissue.



No DR DR



Benign Malignant



Normal Benign *In situ* Invasive

Proposed Method

Online Learning: train the model incrementally with available data

Active Learning: use the model to sample new training data points

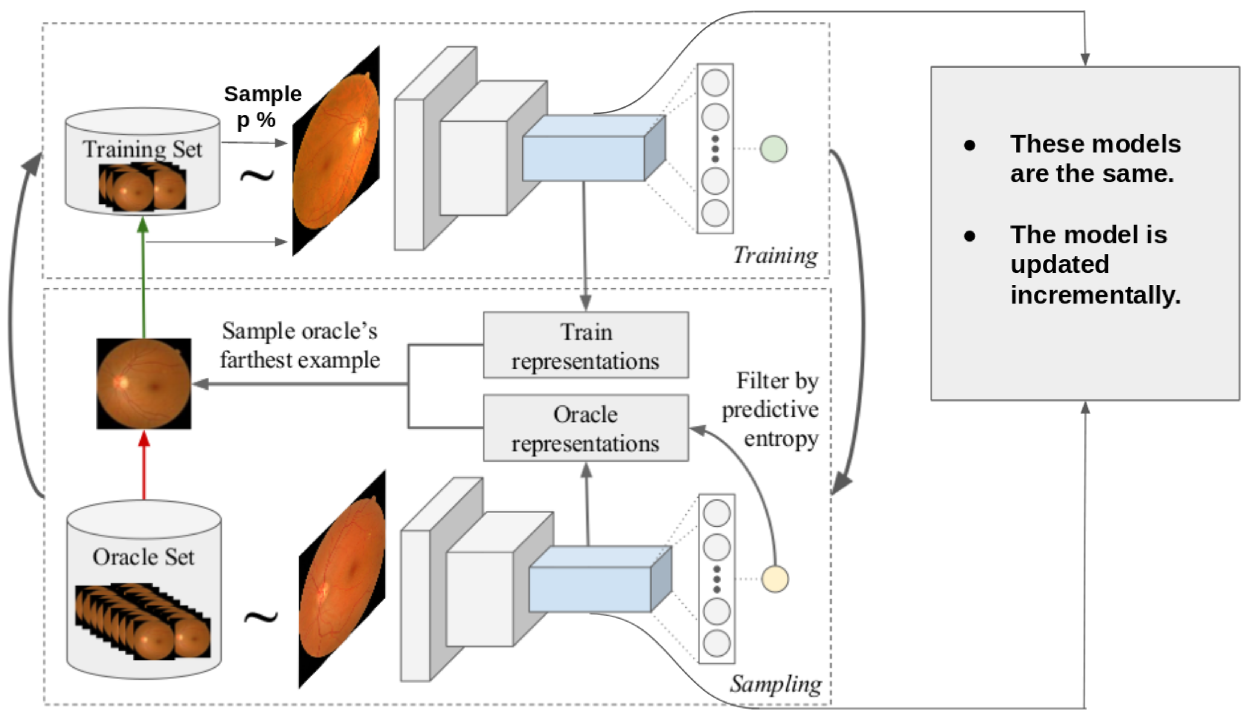
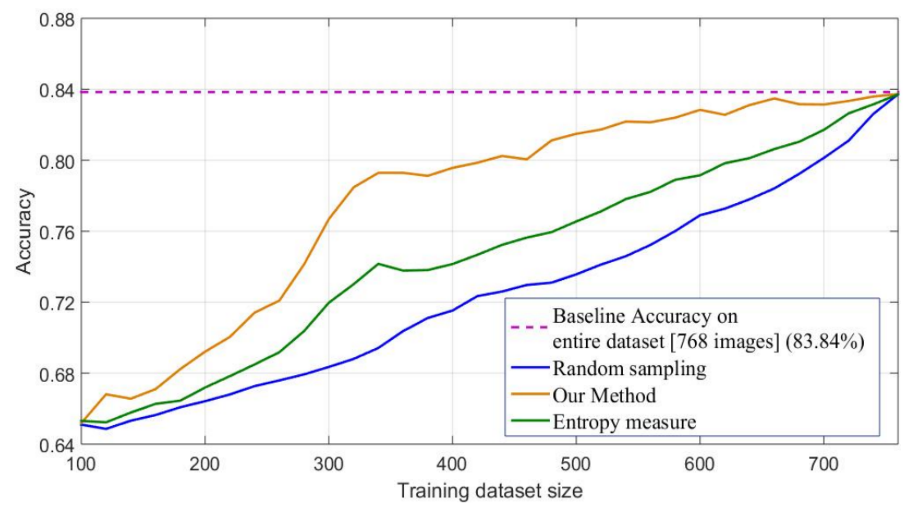


