Combining Image and Region Uncertainty-based Active learning for Melanoma Segmentation

Nicolas Martin, Jean-Pierre Chevallet, Philippe Mulhem, Georges Quénot

Combining Image and Region Uncertainty-based Active learning for Melanoma Segmentation

Nicolas Martin
PEEKTORIA &
Univ. Grenoble Alpes,
CNRS, Grenoble INP§, LIG
F-38000 Grenoble France
nicolas.martin@
peektoria.com

Jean-Pierre Chevallet
Univ. Grenoble Alpes,
CNRS, Grenoble INP§, LIG
F-38000 Grenoble France
jean-pierre.chevallet@
univ-grenoble-alpes.fr

Philippe Mulhem
Univ. Grenoble Alpes,
CNRS, Grenoble INP§, LIG
F-38000 Grenoble France
philippe.mulhem@
univ-grenoble-alpes.fr

Georges Quénot
Univ. Grenoble Alpes,
CNRS, Grenoble INP§, LIG
F-38000 Grenoble France
georges.quenot@
univ-grenoble-alpes.fr

***Abstract*—Segmentation of medical images using learning-based systems remains a challenge in medical computer vision, as training a segmentation model requires exhaustively annotated medical images by experts, which are difficult and expensive to obtain. In the context of melanoma segmentation, we explored active learning methods to combine human annotation for the most uncertain pixels with model predictions for the others. With only around 30% of the images annotated, we achieved performance similar to that obtained with a fully annotated dataset. We also demonstrated that, after a few iterations, experts can focus on annotating only the most uncertain areas of the images, relying on the model for the rest. These approaches pave the way for accelerating the annotation of unlabeled medical datasets and optimizing the use of medical expertise in deep learning projects.**

***Index Terms*—Active Learning, Image segmentation, Medical images, Partial annotation**

I. Introduction

Medical image segmentation is a challenging task in computer vision [1], and correctly segmenting anatomical structures is crucial for computer-aided detection systems [2]. To address this challenge, various deep learning algorithms have been specifically tailored, such as Mask R-CNN [3] and U-Net [4]. This task relies heavily on expert annotations [5], which can be costly and challenging to acquire.

It has been shown that, during the learning of segmentation, not all images are equal, and their annotations influence the training process and the final performance of the segmentation system [6]. Selecting the most informative images should be more beneficial to model performance than random selection of images [7]. This assumption has led to the development of numerous Active Learning (AL) methods designed to select the most informative samples for annotation [6].

In the medical domain, obtaining images for diagnostic or archival purposes has become standard practice, leading to the availability of large datasets [8]. However, these datasets are rarely annotated [9]. Thus, selecting the most informative images using AL methods presents a valuable opportunity to significantly alleviate the annotation workload for experts, thereby promoting the creation of more efficient medical tools based on deep learning algorithms: see [10] for a review of AL for medical images.

Within the images themselves, the need for expertise for annotation also varies: some areas do not require expert annotation (e.g., interior of melanomas), while other areas (e.g., outline of organs) require precise, high-quality annotation [5].

Skin cancer remains a major public health problem, with more than 130,000 cases of melanoma per year [11], which could benefit from computer-assisted diagnosis. Segmenting skin lesions from images is an important step towards achieving this goal and remains a challenging task [12].

The contributions of this paper are:

- a detailed exploration of Active Learning (AL) methods for melanoma image segmentation, combining i) the selection of images to be annotated with ii) a merging of human annotation for uncertain areas and model prediction for others;
- the proposal of a model-agnostic, efficient, and effective AL method for melanoma segmentation.

We compare several methods to select the most uncertain images to minimize annotation cost and maximize segmentation performance. Firstly, we identify the least confidence method as the most efficient approach for image selection, surpassing the segmentation performance achieved with a fully annotated dataset using only 30% of the images. Secondly, we show that it is possible to dramatically lower the number of annotations while preserving segmentation quality: less than 10% of the image pixels need to be annotated after a few AL iterations.

Section 2 presents the related work on active segmentation learning. Section 3 describes the explored methods. Section 4 presents the results, which are further discussed in Section 5.

