LABELING COST SENSITIVE BATCH ACTIVE LEARNING FOR BRAIN TUMOR SEGMENTATION

Maohao Shen[1]*,Jacky Y. Zhang*[2]*,Leihao Chen*[1]*,Weiman Yan*[1]*,Neel Jani*[4]*,Brad Sutton*[3]*,Oluwasanmi Koyejo*[2]

University of Illinois Urbana-Champaign, IL, USA
[1]Dept. of Electrical and Computer Engineering, [2]Dept. of Computer Science,
[3]Dept. of Bioengineering, [4]Carle-Illinois College of Medicine

{maohaos2,yiboz,leihaoc2,weimany2,neeldj2,bsutton,sanmi}@illinois.edu

ABSTRACT

Over the last decade, deep learning methods have achieved state-of-the-art for medical image segmentation tasks. However, the difficulty of obtaining sufficient labeled data can be a bottleneck. To this end, we design a novel active learning framework specially adapted to the brain tumor segmentation. Our approach includes a novel labeling cost designed to capture radiologists' practical labeling costs. This is combined with two acquisition functions to incorporate uncertainty and representation information, ensuring that the active learning selects informative and diverse data. The resulting procedure is a constrained combinatorial optimization problem. We propose an efficient algorithm for this task and demonstrate the proposed method's advantages for segmenting brain MRI data.

Index Terms— Active Learning, Segmentation, Deep Learning, Uncertainty, Approximation Algorithm

1 INTRODUCTION

Semantic image segmentation, a fundamental computer vision task[1], involves assigning class labels to each pixel across an image. One important application area is biomedical image analysis, including the segmentation of 3-D brain tumor magnetic resonance images (MRI). In recent years, many deep learning methods [2, 3, 4] have been designed to achieve state-of-the-art results on brain tumor segmentation. However, deep learning usually relies on vast datasets for training, and access to labeled training data is among the most pressing roadblocks in real-world biomedical image segmentation applications. Specifically, label acquisition requires time-consuming and expensive manual annotation – a high-skill task that must be completed by highly trained and time-constrained physicians.

Active Learning (AL) [5] is an established framework designed to mitigate the problem of scarce labeled data. The standard AL query setting is pool-based sampling [6]. Pool-based active learning is an iterative process where the AL algorithm uses available labeled and unlabeled data to choose the examples which the "labeling oracle" is asked to label. The overall process repeats until a certain performance level is achieved, or the labeling budget is exhausted. Therefore, perhaps the most important task in AL is to decide which sample will be most informative for model training, also known as the query strategy. Many of the standard pool-based AL methods use variants of uncertainty as their query strategy [6]. Examples include max entropy [7] and least confident, margin sampling [8]. Unfortunately, standard AL methods often choose highly correlated and redundant data. Thus, more recently, batch active learning methods have been proposed to mitigate this problem [9, 10, 11], by selecting a large batch of diverse data that's are jointly informative for model training. Batch AL is also ideal for use cases where the labeling oracle must provide multiple labels at once in order to be feasible and/or efficient.

Many AL methods have been developed for image-level classification tasks [12, 13, 11], and a few AL methods have been developed for biomedical image segmentation – including approaches that select data based on model uncertainty with Monte-Carlo (MC) dropout [14, 15]. Relevant work also includes an evaluation of several existing AL algorithms on microscopy image segmentation [16]. Other approaches include conditional GANs to generate diverse samples to help model training [17], and the use of uncertainty and data similarity information provided by FCN, formulating a generalized maximum set cover problem to select the most informative and diverse data [18]. We note that most existing works on active learning often assume that the cost of labeling each sample is identical. This assumption is acceptable for classic ML tasks like image classification. However, for biomedical image segmentation, the labeling process is much more complicated.

Contributions: We aim to develop a novel active learning approach that can select informative and representative data batches. Further, we propose a model for the labeling cost in consultation with radiologists, and incorporate this into active learning. Moreover, we design an AL algorithm that addresses the resulting optimization problem under the outlined conditions and constraints. Finally, we provide experimental results on the BraTS dataset [3, 19, 20]. Our results show that the proposed method outperforms alternatives.

