DECAL: DEployable Clinical Active Learning

Yash-yee Logan YLOGAN3@GATECH.EDU
Mohit Prabhushankar MOHIT.P@GATECH.EDU
Ghassan AlRegib ALREGIB@GATECH.EDU
OLIVES Lab, Georgia Institute of Technology
Atlanta, GA 30332, USA

Abstract

Conventional machine learning systems that operate on natural images assume the presence of attributes within the images that lead to some decision. However, decisions in medical domain are a resultant of attributes within medical diagnostic scans and electronic medical records (EMR). Hence, active learning techniques that are developed for natural images are insufficient for handling medical data. We focus on reducing this insufficiency by designing a deployable clinical active learning (DECAL) framework within a bi-modal interface so as to add practicality to the paradigm. Our approach is a *plug-in* method that makes natural image based active learning algorithms generalize better and faster. We find that on two medical datasets on three architectures and five learning strategies, DECAL increases generalization across 20 rounds by approximately 4.81%. DECAL leads to a 5.59% and 7.02% increase in average accuracy as an initialization strategy for optical coherence tomography (OCT) and X-Ray respectively. Our active learning results were achieved using 3000 (5%) and 2000 (38%) samples of OCT and X-Ray data respectively.

Keywords: Active Learning, Clinical Context, Real-World Deployment

1. Introduction and Related Work

Active learning aims to find the optimal subset of samples from a dataset for a machine learning model to learn a task well (Dasgupta (2011); Settles (2009)). It is studied because of its ability to reduce the costly and laborious burden on experts to provide data annotations. Typical setups focus on acquisition functions that measure the informativeness of samples using constructs from ensemble learning (Beluch et al. (2018)), probabilistic uncertainty (Gal et al. (2017); Hanneke et al. (2014)) and data representation (Geifman and El-Yaniv (2017); Sener and Savarese (2017)). These works were originally developed for the natural image domain and although several studies have adapted these and other techniques to medical imagery (Logan et al. (2022); Melendez et al. (2016); Nath et al. (2020); Otálora et al. (2017); Shi et al. (2019)), they have not been adopted or utilized in real clinical settings.

One reason for this non-adoption is that conventional active learning does not follow the diagnostic process. This is because of the experimental settings in natural images that aided the development of existing active learning algorithms (Ash et al. (2019); Hsu and Lin (2015); Sener and Savarese (2017)). Natural images typically contain homogeneous class attributes that can be extracted from the images themselves. Also, these attributes are usually enough to distinguish between classes. However, in medicine, pathologies manifest

themselves in visually diverse formats across multiple patients. For example, the characteristics of an aged healthy person are visually different from a young healthy person. So how do doctors overcome this? They include clinical data from EMR to assist with their arrival at a diagnostic decision (Brundin-Mather et al. (2018); Brush Jr et al. (2017)). EMR consists of patient ID, demographics, diagnostic imaging and test results that allow a clinician to make a diagnosis. We recommend that active learning frameworks for medical image classification be designed within a bi-modal interface so as to add practicality to the paradigm. With this in mind, we design and evaluate a DECAL framework that integrates EMR data. We show that DECAL aids existing active learning algorithms in finding the best subset for labeling as well as initializing the active learning framework. As such, DECAL is a *plug-in* approach on top of existing active learning based methods.

2. Assessing Active Learning Framework for Medical Domain

We conduct a set of controlled experiments to evaluate the effectiveness of a DECAL framework relative to conventional frameworks.

2.1 Dataset Descriptions

We use images and EMR data from the OCT dataset by Kermany et al. (2018). The dataset consists of grayscale, cross-sectional, foveal scans having varying sizes. We use images from 3 retinal diseases: 10488 choroidal neovascularization (CNV), 36345 diabetic macular edema (DME) and 7756 Drusen annotated at the image level. Samples in training and oracle sets are from 1852 unique patients. The test set consists of 250 images from each diseased class from 486 unique patients.