II. Related work

Active learning for classification has been deeply explored to facilitate annotation labor (see [13] for a review). It can be summarized as the principle of selecting the most informative images to be annotated for a given task [14]. This is a powerful approach to train models from an unlabeled dataset and reduce annotation cost. Such a process consists of a pool of unlabeled

§Institute of Engineering Univ. Grenoble Alpes

data and an expert (also called an "oracle") who is able to annotate selected images. The principle is depicted in Fig. 1:

(1) Evaluate the informativeness (using a previously trained model) of each image in a pool of unlabeled data
(2) Select the most likely informative images
(3) Ask the oracle to annotate the selected images
(4) Train from scratch or fine-tune a new model using the already annotated images and the newly annotated images
(5) Evaluate the new model and repeat the AL process until the desired performance is reached or the annotation budget is consumed

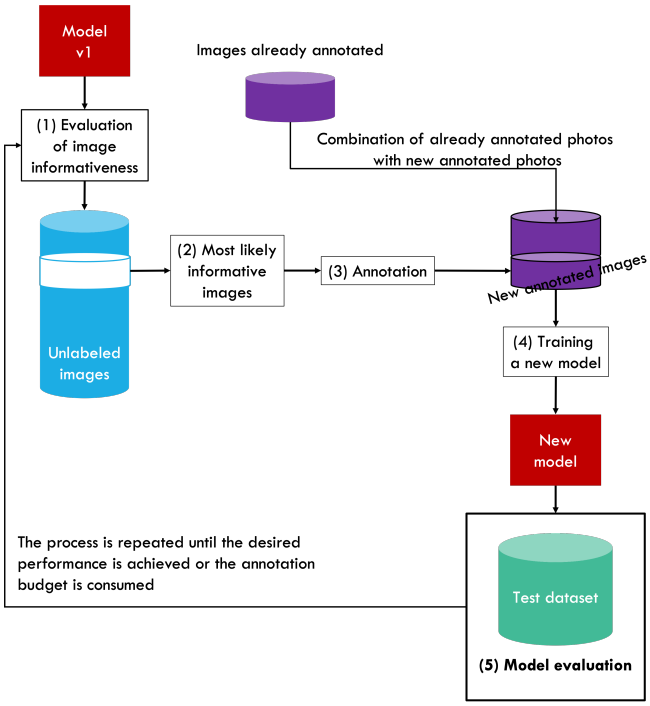

Fig. 1. Active learning principle

Although numerous methodologies for active learning (AL) have been proposed for image classification, the field of active learning for semantic segmentation remains relatively unexplored [15]. While improving the annotation process in the medical domain is crucial due to the rarity and expense of medical expertise, less attention has been devoted to the segmentation of medical images [16]. The majority of existing methods rely on benchmark datasets such as Cityscapes [17], which are characterized by their high diversity, originally curated to provide comprehensive coverage within their respective domains [18].

Three main challenges are encountered in the adoption of active learning (AL) for segmentation [19]: the cold-start problem (i.e., determining the number of samples to annotate initially), the sample selection strategy, and the pixel-wise annotation of selected images. It has been shown that the superiority of AL methods compared to random selection is highly dataset-dependent [20].

In this paper, we focus on the pixel-wise nature of the segmentation problem. We aim to reduce the annotation burden for experts by minimizing not only the number of images to annotate but also the number of pixels to annotate. Learning with partial annotation raises specific challenges because classical loss functions cannot distinguish unannotated pixels from pixels not belonging to the class [21]. Thanks to the iterative process of active learning (AL), the combination of human and model annotation can be explored as the model gradually learns to make predictions, which can potentially be used as pseudo-labels.

In the context of melanoma, several active learning (AL) approaches have been explored using standard datasets such as the ISIC dataset (e.g., [22], [23]). Most of the proposed approaches rely on image-level selection and require full annotation of the selected images. However, several works have explored reducing the amount of annotation required per image.