2 PROBLEM FORMULATION

We denote the three dimensional (3-D) brain image dataset as $\mathcal{D} = \{b_i \in \mathbb{R}^{X\times Y\times Z}\}_{i=1}^{N} = \{s_{i,j} \in \mathbb{R}^{X\times Y}\}_{i=1\dots N}^{j=1\dots Z}$. The dataset consists of N 3-D images b_i, and each b_i consists of Z slices of 2-D image, where we denote $s_{i,j}$ as the j^{th} axial slice of the i^{th} 3-D brain image. We focus on axial slices as these were the easiest to label individually as reported by our expert collaborators. We denote the number of segmentation classes as C, and the prediction of a slice as a label $y \in [C]^{X\times Y}$. We denote the labeled data pool as $\mathcal{D}_l$, the unlabeled data pool as $\mathcal{D}_u$, so the full dataset is $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$.

Our goal is to select a subset of slices $\mathcal{A} \subset \mathcal{D}_u$ from the unlabeled data pool to ask for annotation from radiologists subject to a certain labeling cost budget B, such that the model performance can be improved the most after re-training. We denote the labeling cost function as $g(\cdot)$.

Therefore, the batch active learning problem can be formulated as a constrained combinatorial optimization problem:

$$\mathcal{A}^\star = \underset{\mathcal{A}\subset\mathcal{D}_u}{\arg\max}\, f(\mathcal{A}) \quad \text{s.t.} \quad g(\mathcal{A}) \leq B. \tag{1}$$

However, two key problems remain: how to design $f(\cdot)$; and how to solve the maximization in the combinatorial space. To decide which subset of unlabeled data is the most informative, two acquisition functions are designed from different perspectives: uncertainty and representation. In order to solve the combinatorial optimization problem, we propose a two-stage greedy algorithm that utilizes the two acquisition functions.

3 THE PROPOSED METHOD

Modeling Uncertainty. Given a 2-D slice of a 3-D image, the segmentation model is designed to output pixel-wise predictions. We use $p(y|s;\omega)$ to represent the predictive distribution for segmentation $y \in [C]^{X\times Y}$ of an axial brain slice s, where ω is the model parameter. In this work, we use Monte Carlo (MC) Dropout [12] to measure the uncertainty information from y. Specifically, we apply MC dropout T times during model inference for sampling, and we denote the t^{th} dropout model parameter as w_t. Then, the predictive uncertainty of a slice s can be measured by computing the mutual information between its predictions and model posterior. Denoting $\{y_k \in [C]\}_{k=1\dots X\times Y}$ as the prediction of each pixel of a slice, and employing the standard model which assumes conditional independence of each pixel y_k, i.e., $p(y|s,w) = \prod_k p(y_k|s,w)$, the mutual information can be approximated as:

$$\begin{aligned}
&\mathbb{I}\left[y,\omega|s,\mathcal{D}_l\right] = \mathbb{H}\left[y|s,\mathcal{D}_l\right] - \mathbb{E}_{p(w|D_l)}\left[\mathbb{H}\left[y|w,s\right]\right] \\
&= \sum_k \mathbb{H}\left[y_k|s,\mathcal{D}_l\right] - \sum_k \mathbb{E}_{p(w|D_l)}\left[\mathbb{H}\left[y_k|w,s\right]\right] = \\
&-\sum_{k,c}\left(\frac{1}{T}\sum_t p(y_k=c|s,\omega_t)\right)\log\left(\frac{1}{T}\sum_t p(y_k=c|s,\omega_t)\right) \\
&+\sum_{k,c}\frac{1}{T}\sum_t p(y_k=c|s,\omega_t)\log(p(y_k=c|s,\omega_t)).
\end{aligned}$$

Finally, as the uncertainty score is measured independently for each brain slice, our uncertainty based acquisition function can be formulated as:

$$f_1(\mathcal{A};w,\mathcal{D}_l) = \sum_{s\in\mathcal{A}} \mathbb{I}\left[y,w|s,\mathcal{D}_l\right].$$

The intuition behind this acquisition function is to choose those slices that the model is overall most uncertain about, and those slices that the model disagrees the most with different model parameters. Thus, we expect that by providing true annotation by an oracle for such slices will be helpful for model training at the next iteration.

