# Active Learning Strategies on a Real-World Thyroid Ultrasound Dataset

Hari Sreedhar, Guillaume P R Lajoinie, Charles Raffaelli, Hervé Delingette

# Active Learning Strategies on a Real-World Thyroid Ultrasound Dataset

Hari Sreedhar[1,2][0000−0002−9725−4984], Guillaume P.R. Lajoinie[3][0000−0002−8226−7301], Charles Raffaelli[2], and Hervé Delingette[1][0000−0001−6050−5949]

[1] Centre Inria d'Université Côte d'Azur, 06902 Sophia Antipolis Cedex, France
hari.sreedhar@inria.fr
[2] Centre Hospitalier Universitaire de Nice, 06000 Nice, France
[3] University of Twente, Techmed Center for Technical Medicine, 7522 NB Enschede, The Netherlands

**Abstract.** Machine learning applications in ultrasound imaging are limited by access to ground-truth expert annotations, especially in specialized applications such as thyroid nodule evaluation. Active learning strategies seek to alleviate this concern by making more effective use of expert annotations; however, many proposed techniques do not adapt well to small-scale (i.e. a few hundred images) datasets. In this work, we test active learning strategies including an uncertainty-weighted selection approach with supervised and semi-supervised learning to evaluate the effectiveness of these tools for the prediction of nodule presence on a clinical ultrasound dataset. The results on this as well as two other medical image datasets suggest that even successful active learning strategies have limited clinical significance in terms of reducing annotation burden.

**Keywords:** Thyroid cancer · Active learning · Ultrasound imaging.

## 1 Background

Thyroid nodules are growths disrupting the normal follicular architecture of the thyroid gland, whose evaluation by ultrasound is essential to facilitate thyroid cancer detection and avoid unnecessary interventions. Ultrasound is ideally suited to this task because it is inexpensive and non-invasive, but this technique is limited by the subjective interpretation of the acquired images by the practitioner.

In response to these limitations, many groups have proposed machine learning algorithms to automate thyroid ultrasound evaluation. These approaches apply neural networks to standard B-mode ultrasound images to perform detection and segmentation of nodules [4], benign-malignant classification [1], or combined strategies to reproduce the entire clinical evaluation task [20, 8, 12]. A few groups have even begun to test commercial software for this purpose [3, 17].

As these algorithms are tested and validated for clinical use, they must follow training strategies that respect the limitations inherent to ultrasound imaging. Especially when adapting to ultrasound systems in specific hospital centers,

high-quality annotations drawn by practitioners specifically experienced in thyroid ultrasound are essential; however, the time of these experts is inherently expensive and annotation tasks have a low priority in the patient-oriented workflow. Clinical implementation of these tools will therefore depend on training strategies that make intelligent use of ground truth labels.

### 1.1 Active Learning

This is where active learning holds promise, as a means of efficiently utilizing expert annotations. This approach to machine learning is based on the premise that, for a large pool of unlabeled data, there may exist a smaller subset of observations which would be as effective for supervised learning as the entire image pool. In terms of medical image analysis, this means starting with a collection of unlabeled images, with only a small initial subset selected at random to be annotated by an expert radiologist. This subset of labeled images is used for supervised learning, though the unlabeled images may be used for semi-supervised learning of either the task or of feature representations [6, 13].

Based on the performance of the algorithm trained on this initial labeled set, additional images are selected for annotation. In the context of radiology, this is typically through a pool-based sampling approach in which some criterion guides selection from among the remaining unlabeled images. Once additional images are selected, the algorithm is retrained, and the cycle is repeated (see Fig. 1) [2].

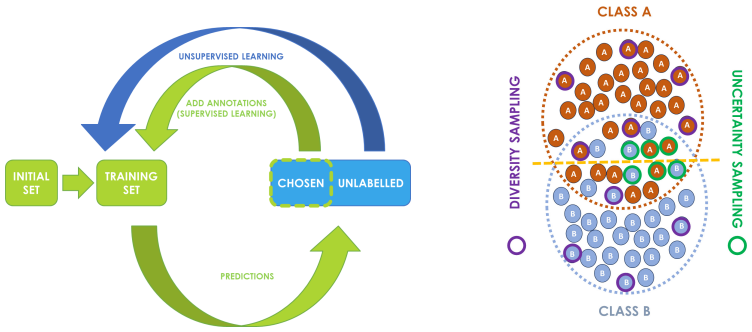


**Fig. 1.** (Left): The basic cycle of pool-based active learning: an initial set of images is randomly chosen for annotation, and used for training. In subsequent iterations, further images are chosen for annotation from the unlabeled image pool to retrain the algorithm. The unlabeled images can also be used for semi-supervised strategies. (Right): The two main categories of active learning criteria: uncertainty and diversity.