Such reduction may rely on rough annotations, such as bounding boxes or image-level labels, which have been explored to reduce the need for pixel-wise annotations but still present insufficient results for medical applications [19]. This has led to the consideration of pixel-wise human annotations. For instance, the authors of [24] estimate uncertainty using Monte Carlo Dropout methods [25], [26]. During the AL process, at the image level, they propose manually annotating the most uncertain images and pseudo-annotating, with the trained model, some images (all pixels) when the uncertainty is below a threshold. [24] achieved a segmentation quality, according to the Dice similarity evaluation [22], [23], of 74% on the ISIC dataset after 9 AL iterations. [19] proposed a multi-stage method to reduce the need for pixel-wise annotation. They rely on three stages: (1) the selection of the most valuable images to annotate based on pixel entropy uncertainty, region consistency, and image diversity; (2) expert annotation of the most uncertain super-pixels for the selected images; and (3) iteration of the super-pixel annotation process for the selected images. [27] proposed an interactive process based on the combination of model prediction (U-Net model with conditional random fields) and human correction. The correction of the proposed annotation is repeated until a satisfactory annotation is obtained.

Other approaches have been studied in different medical domains, such as [18]. Instead of annotating the whole image during the AL process, their approach aims to select the most uncertain image patches. The uncertainty is estimated by calculating the difference between the prediction from the original images and a transformed version of these images.

Although this approach reduces the annotation effort, it requires annotating entire patches and does not rely on the model's predictions for pseudo-labels.

All of these approaches reduce the quantity of annotation required but: i) do not directly exploit the combination of within-image pseudo-labeling (i.e., with the trained model) and human annotations for the most uncertain areas, ii) require several iterations per selected image, and iii) do not rely on

AL to leverage this progressive learning. In this paper, we explore, in the context of melanoma, the selection of the most informative images (summarizing the uncertainty at the pixel level) and the combination of human annotation for the most uncertain areas and model annotations for the other areas.

III. Proposed framework

A. Uncertainty estimation at pixel and image levels

We focus on image selection methods that consider the informativeness of images (step 1 in Fig. 1), particularly those relying on uncertainty estimation [28]. Uncertainty-based methods aim to select the most uncertain samples for annotation [14]. In segmentation tasks, the uncertainty estimation of an image depends on the uncertainty estimation of each of its pixels.

We consider three major informativeness estimations[1]: least confidence [19], entropy [19], and Monte Carlo Dropout [25].

The **least confidence** approach involves selecting the images for which the model is the least confident. The least confidence score for one pixel is defined in Equation (1):

$$Uncertainty_{LeastConfidence}(\hat{y}) = |0.5 - \hat{y}| \tag{1}$$

where $\hat{y}$ is the predicted value by the model for one pixel y of an input image.

The **entropy** approach consists in selecting the images with the maximum of information. The entropy score is defined, for one pixel, as:

$$Uncertainty_{Entropy}(\hat{y}) = -\hat{y}\log_2\hat{y} \tag{2}$$

The **Monte Carlo (MC) Dropout** method is based on prior work showing that Dropout can be used to estimate model uncertainty [26], especially for semantic segmentation [25]. Similar to [24], we compute the standard deviation of N predictions of the model at the pixel level. The Monte Carlo (MC) Dropout uncertainty score is defined in Equation (3):

$$Uncertainty_{MCDropout}(\hat{y}) = \sqrt{\frac{1}{N}\sum_{j}^{N}(\hat{y_j} - \bar{y})^2} \tag{3}$$

where N is the number of Dropout samples and $\hat{y_j}$ one of the N predictions for one pixel, with $\hat{y} = \{\hat{y_j}\}$.

The uncertainty scores mentioned above are calculated at the pixel level. These pixel-level uncertainties are summarized at the image level using classical metrics such as mean, median, or sum [29]. In this paper, we focus on summing the pixel values at the image level.

[1]Methods based on margin are not considered in this paper, as they require at least two classes in the images to be applied.

B. Partial human annotation of images

In the majority of proposed AL methods for segmentation, entire images or image patches are annotated (step 3 in Fig. 1). A qualitative exploration of the uncertainty maps during the AL process showed that only limited areas have high uncertainty. We found that these areas are generally located on the border between the two classes: background and melanoma.