We also use images and EMR data from the X-Ray dataset also by Kermany et al. (2018). The X-rays are grayscale, cross-sectional chest scans from children belonging to a healthy class and 2 types of pneumonia: viral and bacterial annotated at the image level. We use 1349 healthy, 1345 viral and 2538 bacterial samples in the combined training and oracle sets from 2650 unique patients. The test set consists of 234 healthy, 148 viral and 242 bacterial images from 431 unique patients. There is also no overlap in patients or imagery in train or test sets for both datasets. This means the imagery in the train and test sets come from different patient cohorts. EMR data used for our analysis was patient identity from both datasets.

Table 1: Information about dataset size and the number of samples added after each training round.

Details	**Dataset**	
	OCT	X-Ray
Oracle + Training Set Images	54589	5232
Images in Initial Training Set	128	128
Images Queried per Iteration	128	128
Unique Patients per Iteration	128	128
Total Unique Patients	2338	3081

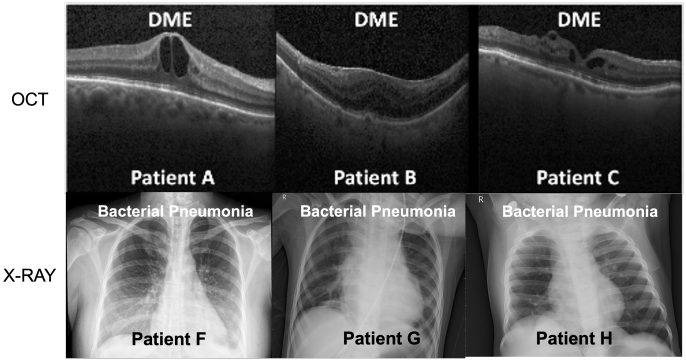


Figure 1: Sample imagery of visual characteristics of disease states across patient cohorts.

2.2 Active Learning with EMR Data

Figure 1 shows sample scans from each dataset with patients having the same disease. The visual characteristics across patients are noticeably different. This intra-class diversity is a typical occurrence in medical datasets. Existing active learning paradigms fail to properly account for disease manifestations from whom there is less data. This oversight is very dangerous for safety critical domains like medicine. Thus, we posit that EMR data, in the form of patient identity, be leveraged to account for the intra-class diversity present in medical datasets. We use patient identity as a *plug-in* constraint that can be applied prior to sample selection with any query acquisition function. The next batch of informative samples will have unique patient identity from the unlabeled pool and be appended to the training set. This process is repeated to determine the minimum number of labeled samples needed to maximize model performance.

2.3 Experiments

Implementation Details We assess our active learning framework on Resnet-18, Resnet-50 and Densenet-121 (He et al. (2016); Huang et al. (2017)). We do not use pre-trained models in any of our analysis. We use the Adam optimizer with a learning rate of 1.5e-4. Hyper-parameters are tuned based on the OCT dataset and then the same parameters are used for the X-Ray dataset. For each round, the Resnet and Densenet models are trained until 98% and 94% accuracy is achieved on the training set respectively. Following each round, the model's weights are reset and randomly initialized. This is repeated with five different random seeds. We aggregate and report average accuracy and standard deviation. All images are resized to 128×128 and OCT scans are normalized with $\mu = 0.1987$ and $\sigma = 0.0786$ while X-Rays with $\mu = 0.4823$ and $\sigma = 0.0379$. Table 1 shows more implementation details for each dataset.

2.3.1 Initializing Active Learning with EMR Data

Existing frameworks typically start active learning by randomly selecting a small amount of samples to train the initial model. Subsequently, they apply methods of ranking sample informativeness. By doing this they naively assume that the data distribution is even, which

is hardly the case in medical datasets as shown in Figure 1. Randomly selecting from an unbalanced distribution is not guaranteed to gather a representative sample of the classes present (Zhu et al. (2008)). Therefore, we recommend the integration of EMR data from the outset to circumvent this.