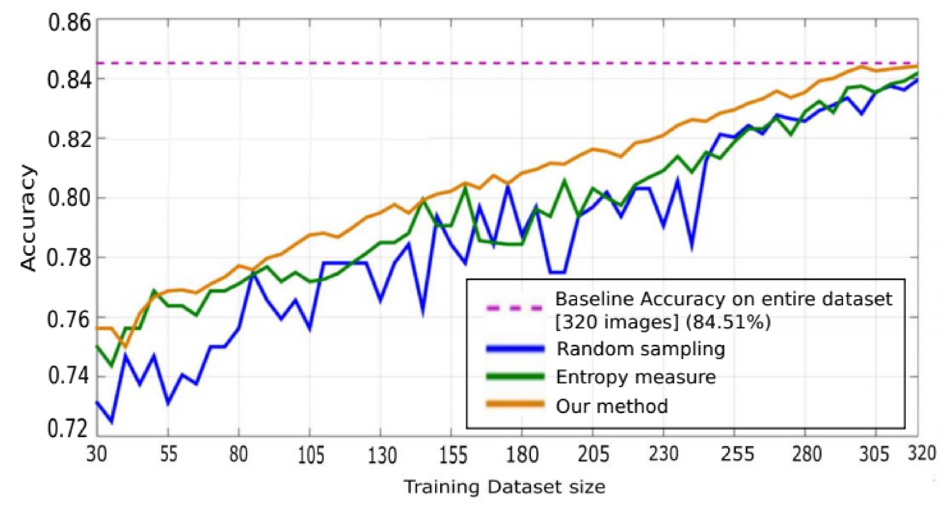
Diagram describes how to **iteratively perform online active learning.** Assume presence of a Deep Network (DN), labeled training data D_{train} (top left of diagram), and unlabeled data D_{oracle} (bottom left of diagram). First, **perform an active learning step**. Use the DN to compute a feature embedding for all labeled examples in D_{train} and the top M unlabeled examples from D_{oracle} with highest predictive entropy. Select and label oracle examples furthest in feature space from the centroid of all labeled examples. The oracle examples are selected one at a time, and the centroid updated after each labeling. Next, **perform the online learning step** by training the model on the expanded training set. In the online setting, the model weights are not reset between iterations and we use only the newly labeled examples and a subset of previously labeled examples. **Repeat**.