Measuring Representativeness. While uncertainty based acquisition is used to select data that are most uncertain, it does not consider the correlation between samples, which could result in a highly redundant set of data. For example, simply selecting the highest entropy unlabeled data can be worse than random selection [9]. We propose additional acquisition functions $f_2(\cdot)$ to overcome this problem by quantifying the diversity of the selected subset. In other words, $f_2(\cdot)$ indicates how well a small subset represents the whole unlabeled dataset.

To this end, we consider pairwise similarity. We define a similarity function using kernel distance $h(x,y) = e^{-\frac{\|x-y\|}{\sigma}}$. In order to quantify the similarity between two slices s and s', one way is directly using $h(s,s')$, but the computation of the distance between raw images is cumbersome. One better solution is to extract the features from raw images using an image descriptor. Our model has a U-net structure [2] with an encoder to extract image features when downsampling the input images, and a decoder to reconstruct the segmentation mask. Thus, the similarity can be computed by $h(\phi(s),\phi(s'))$, where $\phi(s)$ and $\phi(s')$ are the spatially smaller feature map output by the last layer of encoder given input slices s, s' respectively. Using this pairwise similarity, we design two alternative acquisition functions to quantify the representatives of a given subset $\mathcal{A}$.

First we consider maximizing the outer similarity, i.e., the similarity between data in $\mathcal{A}$ and other unlabeled data outside of $\mathcal{A}$, while minimizing the inner similarity, i.e., the similarity of data within the $\mathcal{A}$. Formally, this is given by:

$$f_2(\mathcal{A};\mathcal{D}_u) = \frac{\sum_{s\in\mathcal{A}}\sum_{s'\in\mathcal{D}_u\setminus\mathcal{A}} h(\phi(s),\phi(s'))}{\sum_{s\in\mathcal{A}}\sum_{s'\in\mathcal{A}} h(\phi(s),\phi(s'))}. \tag{2}$$

An alternative approach is to maximize the sum of pairwise similarities between each sample in unlabeled data pool and

its most similar sample in the subset $\mathcal{A}$. Formally,

$$f_2(\mathcal{A}; \mathcal{D}_u) = \sum_{s \in \mathcal{D}_u \setminus \mathcal{A}} \max_{s' \in \mathcal{A}} h(\phi(s), \phi(s')). \quad (3)$$

Maximizing this function is equivalent to the *K-Median* problem, except that *K-Median* problem is minimizing the distance, whereas we are interested in maximizing the similarity.

Combined Cost Function. Unlike the standard active learning setting where the cost of labeling is simply the number of samples to be labeled, the labeling process for 3-D brain images is more complicated. Given the labeled dataset $\mathcal{D}_l$ and the subset of data to be labeled by radiologists $\mathcal{A} \subseteq \mathcal{D}_u$, we aim to design a cost function that can approximate the true labeling cost:

$$g(\mathcal{A}; \mathcal{D}_l),$$

which depends on both subset $\mathcal{A}$ and the current labeled data pool $\mathcal{D}_l$. We use $\mathcal{A}(i)$ to represent the axial slices in the subset $\mathcal{A}$ that are taken from i^{th} brain individual. Similarly, we use $\mathcal{D}_l(i)$ to represent the axial slices in $\mathcal{D}_l$ that are taken from i^{th} brain individual. In consultation with radiologists, we propose the following two assumptions that capture key properties of the real labeling process.

Assumption 1 *The cost incurred by labeling for one individual is independent from labeling for other individuals. Formally,*

$$g(\mathcal{A}; \mathcal{D}_l) = \sum_i g(\mathcal{A}(i); \mathcal{D}_l) = \sum_i g(\mathcal{A}(i); \mathcal{D}_l(i)).$$

Assumption 2 *The cost of labeling a new slice depends on the distance to its nearest labeled slice or slices already in subset $\mathcal{A}$. Formally, for a slice $s \in \mathcal{D}_u(i)$,*

$$g(\mathcal{A} \cup \{s\}; \mathcal{D}_l) - g(\mathcal{A}; \mathcal{D}_l) = \begin{cases} c(d(s, s')), & \exists s' \in \operatorname{argmin}_{s' \in \mathcal{D}_l(i) \cup \mathcal{A}(i)} d(s, s') \\ c_0, & otherwise, \end{cases}$$

where $c(\cdot)$ is a monotone increasing function representing the cost of labeling a new slice based on the distance to its nearest labeled slice, and $d(s, s')$ denotes the distance between two slices, and c_0 is the cost of labeling the first slice of an individual whose slices have not yet been labeled.