The criteria for selecting images for annotation vary between strategies. The most commonly considered criterion is uncertainty, i.e. selecting cases in which the algorithm's predictions are uncertain in order to improve its performance [2, 11]. Relying solely on this measure, however, risks overrepresenting a subset of

cases, rather than the entire distribution of images. Therefore, diversity strategies seek to include images dissimilar to each other or to already-labeled images, to prioritize the "representativeness" of the selected instances (see Fig. 1) [18, 14].

Whichever specific strategy is chosen, active learning translates logically to the analysis of ultrasound images, because of the cost of manual annotation by expert radiologists. Zhou et al. demonstrated this by combining active learning with transfer learning to fine-tune a convolutional neural network for carotid intima-media thickness interpretation [21]. More recently, Huang et al. proposed a framework for segmentation of breast and knee cartilage ultrasound that combined active learning criteria with semi-supervised learning to better adapt to different ultrasound datasets, along with an uncertainty selection strategy modified to avoid redundant image selection [6].

Despite these advances, many active learning strategies struggle to outperform the baseline of randomly selecting images for annotation [9]. Gaillochet et al., applying active learning to MRI images, addressed this problem with a stochastic batch selection strategy to harness the power of random sampling on small-scale datasets [5]. These examples call into question the feasibility of practical implementation of active learning strategies in a clinical context.

### 1.2 Active Learning Applied to Thyroid Ultrasound

With this in mind, we have applied active learning on a clinical dataset of ultrasound images. Since clinical thyroid images are not always acquired following standardized protocols (as is often the case for AI studies), we have chosen to assess the potential of active learning techniques on these unmodified, real-life examples. We present therefore an example of binary classification of the presence or absence of thyroid nodules in these images with the following contributions:

1) A novel and simple weighted selection active learning strategy to respect the representative power of random selection with small annotation budgets.

2) A real-world implementation adapted to the difficulties of learning on an actual clinical ultrasound dataset, including using semi-supervised feature extraction to facilitate active learning strategies. The results are assessed with a higher number of repetitions than is typically tested [5, 13, 19] to ensure statistical relevance.

## 2 Materials and Methods

### 2.1 Image Datasets

Ultrasound images for the study were collected from the stored images of thyroid examinations conducted in the course of routine clinical practice by radiologists at the Centre Hospitalier Universitaire de Nice from August 2021 to June 2022. All scans had been acquired on a Siemens S3000 ultrasound system (Siemens Healthineers, Erlangen, Germany) in accordance with standard practice for our institution. All images from ultrasound examinations of the thyroid were exported in DICOM format and de-identified. The images were then automatically

filtered to include only B-mode images with no Doppler or elastography overlays. Finally, images were filtered to only include those in axial views, with recognizable anatomical landmarks of the trachea or the carotid vessels. The resulting 1048 images from 269 patients were then annotated by a non-expert reader, who manually segmented solid, cystic, and mixed solid and cystic nodules. Spongiform lesions were excluded. These annotations, examples of which can be seen in Supplementary Figure 1, were then converted into equivalent labels of nodule presence (602 images) or absence (446 images).

**External Datasets** Given the non-expert annotations and potential difficulties of learning from our dataset, we conducted equivalent tests on two public medical imaging datasets randomly downsampled to an equivalent size. The PneumoniaMNIST dataset contains pediatric chest X-ray images with labels for pneumonia vs normal binary classification [7]. The BreaKHis dataset contains histopathological images in the context of breast cancer, with labels for benign and malignant diagnoses [15].

### 2.2 Rigged Draw Strategy

Inspired by Gaillochet et al., we sought to harness the power of random selection to represent a small dataset [5]. In order to do this while controlling the relative contribution of the uncertainty criterion, we proposed a weighted selection strategy called rigged draw. In this strategy, the relative weight $w_n$ for selecting any sample in an active learning round is:

$$w_n(\alpha) = 1 + \alpha \frac{c_n}{c_{90}} \tag{1}$$

where $c_n$ is the value of the uncertainty-based criterion for the $\mathrm{n}^{\mathrm{th}}$ sample, $c_{90}$ is the $90^{\mathrm{th}}$ percentile value of the criterion across all unlabeled images, and $\alpha$ is a factor weighting the importance of the uncertainty criterion relative to random selection. The choice to normalize relative to the $90^{\mathrm{th}}$ percentile was to avoid the effects of outlier maximum values with certain selection strategies.