To do this, we first compute the distribution of patients throughout the unlabeled pool. Then, we select a fixed number of images from unique patients IDs and pair it with its annotation for the initial training set. The intuition behind this strategy is for the first training samples to have maximally dissimilar images. These samples are then used to start our DECAL paradigm. We present two experimental modalities in the initialization phase depending upon the availability of data.

Large Initial Training Set We select 1000 samples at random from the unlabeled pool and train a model for each architecture and dataset for the first round only as our baseline. Then, we perform DECAL initialization by selecting one image from a 1000 unique patients in the unlabeled pool. We then train a model for each architecture and dataset for the first round only and compare it to the baseline by reporting the average accuracy and standard deviation on the test set. The results are presented in Section 3 Table 2.

Small Initial Training Set We select 128 samples with DECAL initialization then start both conventional active learning and DECAL methods and record the earliest round where average accuracy is greater than random chance (33%). Next we compute the percentage increase/decrease that DECAL achieves relative to the corresponding baseline. The results are presented in Section 3 Table 3.

2.3.2 Baseline Sample Acquisition Algorithms

We apply patient ID as a modular "plug-in" constraint prior to sample selection with each of these baseline algorithms to make our framework clinically deployable. The first baseline is standard random sampling, the next three are margin, least confidence and entropy uncertainty based sampling (Settles (2009)) and the last is an amalgamation of diversity and uncertainty-based sampling approaches known as BADGE (Ash et al. (2019)).

3. Results

Initializing Active Learning with EMR Data First, we show results to validate the importance of integrating patient ID from the onset. When a large initial train set is used, Table 2 shows in yellow that DECAL initialization always leads to higher accuracy regardless of architecture or dataset type. DECAL initialization lead to a 5.59% and 7.02% increase in average accuracy for OCT and X-Ray data respectively. However, evaluating the impact of DECAL initialization with a small initial training set mandates a different approach. Since we use non-pre-trained models, training with a small set will unsurprisingly result in over-fitting. This means generalization on the test set will be that of random chance until the training pool becomes large enough. Table 3 highlights in yellow the instances when DECAL initialization lead to a percentage increase in average accuracy during early rounds of training. From these results we see DECAL initialization leads to better generalization for at least 5 of the 10 query strategies across all architectures and datasets.

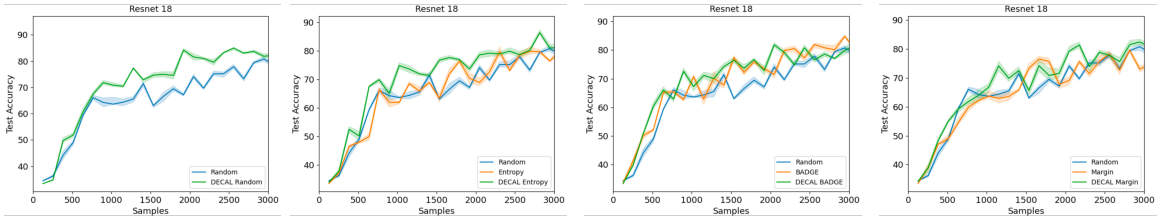
Active Learning with EMR Data We use learning curves to evaluate how DECAL can better characterize disease states. Due to space constraints, we show learning curves for only Resnet-18 and Densenet-121 architectures in Figure 2 and Figure 3. In these plots `x-axis` corresponds to the number of samples in the train set and `y-axis` corresponds to the performance accuracy on the test set. Each colored curve is the average of five trials, with standard errors being shown by the shaded regions. We see that DECAL consistently matches or surpasses the baseline algorithms. Furthermore, we see DECAL often having an edge over baseline algorithms from the early rounds of training onward like in Figures 2b, 3a. This is an indicator that DECAL methods not only generalize better but also generalize faster than the baselines.

Table 2: DECAL vs random initialization using large train set.