Evaluation

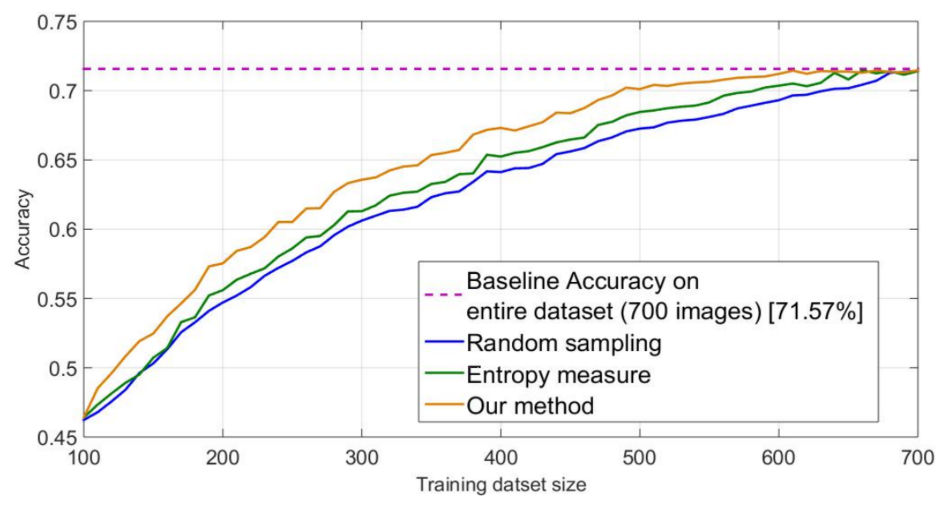
Our method consistently outperforms competing methods.