### 2.3 Supervised and Unsupervised Active Learning Strategies

We tested supervised learning using only labeled images with a ResNet18 pretrained on natural images. We compared random selection, LeastConfidence (an uncertainty strategy), and KMeans (a diversity strategy) as implemented in Zhan et al. [19, 16, 10]. We also tested rigged draw sampling, defining the uncertainty criterion $c_n$ as the positive entropy contribution of sample $n$:

$$c_n(p_n) = -p_n \log_2(p_n) \tag{2}$$

where $p_n$ is the probability of nodule presence as predicted by the network (between 0 and 1). With this choice, we would preferentially weight images with a predicted probability close to 0.5.

As suggested by Huang et al., learning from ultrasound data may be difficult for active learning strategies that begin with few labeled images [6]. We therefore also tested semi-supervised learning using the network architecture proposed by Shui et al. for their two-stage WAAL active learning strategy [13]. This strategy depends on a network which conducts classification upon a feature representation which is in turn trained with a loss function seeking to reduce the distance between labeled and unlabeled images.

Our motivation for using this network was to imitate its approach to learning a useful feature representation from the images that would increase the effectiveness of active learning strategies. In addition to testing the entire WAAL strategy, this network structure was also used separately to test the previously mentioned active learning strategies.

## 3 Results

The active learning strategies were tested with both the supervised and semi-supervised strategies using the DeepAL+ toolkit from Zhan et al. [19]. For each test, a base set of 50 images was taken from a training set of 850 images and used to train the network for a fixed number of epochs (60), with subsequent batches of 50 being selected from among the unlabeled images, up to the maximum size of 750 images. A balanced test set on our dataset was established using 199 images from patients not represented in the training set (102 with nodules, 97 without); on the other two datasets test sets were slightly larger (624 for PneumoniaMNIST and 364 for BreaKHis, as noted in Supplementary Table 1). In order mitigate the effects of different starting sets and the stochastic nature of certain selection strategies, approximately 20 repetitions were used; as seen in Fig. 3, the starting set can create a high degree of variability in strategy performance.

The rigged draw strategy was tested using weights of $\alpha = 5$, $\alpha = 25$, and $\alpha = 50$ to give different importance to the uncertainty criterion during selection. The results with the most effective weight, $\alpha = 25$, are reported here, with the others given in Supplementary Tables 2 and 3.

### 3.1 Supervised Strategies

We used AUC under the ROC as a measure of classification performance independent of decision threshold. The median binary classification AUC values as a function of the cumulative active learning budget using the supervised strategy are given in Fig. 2, with the distributions of AUC values at different budgets for our strategy given in Fig. 3. The AUC values achieved with different budgets on ultrasound data also varied greatly with different starting sets (see Fig. 3)

The area under the budget curve (AUBC) values, calculated as the area under the curve of classification AUC value vs normalized cumulative budget (from 0 to 1), serves as a measure of the efficacy of the active learning strategies [19].

A summary of these AUBC values for the supervised strategies is given in Table 1. When the AUBC values from the repeated trials with the rigged draw strategy were compared to random selection, no statistically significant difference was found with the two-sample Kolmogorov-Smirnov test. In addition, the AUCs achieved at all budget sizes for the ultrasound dataset were substantially lower than those achieved on the PneumoniaMNIST and BreaKHis datasets (see Fig. 2).

**Table 1.** Supervised learning AUBC values. Values closer to 1 indicate a more effective strategy.

| Dataset | Test Set Size | Measure | Random | LeastCertain | KMeans | RiggedDraw |
|---|---|---|---|---|---|---|
| US Dataset | 199 | Mean | 0.643 | 0.642 | 0.641 | 0.639 |
| | | Median | 0.642 | 0.646 | 0.641 | 0.639 |
| | | STD | 0.010 | 0.011 | 0.009 | 0.012 |
| Pneumonia MNIST | 624 | Mean | 0.918 | 0.917 | 0.914 | 0.919 |
| | | Median | 0.917 | 0.917 | 0.916 | 0.920 |
| | | STD | 0.006 | 0.005 | 0.005 | 0.004 |
| BreaKHis | 364 | Mean | 0.832 | 0.828 | 0.826 | 0.832 |
| | | Median | 0.831 | 0.829 | 0.823 | 0.836 |
| | | STD | 0.015 | 0.020 | 0.017 | 0.022 |

### 3.2 Semi-supervised Strategies

For the semi-supervised strategies using the feature representation learned from all images, the median binary classification AUC values as a function of the cumulative active learning budget for each of the strategies and datasets are given in Fig. 2, with the distributions of AUC values at different budgets for our strategy given in Fig. 3. The AUBC values are reported in Table 2. When the AUBC values from the repeated trials with the rigged draw strategy were compared to random selection, a p-value of 0.0082 was found via the Kolmogorov-Smirnov test.