We expect the cost function $c(\cdot)$ to be monotone increasing w.r.t. distance d, and be always smaller than c_0. These properties of $c(\cdot)$ reflect the real scenario that labeling a slice near labeled slices costs cheaper than labeling a slice far away from labeled slices, as labeled slices could be used as references when labeling around. We found that one reasonable $c(\cdot)$ is:

$$\begin{cases} c(d) &= \log(1 + d) \\ c_0 &= \alpha \cdot \log(1 + \textit{MaxDistance}), \quad \alpha > 1, \end{cases}$$

where *MaxDistance* is the largest possible distance between two slices of a brain.

Note that the definition of $g(\mathcal{A}, \mathcal{D}_l)$ depends on the order of labeling for slices in $\mathcal{A}$, but as an approximation, we ignore the difference incurred by labeling order. Therefore, we use the natural order generated by our algorithm, and it is always possible to implement the algorithm with a different labeling order, e.g., the best order that minimizes $g(\mathcal{A}, \mathcal{D}_l)$.

3.1 Active Learning Algorithm

Algorithm 1: Greedy Active Learning

1 **Input:** Data pool $\mathcal{D}_l, \mathcal{D}_u$; Cost constraint B.
2 $\mathcal{A} \leftarrow \emptyset$, $s_0 \leftarrow \emptyset$, and $n \leftarrow 0$
3 $\mathcal{D}_c^\star \leftarrow \arg\max_{\mathcal{D}_c \subseteq \mathcal{D}_u, |\mathcal{D}_c| \leq M} \sum_{s \in \mathcal{D}_c} f_1(s; \mathcal{D}_l)$
4 **while** $g(\mathcal{A} \cup s_n; \mathcal{D}_l) \leq B$ **do**
5 $\quad \mathcal{A} \leftarrow \mathcal{A} \cup s_n$
6 $\quad n \leftarrow n + 1$
7 $\quad s_n \leftarrow \arg\max_{s \in \mathcal{D}_c^\star \setminus \mathcal{A}} \frac{f_2(\mathcal{A} \cup s; \mathcal{D}_c^\star) - f_2(\mathcal{A}; \mathcal{D}_c^\star)}{g(\mathcal{A} \cup s; \mathcal{D}_l) - g(\mathcal{A}; \mathcal{D}_l)}$
8 **Output:** a subset of data $\mathcal{A} = \{s_1, ..., s_{n-1}\}$

As the computation of $f_2(\cdot)$ is $O(|\mathcal{D}_u|)$ times more complex than the computation of $f_1(\cdot)$, it is necessary to design a complexity efficient and adaptive way to combine the two metrics. Thus, we design a hierarchical selection procedure: first, $f_1(\cdot)$ is used to select a large subset of data uncertain to the model, with the number of slices no more than M ; then, $f_2(\cdot)$ is used to find the most representative batch of data among the uncertain samples, with its cost no more than B. That is, equation (1) is reformulated into a two stage optimization problem:

$$\mathcal{D}_c^\star = \arg\max_{\mathcal{D}_c \subseteq \mathcal{D}_u} f_1(\mathcal{D}_c; \mathcal{D}_l) \quad \text{s.t.} \quad |\mathcal{D}_c| \leq M$$
$$\mathcal{A}^\star = \arg\max_{\mathcal{A} \subseteq D_c^\star} f_2(\mathcal{A}; \mathcal{D}_c^\star) \quad \text{s.t.} \quad g(\mathcal{A}) \leq B$$

The first sub-problem can be solved easily by choosing the slices with top $f_1(\cdot)$ values (line 3 in Algorithm 1), since the uncertainty score $f_1(\cdot)$ is evaluated independently for each slice. Unfortunately, the second sub-problem is a combinatorial optimization problem, solving which optimally is NP-hard. However, we can efficiently approximate the solution by a greedy procedure shown in line 4-7 in Algorithm 1: at each iteration, we look for one slice s that maximizes the marginal gain per cost increase.