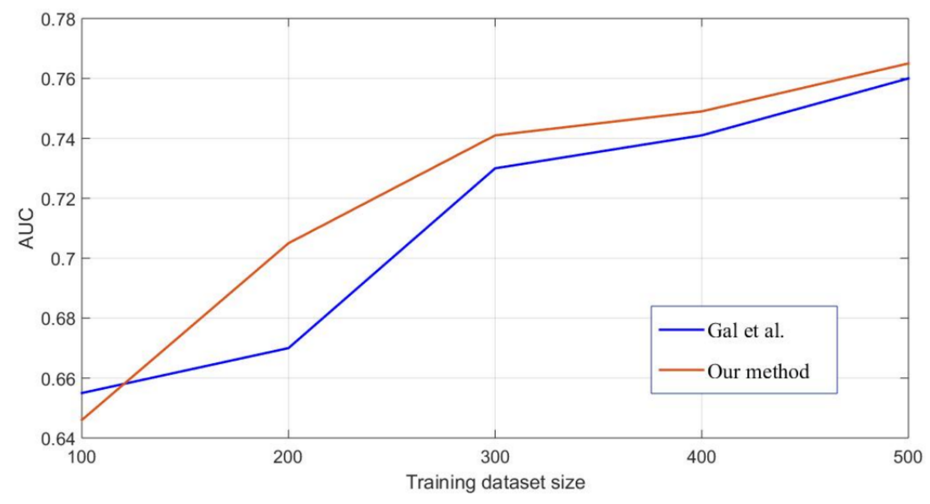
Messidor



Breast Cancer

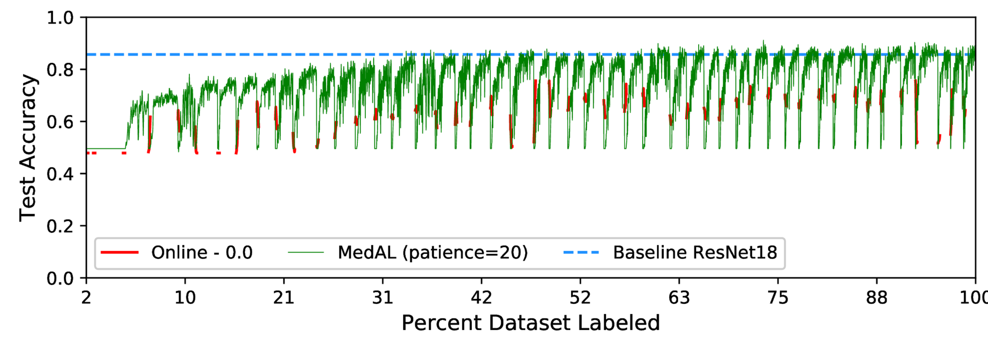


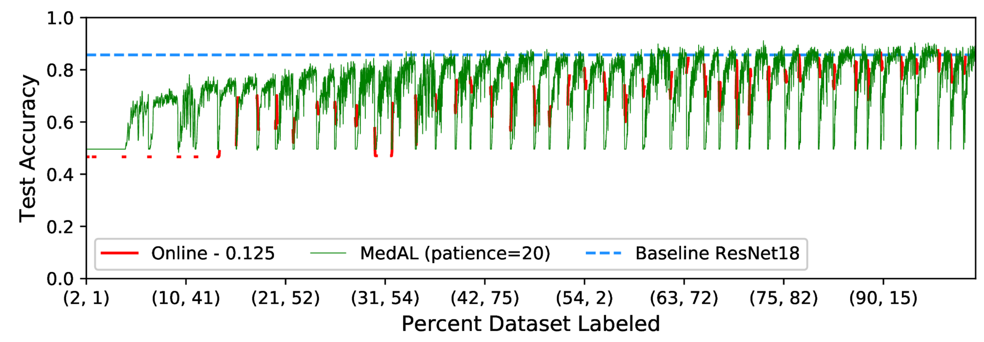
Skin Cancer

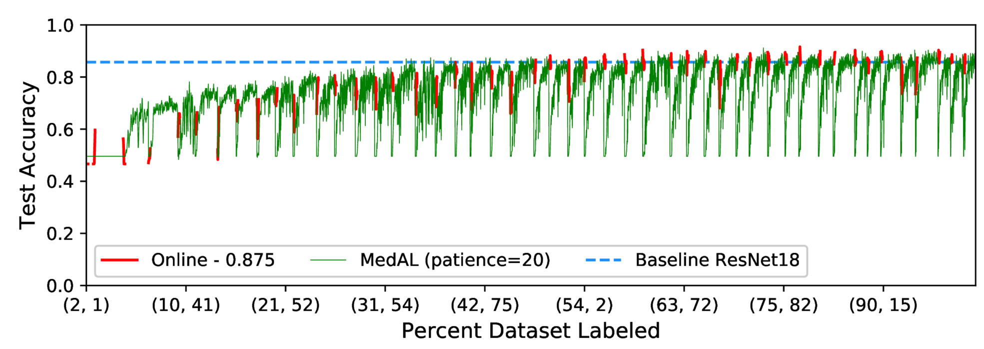


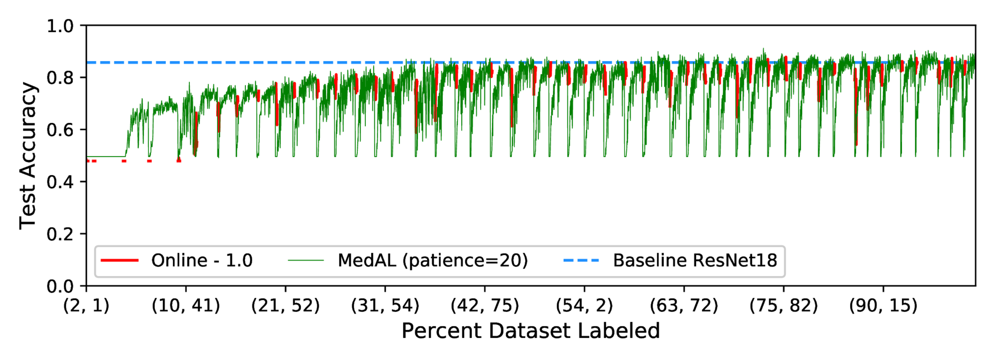
Skin Cancer AUC

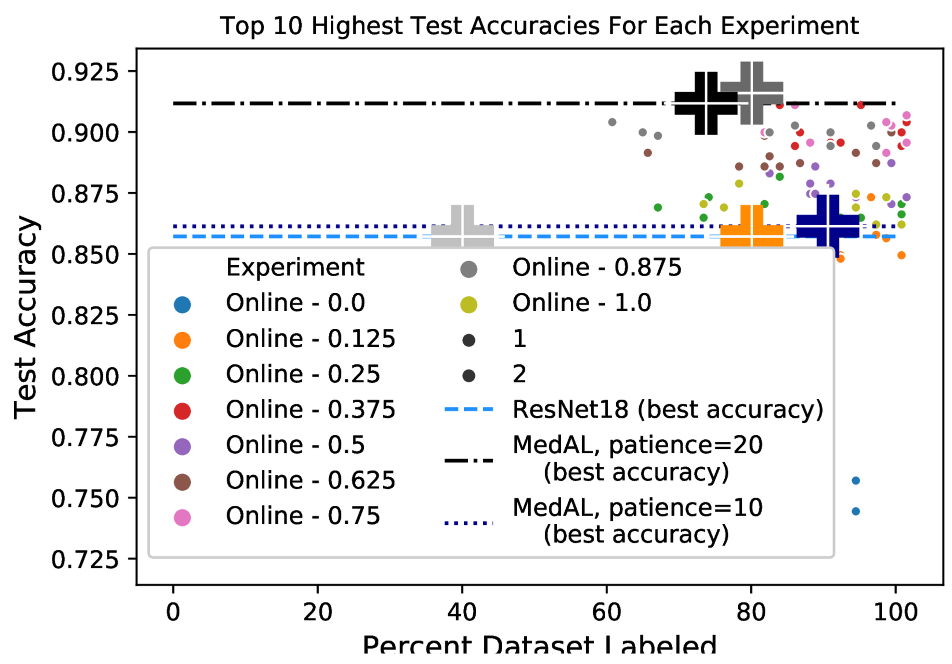
We introduce a hyperparameter, *p*, specifying how much previously labeled data to incorporate