Based on these results, we explore the partial annotation of images selected by the AL process. To do this, the expert only annotates the most uncertain areas. The following steps are followed to combine model and human annotations (see Fig. 2):

1) Produce the model probabilities
2) Select the most uncertain images
   a) Binarize the uncertainty mask with a defined threshold τ_u (e.g., .25)
   b) Binarize the predicted mask with a defined threshold τ_b (e.g., .50)
   c) Annotate the most uncertain areas with human annotation ("Selected human annotation" in Fig. 2) and annotate the most certain areas with the binarized prediction mask ("Selected model annotation" in Fig. 2)
3) Repeat the AL process until desired performance is reached or the annotation budget is consumed

IV. Experiments

A. Dataset

The dataset from the ISIC Challenge 2018 [30], [31] is utilized in the following experiments. It contains RGB images of melanoma along with the corresponding segmentation masks (see an example image and related annotations in Fig. 3). For this paper, we used the images and annotations from Task 1 of the 2018 challenge: it includes 2,594 training images, 100 validation images, and 1,000 test images.

B. Evaluation measure

Our experiments consider binary segmentation. In this case, the ground truth segmentation Sg_i for one image i may be represented by a pair $< Sg_i^m, Sg_i^b >$ where Sg_i^m is the set of pixels labeled as $melanoma$ and Sg_i^b is the set of pixels labeled as $background$. The automatic segmentation Sa_i for the same image i contains $< Sa_i^m, Sa_i^b >$ where Sa_i^m is the set of pixels classified as $melanoma$ and Sa_i^b is the set of pixels classified as $background$. The performance is evaluated using the Dice similarity coefficient, defined in equation (4):

$$Dice(Sg_i, Sa_i) = \frac{2|Sg_i^m \cap Sa_i^m|}{|Sg_i^m| + |Sa_i^m|} \tag{4}$$

The score is computed per image and averaged over the images in the dataset. It varies between 0 and 1, with a higher value indicating better segmentation performance.