OCT Dataset			
Initialization	Resnet-18	Resnet-50	Densenet-121
Random	61.38 ± 4.53	64.4 ± 6.75	76.93 ± 4.64
DECAL	**64.53 ± 9.83**	**69.28 ± 3.41**	**80.16 ± 5.34**
X-Ray Dataset			
Initialization	Resnet-18	Resnet-50	Densenet-121
Random	52.24 ± 9.47	66.95 ± 3.98	70.12 ± 7.44
DECAL	**58.2 ± 9.42**	**71.08 ± 4.17**	**72.82 ± 7.44**

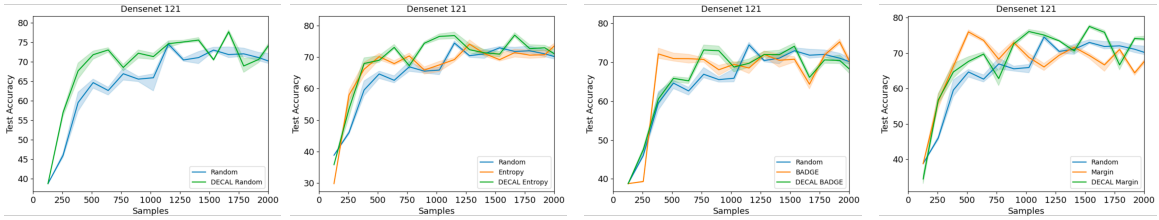
Table 3: DECAL vs random initialization during early rounds of training. +/- numbers show the percentage increase/decrease in average accuracy relative to random.

	Dataset	OCT			X-Ray		
	Round	4	2	1	2	2	1
	Query	Resnet-18	Resnet-50	Densenet-121	Resnet-18	Resnet-50	Densenet-121
Random Initialization	Random	54.37	46.18	42.66	63.91	69.64	46.37
	Entropy	53.62	46.88	49.97	55.88	51.18	46.18
	BADGE	60.53	48.69	35.3	57.50	65.8	40.32
	Margin	55.57	47.01	47.52	56.28	68.34	54.03
	Least Conf	49.49	49.46	46.85	55.03	57.30	47.46
	DECAL Random	61.44	54.5	48.48	64.16	63.42	46.37
	DECAL Entropy	64.48	50.72	52.66	51.92	67.27	51.57
	DECAL BADGE	63.97	53.44	46.9	64.8	69.55	47.08
	DECAL Margin	58.88	48.64	50.4	57.14	62.5	47.03
	DECAL Least Confidence	60.48	49.30	49.78	60.06	64.42	40.99
DECAL Initialization	Random	**+9.21%**	**+3.76%**	**+5.95%**	**+2.03%**	-5.11%	-0.75%
	Entropy	-6.95%	**+3.45%**	-10.74%	**+11.90%**	**+24.73%**	**+25.52%**
	BADGE	**+5.73%**	**+8.64%**	**+21.38%**	**+9.91%**	**1.70%**	-2.48%
	Margin	-1.87%	**+6.53%**	-16.45%	-3.02%	-10.43%	**+2.17%**
	Least Confidence	**+17.61%**	-6.08%	-12.23%	**+5.23%**	**+9.68%**	-6.27%
	DECAL Random	-1.17%	-6.01%	**+14.35%**	**+4.09%**	**+8.04%**	**+22.88%**
	DECAL Entropy	**+4.67%**	-0.53%	-16.71%	**+4.31%**	-5.90%	**+3.83%**
	DECAL BADGE	**+3.25%**	-7.99%	**+0.80%**	**+4.55%**	**+1.01%**	**+0.89%**
	DECAL Margin	**+0.66%**	**+5.53%**	**+5.55%**	**+9.87%**	**+2.91%**	**+20.69%**
	DECAL Least Confidence	-4.19%	-2.86%	-3.37%	**+3.15%**	-0.24%	**+10.85%**



(a) (b) (c) (d)

Figure 2: Test accuracy vs sample count during training for the OCT on Resnet-18.