4 EXPERIMENTAL RESULTS

We implemented our experiments using an encoder-decoder based 3-D U-net-like network [2]. The training and testing were done on the BraTS 2018 [3, 19, 20] training dataset (285 brain cases), where we randomly split the dataset into a training set (233 cases) and a testing set (52 cases). Each brain image is a 3-D image with four MRI modalities (T1,

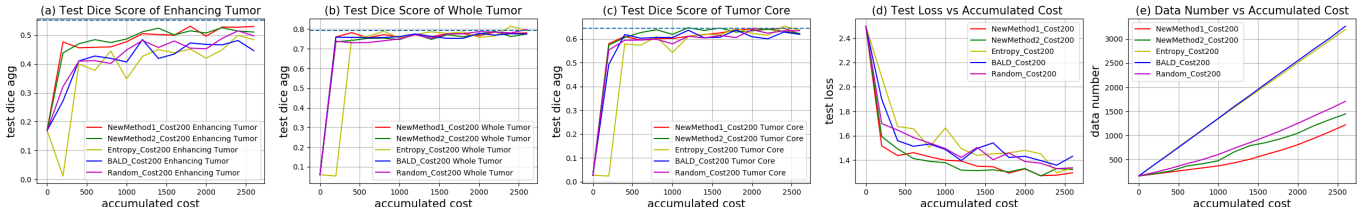

Fig. 1: Results with Cost Budget 200: AL learning curves of three tumor subregions: (a)ET, (b)WT, (c)TC; (d)Test loss as a function of accumulated cost; (e)Data number as a function of accumulated cost.The NewMethod1 (red curve) and NewMethod2 (green curve) are the proposed Algorithm 1 with (equation (2)) and (equation (3)) respectively. Our proposed methods have better general performance as shown in (d), and are especially good at the segmentation of ET as shown in (a). Also, they require the fewest slices for training as shown in (e).

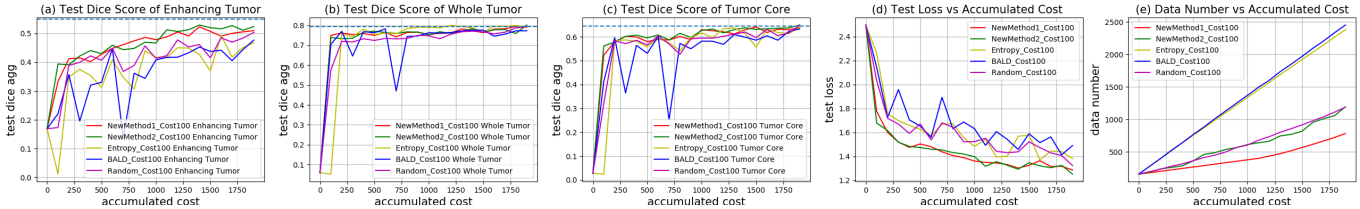

Fig. 2: Results with Cost Budget 100: AL learning curves of three tumor subregions: (a) ET, (b) WT, (c) TC; (d) Test loss as a function of accumulated cost; (e) Data number as a function of accumulated cost. The NewMethod1 (red curve) and NewMethod2 (green curve) are the proposed Algorithm 1 with (equation (2)) and (equation (3)) respectively. Our propped methods have better general performance as shown in (d), and are especially good at segmentation of ET as shown in (a). Also, they require the fewest slices for training as shown in (e).

T1Gd, T2, FLAIR) and input size of $240 \times 240 \times 155$. In our experiments, we used downsampled and cropped data to fit into GPU memory. The model outputs a segmentation mask with three channels to represent three different kinds of tumor subregions: Enhancing Tumors (ET), Whole Tumor (WT), and Tumor Core (TC). The model training and active learning process are only performed on training data, and the performance of the trained model is evaluated on testing data.

Initially, the model was trained on 5 fully labeled brain individuals as seed examples; after the model was well trained, its performance was evaluated on testing data. New slices were selected for labeling by the query strategy at each iteration. Then, such new labeled slices were added into labeled data to perform the next training – this process is repeated until the accumulated total cost reaches a maximum certain threshold.