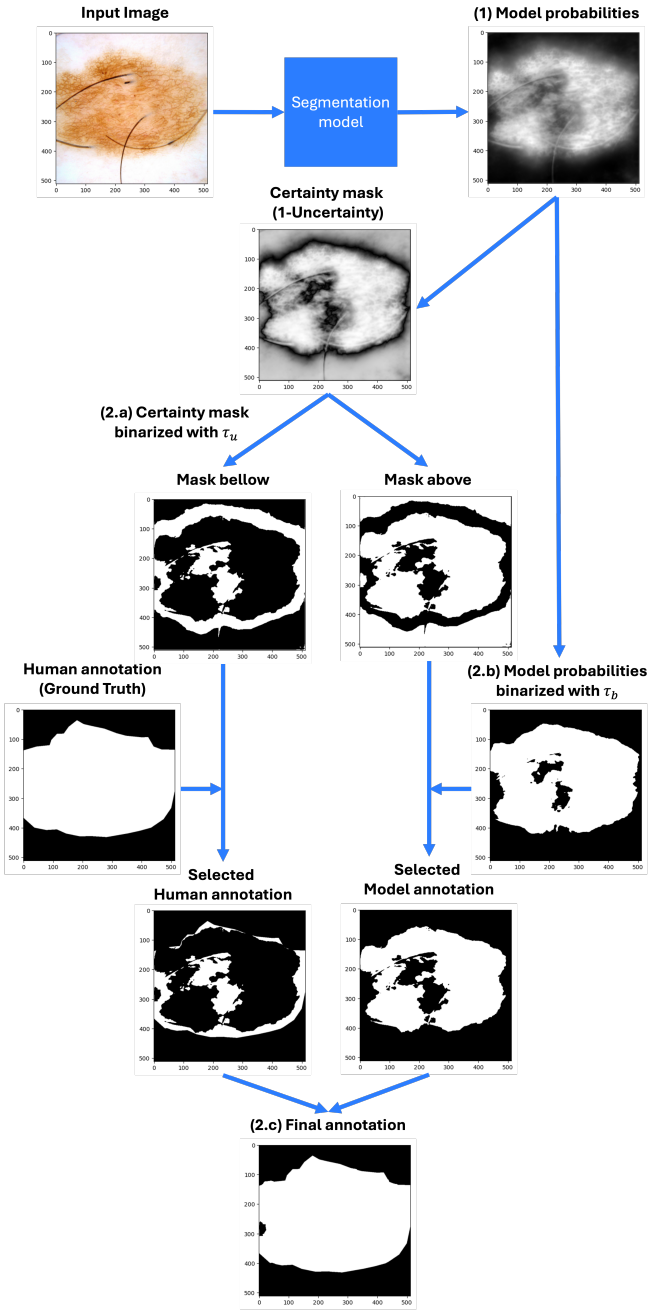

Fig. 2. Partial annotation principle

C. Network architecture

Several architectures [1] have been proposed for image segmentation, including Mask R-CNN [3] and U-Net [4]. U-Net is especially suitable for medical image segmentation as it delivers excellent performance even with limited data [32]. A modified version of U-Net has been chosen, but the proposed approach is model-agnostic. Compared to the original U-Net architecture, a ResNet-18 [33] backbone pre-trained on ImageNet [34] was used. Furthermore, before each increase in the number of filters in the U-Net architecture, 1x1 convolutions are incorporated, and batch normalization [35] is added after each convolution layer to enhance performance and generalization.

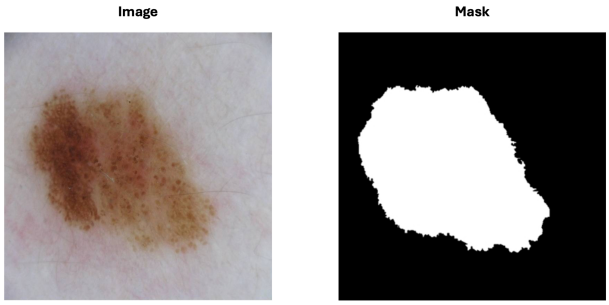

Fig. 3. One image and its annotations, from ISIC 2018 [30], [31] - Task 1 (ISIC_0000009)

D. Hyper-parameters

The following hyper-parameters were used to train the networks during the experiments:

- Image size: 512 x 512 pixels
- Batch size: 32
- Cold start[2]: 128 fully annotated images
- Number of most informative images selected at each AL iteration[3]: 32
- Number of AL iteration[4]: 40
- Data augmentation:
  - Random rotation ($\pm$ 30%)
  - Random horizontal flip (50% of the time)
  - Random vertical flip (50% of the time)
  - Random Color Jitter: Brightness ($\pm$ 20%), Contrast ($\pm$ 20%), Saturation ($\pm$ 20%)
  - Random Blur: σ_{min}=0.01, σ_{max}=0.2, kernel size=5
  - Normalization [5]: $Mean_{Red,Green,Blue}$ = (0.485, 0.456, 0.406) $SD_{Red,Green,Blue}$ = (0.229, 0.224, 0.225).
- Learning rate: .01
- Number of epochs (per AL iteration): 10
- Optimizer: Stochastic gradient descent with learning rate=0.01, momentum=0.9, and weight decay=0.0001

All the experiments were carried out on a NVIDIA A6000 48 Go.

E. Simulated active learning

Comparing AL methods in a prospective way (i.e., actually asking an expert to annotate the selected images) is problematic because the selection will influence future item selection and, therefore, the results. To compare the approaches, similar to [36], we simulate the active learning process using the fully annotated dataset of melanoma ISIC 2018 [30], [31].

[2]Number of randomly selected images to initialize model training.
[3]To avoid incomplete batch size, we select the same size as batch size.
[4]To reduce the computational cost of the experiments, we limited the number of AL iterations to around 50% of the full dataset.
[5]Values used for the pretraining on Imagenet.

TABLE I
COMPARISON OF UNCERTAINTY ESTIMATION METHODS AT IMAGE-LEVEL (DICE SCORE IN %)

AL Method	Val. dataset[7]	Test dataset
Random selection	87	85
Full dataset	84	83
Least confidence	88	87
Entropy	87	86
MC Dropout	86	85
ALS+UNET [23]	-	81
SA+AS [22]	-	86

At each AL iteration, the model was fine-tuned using all available annotated data (previously and newly annotated images), in accordance with the survey by Budd et al. [10].

V. RESULTS

At each iteration, we selected the 32 most uncertain images to annotate based on the described uncertainty metrics. Using these newly annotated data, we fine-tuned the model, calculated the segmentation performance on the validation dataset (100 images), and then repeated the process until the annotation budget was reached. Finally, we evaluated the last trained model on the test dataset (1000 images).