Performance for the rigged draw strategy improved substantially for the ultrasound images using the semi-supervised approach (see Fig. 3). However, the AUBC values for the rigged draw strategy, while greater than random selection, were not substantially different in terms of magnitude (see Table 2), and once again there was considerable variation in AUC values at each budget size (see Figure 3). In addition, for the PneumoniaMNIST and BreaKHis datasets, high AUC values were reached with very few images, and thus no meaningful differences could be observed between strategies (see Fig. 2).

## 4 Discussion

Overall, the results using supervised learning did not show a significant advantage for any active learning strategy compared to random selection on any of the

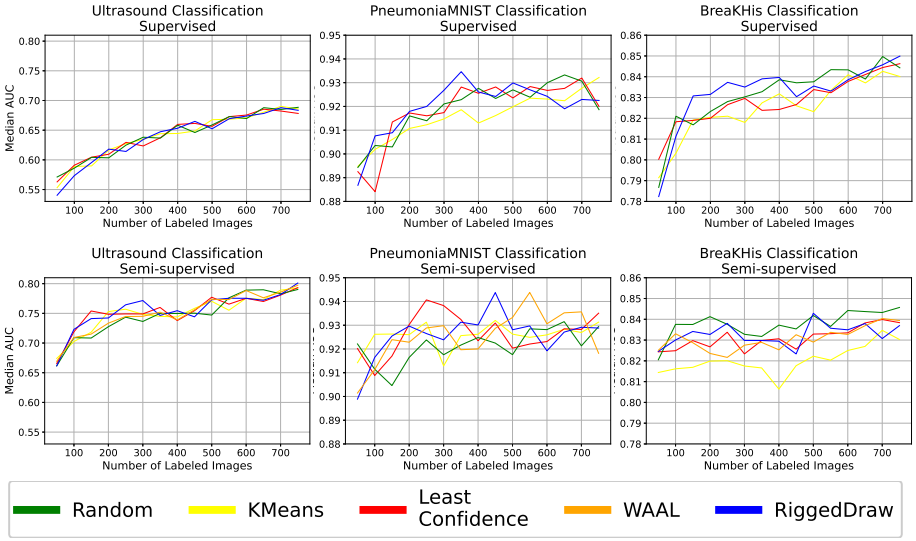


**Fig. 2.** Median AUC values for different active learning strategies on the three datasets. (Top Row): Supervised strategy. (Bottom Row): Semi-supervised strategy.

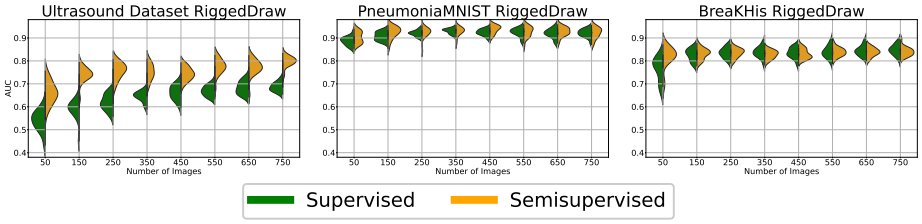


**Fig. 3.** Violin plots of classification AUC values on the at different label budgets with the rigged draw strategy.

**Table 2.** Semi-supervised learning AUBC values. Values closer to 1 indicate a more effective strategy, with * indicating p-values $< 0.05$ when compared to random selection

| Dataset | Measure | Random | LeastCertain | KMeans | WAAL | RiggedDraw |
|---|---|---|---|---|---|---|
| US Dataset | Mean | 0.747 | 0.751 | 0.749 | 0.751 | 0.754* |
| | Median | 0.748 | 0.752 | 0.750 | 0.751 | 0.755 |
| | STD | 0.009 | 0.008 | 0.009 | 0.008 | 0.007 |
| Pneumonia MNIST | Mean | 0.918 | 0.923* | 0.923 | 0.923 | 0.924* |
| | Median | 0.919 | 0.925 | 0.922 | 0.924 | 0.923 |
| | STD | 0.008 | 0.004 | 0.008 | 0.006 | 0.006 |
| BreaKHis | Mean | 0.836 | 0.828 | 0.823 | 0.831 | 0.833 |
| | Median | 0.841 | 0.830 | 0.820 | 0.830 | 0.834 |
| | STD | 0.017 | 0.026 | 0.022 | 0.017 | 0.018 |