The quality of segmentation performed by the trained model is evaluated by the Dice score:

$$\frac{2 \times \sum p(y|s) \cdot p_{true}}{\sum p(y|s)^2 + \sum p_{true}^2},$$

where $p(y|s)$ is the segmentation mask output of model, and p_{true} is the ground truth mask.

To compare our proposed active learning methods with other baselines, we kept the initial seed example and the cost budget B identical across every method. The baseline methods include Random Sampling, Max Entropy [7], and BALD [12]. At each iteration of active learning, we collected the Dice score evaluated on testing data. In the end, several results were reported: Testing Dice Score vs. Accumulated Cost across three tumor subregions, Testing Loss vs. Accumulated Cost, and Accumulated Data Number vs. Accumulated Cost. The testing Dice score obtained using fully labeled training data were further reported as a dashed line in the graph, as the best performance the model can achieve. To compare the performance of active learning methods under different labeling cost conditions, we report results of different cost budget: The results of cost budget $B = 200$ are shown in Fig.1, and the results of cost budget $B = 100$ are shown in Fig.2.

It can be seen that our proposed method outperforms the baselines, especially for the segmentation of Enhancing Tumors (ET), which is the subregion that is most clinically relevant for segmentation among the three tumor classes, as it shows an active tumor growing region which is the primary target of therapy. To better visualize the general performance of active learning methods on three tumor subregions, the test loss curve also showed that the testing loss of our proposed methods converged much faster than other baseline methods. Meanwhile, while our proposed methods have better performance, they also require the least number of slices for training.

5 CONCLUSION

In this paper, we introduce active learning to help mitigate the scarce labeling problem of brain tumor segmentation. While most existing methods assume cardinality labeling cost, we present a novel active learning framework that incorporates labeling cost to model the practical annotation cost of radiologists. We combine two metric functions from both uncertainty and representation perspectives to select diverse data. In addition, we provide an efficient algorithm adapted to our active learning setting. These are several interesting directions for future work, including further improved uncertainty metrics to incorporate geometric correlations, e.g., using a graphical model, training a regression model using real collected cost data to better approximate the labeling cost, and considering more general datasets and other tasks.

6 Compliance with Ethical Standards

This research was conducted with data made available in open access by BraTs 2018. Ethical approval was not required as confirmed by the license attached with the open-access data.

7 Acknowledgments

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

8 References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[3] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[4] M. W. Nadeem et al., "Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges," *Brain Sciences*, vol. 10, no. 2, pp. 118, 2020.

[5] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.

[6] B. Settles, "Active learning literature survey," Tech. Rep., University of Wisconsin-Madison Department of Computer Sciences, 2009.

[7] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication, by CE Shannon (and Recent Contributions to the Mathematical Theory of Communication), W. Weaver*, University of illinois Press, 1949.

[8] N. Roy and A. McCallum, "Toward optimal active learning through monte carlo estimation of error reduction," *ICML, Williamstown*, pp. 441–448, 2001.

[9] A. Kirsch, J. van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 7026–7037.

[10] R. Pinsler, J. Gordon, E. Nalisnick, and J. M. Hernández-Lobato, "Bayesian batch active learning as sparse subset approximation," in *Advances in Neural Information Processing Systems*, 2019, pp. 6359–6370.

[11] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.

[12] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," *arXiv preprint arXiv:1703.02910*, 2017.

[13] K. Wang et al., "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[14] M. L. di Scandalea et al., "Deep active learning for axon-myelin segmentation on histology data," *arXiv preprint arXiv:1907.05143*, 2019.

[15] M. Gorriz et al., "Cost-effective active learning for melanoma segmentation," *arXiv preprint arXiv:1711.09168*, 2017.

[16] J. Roels and Y. Saeys, "Cost-efficient segmentation of electron microscopy images using active learning," *arXiv preprint arXiv:1911.05548*, 2019.

[17] D. Mahapatra et al., "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 580–588.

[18] L. Yang et al., "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 399–407.

[19] S. Bakas et al., "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, pp. 170117, 2017.

[20] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.