A. Comparison of uncertainty estimation methods at image-level

We first compared the selection methods for data annotation at the image level. We evaluated three AL methods (Least Confidence, Entropy, Monte Carlo Dropout) and two baselines:

- Full dataset: the model was trained on the fully annotated training dataset. For a fair comparison, the same number of iterations as for 40 AL iterations was used: 8140[6].
- Random selection: the images to annotate were randomly selected at each AL iteration.

The Dice scores are presented in Table I. The evolution of segmentation performance over the AL iterations is presented in Fig. 4.

The least confidence method showed the best performance. It resulted in a 4.8% improvement on the test dataset compared to training on the fully annotated dataset, and a 2.4% improvement compared to random selection.

B. Combination of human and model annotation

Based on the best selection method at the image level (i.e., least confidence), we evaluated the segmentation performance based on the combination of human annotation for the most uncertain pixels and model prediction for the others during training, following the principle depicted in Fig. 2. Apart from the performance of the model, the distribution of annotation between the expert and the model is determined by two parameters, τ_u and τ_b:

[6]At each AL iteration, the model is trained for 10 epochs. The number of iterations per epoch increased at each AL iteration, as the number of training samples increased. The number of iterations can be considered cumulative as the model weights adjusted in the prior AL iteration are reused for the next AL iteration training.

[7]Performance on the validation dataset at the last AL iteration.

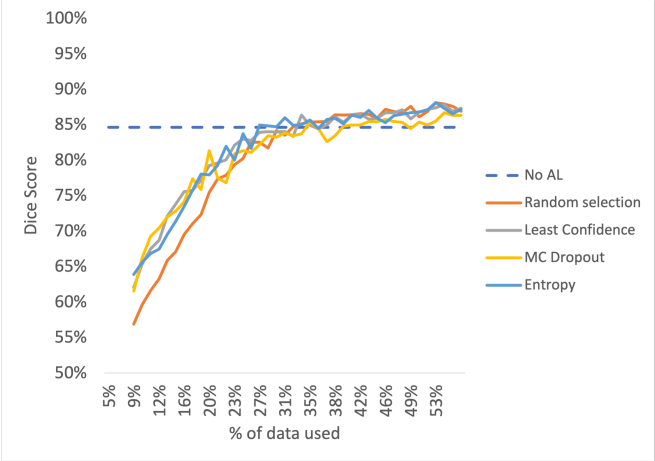

Fig. 4. Dice score over number of annotated images - Full annotation

- τ_u is the uncertainty threshold, determining from which probability value the model's prediction is considered certain.
- τ_b is the threshold for binarizing model predictions, determining the probability value at which the prediction switches from one class to another.

We evaluated three values for each parameter. The higher the value, the higher the confidence requirement of the model in the prediction:

- τ_u: 0.125, 0.25, 0.375
- τ_b: 0.25, 0.5, 0.75

To reduce computational cost and training time, we limited the number of AL iterations to 20. Moreover, to estimate the reduction in the number of pixels to be annotated by combining human annotation and model predictions, we calculated the percentage of pixels that the user should have annotated over the AL process. We distinguished between positive pixels (which require annotation by humans) and all pixels (both negative and positive). For a fair comparison, we also calculated the percentage of positive pixels annotated in the full training dataset, which is around 21%.

The results are presented in Table II. The more we increase the certainty threshold (τ_u), the more surface area the expert will have to analyze and annotate, but this does not improve performance. Conversely, reducing the uncertainty threshold will speed up annotation time but reduce performance. The binarization threshold (τ_b) does not seem to have a significant influence. Furthermore, as shown in Fig. 5, during the first iterations, the expert should positively annotate around 10% of the images, then the percentage drastically drops to less than 5% to positively annotate.

VI. DISCUSSION

In this paper, we explored several methods of active learning applied to melanoma segmentation. For this purpose, we firstly compared various selection methods, including random selection, least confidence, entropy, and Monte Carlo dropout.