(a) (b) (c) (d)

Figure 3: Test accuracy vs sample count during training for the X-Ray on Densenet-121.

4. Conclusion and Next Steps

In this work we motivate the design of a DECAL framework for medical applications. Augmenting active learning paradigms with EMR data creates the perfect setting for which active learning can be of true utility within the medical domain. In this study we have demonstrated this by designing an active learning framework constrained on a medically grounded prior gleaned from clinical EMR data about patient identity.

Determining what other forms of EMR best serve active learning paradigms remains an open research question. Our next steps include investigating additional EMR data to inject into our framework for a more holistic analysis. Towards these efforts, we have collaborated alongside Retina Consultants of Texas (Houston, TX, USA) to create our own dataset (Logan* et al. (2022 Under Review)) that contains a plethora of EMR specific to ophthalmology. The clinical information consists of patient identity, general demographics, ocular disease state (Best Corrected Visual Acuity, Central Sub-field Thickness) and detailed ocular imaging in the form of spectral domain OCT, fundus photography and fluorescein angiography collected per the protocol. These were measured and recorded during routine visits to the clinic. In addition to DECAL, clinical context also aids research in supervised contrastive learning (Kokilepersaud et al. (2022)).

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650044.

References

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9368–9377, 2018.

Rebecca Brundin-Mather, Andrea Soo, Danny J Zuege, Daniel J Niven, Kirsten Fiest, Christopher J Doig, David Zygun, Jamie M Boyd, Jeanna Parsons Leigh, Sean M Bagshaw, et al. Secondary emr data for quality improvement and research: a comparison of manual and electronic data collection from an integrated critical care electronic medical record system. *Journal of critical care*, 47:295–301, 2018.

John E Brush Jr, Jonathan Sherbino, and Geoffrey R Norman. How expert clinicians intuitively recognize a medical diagnosis. *The American journal of medicine*, 130(6): 629–634, 2017.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19): 1767–1781, 2011.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Twenty-Ninth AAAI conference on artificial intelligence*, 2015.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

Kiran Kokilepersaud, Mohit Prabhushankar, Ghassan AlRegib, Stephanie Trejo Corona, and Charles Wykoff. Gradient-based severity labeling for biomarker classification in oct. In *International Conference on Image Processing (ICIP)*. IEEE, 2022.

Yash-yee Logan, Ryan Benkert, Ahmad Mustafa, and Ghassan AlRegib. Patient aware active learning for fine-grained oct classification. In *International Conference on Image Processing (ICIP)*. IEEE, 2022.

Yash-yee Logan*, Kiran Kokilepersaud*, Stephanie Trejo Corona, Mohit Prabhushankar, Ghassan AlRegib, and Charles Wykoff. Olives: Optical labels for investigating visual eye semantics. In *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*. IEEE, 2022 Under Review.

Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick H. H. M. Philipsen, Helen Ayles, and Clara I. Sánchez. On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis. *IEEE Transactions on Medical Imaging*, 35 (4):1013–1024, 2016. doi: 10.1109/TMI.2015.2505672.

Vishwesh Nath, Dong Yang, Bennett A Landman, Daguang Xu, and Holger R Roth. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2534–2547, 2020.

Sebastian Otálora, Oscar Perdomo, Fabio González, and Henning Müller. Training deep convolutional neural networks with active learning for exudate classification in eye fundus images. In *Intravascular imaging and computer assisted stenting, and large-scale annotation of biomedical data and expert label synthesis*, pages 146–154. Springer, 2017.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles. Active learning literature survey. 2009.

Xueying Shi, Qi Dou, Cheng Xue, Jing Qin, Hao Chen, and Pheng-Ann Heng. An active learning approach for reducing annotation cost in skin lesion analysis. In *International Workshop on Machine Learning in Medical Imaging*, pages 628–636. Springer, 2019.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, 2008.