into the online training step. We find that higher *p* achieves better performance than the baseline model

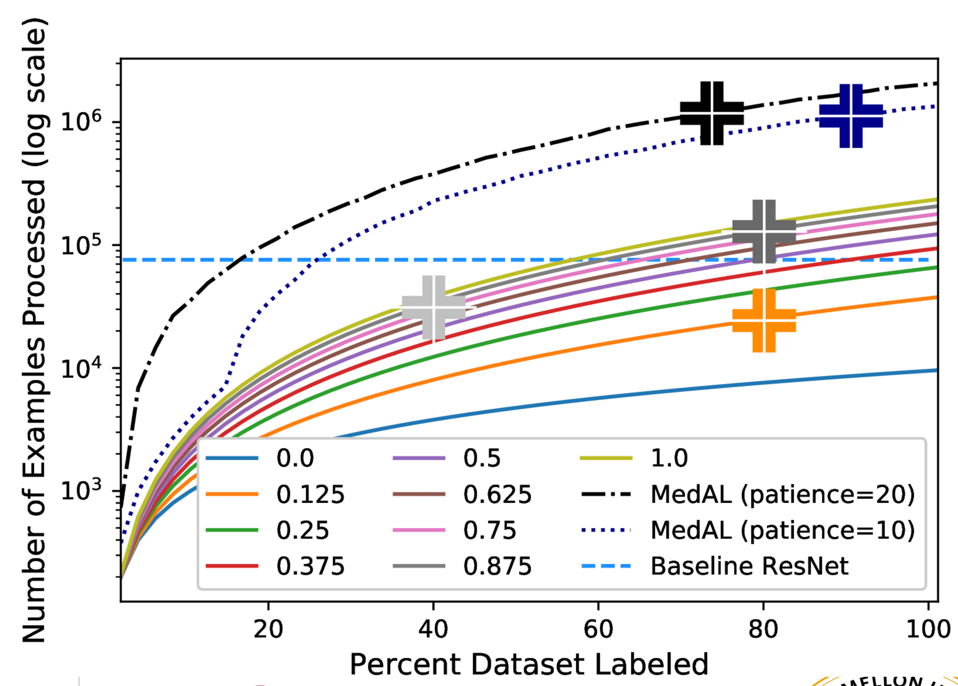












U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

FCT Fundação para a Ciência e a Tecnologia

Carnegie Mellon Portugal

Acknowledgment: We would like to acknowledge the FCT scholarship for Alex Gaudio, Dual Degree PhD student.