TABLE II
DICE SCORES AND PERCENTAGE OF ANNOTATED PIXELS ACCORDING TO CERTAINTY (τ_u) AND BINARIZATION (τ_b) THRESHOLDS

τ_u	τ_b	Dice (%)	Mean % of annotated pixels	Mean % of pixels to verify/annotate
0.125	0.25	80	2	12
	.50	81	5	11
	.75	81	3	11
.25	0.25	84	5	19
	0.50	83	6	18
	0.75	83	5	18
0.375	0.25	84	9	32
	0.50	82	9	32
	0.75	82	10	32
Full dataset		83	21	100

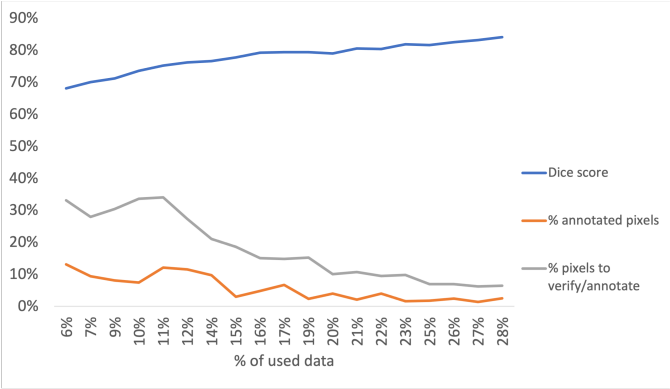

Fig. 5. Dice score and Percentage of pixels to be annotated during the AL process - τ_u=0.25 & τ_b = 0.25

The results showed that random selection is challenging to surpass, as also noted by [16]: Only the Least Confidence and Entropy methods outperformed random selection.

Based on the most effective approach (i.e., least confidence), we showed that the combination of model and human annotation significantly reduces the need for expert annotation. After a few iterations, less than 5% of the pixels needed to be annotated. Over all 20 AL iterations, the expert only needed to annotate an average of 5% of the pixels compared to 21% without our approach. Additionally, they only needed to analyze 19% of the pixels (including both melanoma and background pixels), compared to 100% of the pixels in the standard approach.

While this approach allows for performance similar to training on a fully annotated dataset, it remains inferior to the segmentation performance achieved with a least confidence selection where all pixels in the selected images are annotated. This could be due to poor segmentation performance or overconfidence of the model during the initial AL queries, leading to poor quality annotations for part of the training dataset. Updating the initial (model) annotations using the more efficient segmentation model obtained through the AL process could be explored, particularly by studying when to carry out these re-annotations (e.g., performance improvement threshold on a validation dataset). Such an approach could indeed improve performance, but at the expense of longer training time.

The experiments reported here rely on human annotation using annotations available from a fully annotated dataset. It would be interesting to evaluate, under real annotation conditions with experts, the time saved using this pseudo-annotation method. We could also compare the benefit of providing pseudo-annotations even for uncertain areas.

Several limitations can be pointed out. First, the performance of this approach must be evaluated on other datasets. Second, the dataset used contains only one class (0: background; 1: melanoma), limiting the annotation complexity and the segmentation problem. Finally, although the proposed approach is model-agnostic, it has only been evaluated with the U-Net architecture.

VII. CONCLUSION

Building a good medical segmentation tool using deep learning is often limited by the availability of annotated images. When these images do exist, they are usually few in number and not fully annotated, primarily because annotation must be done by medical specialists and is very time-consuming. In this paper, we explored Active Learning methods to reduce both the number of images to annotate and the number of pixels to annotate per image. In the context of melanoma segmentation, we showed that it is possible to obtain similar results by annotating only 30% of the images and by annotating less than 5Hence, this paper paves the way to facilitate the segmentation of medical images in fields where data is available but not annotated. Future work will aim to extend this approach to more complex medical data, such as 3D images.

ACKNOWLEDGMENT

This work has been supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

REFERENCES

[1] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, arXiv: 1702.05747.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870 [cs]*, Jan. 2018, arXiv: 1703.06870.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015, arXiv: 1505.04597.

[5] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation," *arXiv:1706.04737 [cs]*, Jun. 2017, arXiv: 1706.04737.

[6] B. Settles, "Active learning literature survey," 2009.

[7] D. Yoo and I. S. Kweon, "Learning Loss for Active Learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 93–102.

[8] X. Zeng, L. Wen, Y. Xu, and C. Ji, "Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models," *Computer methods and programs in biomedicine*, vol. 197, p. 105700, 2020.

[9] L. Dao and N. Q. Ly, "A Comprehensive Study on Medical Image Segmentation using Deep Neural Networks," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023.

[10] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, Jul. 2021.

[11] Q. Liang, H. Qin, H. Zeng, J. Long, W. Sun, D. Zhang, and Y. Wang, "Active Learning Integrated Portable Skin Lesion Detection System Based on Multimodel Fusion," *IEEE Sensors Journal*, vol. 23, no. 9, pp. 9898–9908, May 2023, conference Name: IEEE Sensors Journal.

[12] Z. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M. E. Celebi, and G. Hamarneh, "A survey on deep learning for skin lesion segmentation," *Medical Image Analysis*, vol. 88, p. 102863, Aug. 2023.

[13] B. Settles, "Active Learning Literature Survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[14] S. Mittal, J. Niemeijer, J. P. Schäfer, and T. Brox, "Best Practices in Active Learning for Semantic Segmentation," Mar. 2023, arXiv:2302.04075 [cs].

[15] S. A. Golestaneh and K. M. Kitani, "Importance of Self-Consistency in Active Learning for Semantic Segmentation," Aug. 2020, arXiv:2008.01860 [cs].

[16] M. Gaillochet, C. Desrosiers, and H. Lombaert, "Active learning for medical image segmentation with stochastic batches," *Medical Image Analysis*, vol. 90, p. 102958, 2023.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] S. Mittal, J. Niemeijer, J. P. Schäfer, and T. Brox, "Best Practices in Active Learning for Semantic Segmentation," in *Pattern Recognition*, U. Köthe and C. Rother, Eds. Cham: Springer Nature Switzerland, 2024, pp. 427–442.

[19] X. Li, M. Xia, J. Jiao, S. Zhou, C. Chang, Y. Wang, and Y. Guo, "HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image segmentation," *Medical Image Analysis*, vol. 88, p. 102862, Aug. 2023.

[20] V. Nath, D. Yang, B. A. Landman, D. Xu, and H. R. Roth, "Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2534–2547, Oct. 2021.

[21] N. Martin, J.-P. Chevallet, and G. Quénot, "Segmenting partially annotated medical images," in *International Conference on Content-based Multimedia Indexing*. Graz Austria: ACM, Sep. 2022, pp. 111–115.

[22] X. Shi, Q. Dou, C. Xue, J. Qin, H. Chen, and P.-A. Heng, "An Active Learning Approach for Reducing Annotation Cost in Skin Lesion Analysis," Sep. 2019, arXiv:1909.02344 [cs, eess, stat].

[23] X. Shu, Y. Yang, R. Xie, J. Liu, X. Chang, and B. Wu, "Als: Active Learning-Based Image Segmentation Model for Skin Lesion," Rochester, NY, Jun. 2022.

[24] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i Nieto, "Cost-Effective Active Learning for Melanoma Segmentation," Nov. 2017, arXiv:1711.09168 [cs].

[25] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," Oct. 2016, arXiv:1511.02680 [cs].

[26] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," Oct. 2016, arXiv:1506.02142 [cs, stat].

[27] Y. Zhang, J. Chen, X. Ma, G. Wang, U. A. Bhatti, and M. Huang, "Interactive medical image annotation using improved Attention U-net with compound geodesic distance," *Expert Systems with Applications*, vol. 237, p. 121282, Mar. 2024.

[28] J. Guo, Q. Wang, S. Su, and Y. Li, "Informativeness-guided active learning for deep learning–based façade defects detection," *Computer-Aided Civil and Infrastructure Engineering*, p. mice.12998, Mar. 2023.

[29] I. C. Saidu and L. Csató, "Active Learning with Bayesian UNet for Efficient Semantic Image Segmentation," *Journal of Imaging*, vol. 7, no. 2, p. 37, Feb. 2021.

[30] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Apr. 2018, pp. 168–172, iSSN: 1945-8452.

[31] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, p. 180161, Aug. 2018, publisher: Nature Publishing Group.

[32] O. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," *arXiv:1606.06650 [cs]*, Jun. 2016, arXiv: 1606.06650.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[35] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Feb. 2015, arXiv: 1502.03167.

[36] S. Ayache and G. Quénot, "Video Corpus Annotation Using Active Learning," in *Advances in Information Retrieval*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, vol. 4956, pp. 187–198, series Title: Lecture Notes in Computer Science.