datasets. In addition, classification performance on the ultrasound dataset was poorer than for the others; AUC improvement on the external datasets began to reach a plateau with budgets of only around 300 out of the total 750 images. This difference could be due to limitations inherent to the non expert annotations or the complexity of the classification task. It could also be related to the differences between our clinical ultrasound images and the public dataset images from different imaging modalities.

Performance on the ultrasound dataset was greatly improved, however, by a semi-supervised approach to learn a feature representation to reduce the distance between labeled and unlabeled images. Better results than were possible with the supervised network were attained with only 150 out of the total 750 images. This suggests that some degree of semi-supervised learning is preferable for training on image sets like ours; in an active learning scenario it makes prudent use of unlabeled data for which annotations are expensive.

The semi-supervised approach also showed a statistically significant advantage for the rigged draw strategy over random selection. This was not true of any of the other strategies tested on ultrasound data. However, the magnitude of the differences in classification AUC remained minimal, especially in light of the variability within each strategy. This is particularly important as we did test many repetitions of each strategy to compensate for the effects of different starting sets, unlike other comparisons which have used as few as 3 or 5 repetitions [19, 5, 13]. In light of the standard deviation of AUBC values as well as the range of AUC values at individual budget sizes, the impact of active learning on ultrasound data at this scale is unlikely to be clinically relevant.

It should be acknowledged that using non-expert annotations likely contributed to poor performance on our dataset. More specialized networks or pre-training on ultrasound images could also improve overall performance; however, this would not necessarily increase the relative advantage of active learning strategies. Rigorous optimization of the rigged draw strategy (such as the weight or the percentile for normalization) and of the annotation budget per round could have improved active learning results specifically; however, the need to fine-tune

strategies to this extent further suggests that they would not be suitable for real clinical thyroid ultrasound applications.

Therefore, at the scale of a thyroid ultrasound dataset from our clinical department, the benefits of existing active learning strategies appear to be limited. Semi-supervised approaches, and strategies like rigged draw that harness the power of random selection increase effectiveness; however, further refinement will be necessary to meaningfully reduce annotation burden. Future practical implementation will only be possible with more robust versions of these active learning tools that work consistently in a real hospital setting.

**Acknowledgements** The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 847581. G.L. acknowledges funding from the 4TU Precision Medicine program supported by High Tech for a Sustainable Future. G.L. also acknowledges funding by the European Union (ERC stg grant, Super-FALCON, project number 101076844).

## References


1. Buda, M., Wildman-Tobriner, B., Hoang, J.K., Thayer, D., Tessler, F.N., Middleton, W.D., Mazurowski, M.A.: Management of thyroid nodules seen on us images: Deep learning may match performance of radiologists. Radiology **292**(3), 695–701 (2019). https://doi.org/10.1148/radiol.2019181343
2. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis **71**, 102062 (2021). https://doi.org/10.1016/j.media.2021.102062
3. Chambara, N., Liu, S.Y.W., Lo, X., Ying, M.: Diagnostic performance evaluation of different ti-rads using ultrasound computer-aided diagnosis of thyroid nodules: An experience with adjusted settings. Plos One **16**(1) (2021). https://doi.org/10.1371/journal.pone.0245617
4. Chen, H., Song, S., Wang, X., Wang, R., Meng, D., Wang, L.: Lrthr-net: A low-resolution-to-high-resolution framework to iteratively refine the segmentation of thyroid nodule in ultrasound images. Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data p. 116–121 (2021). https://doi.org/10.1007/978-3-030-71827-5_15
5. Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. arXiv preprint arXiv:2301.07670 (2023)
6. Huang, K., Huang, J., Wang, W., Xu, M., Liu, F.: A deep active learning framework with information guided label generation for medical image segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1562–1567 (2022). https://doi.org/10.1109/BIBM55620.2022.9995046
7. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5) (2018). https://doi.org/10.1016/j.cell.2018.02.010
8. Lu, J., Ouyang, X., Liu, T., Shen, D.: Identifying thyroid nodules in ultrasound images through segmentation-guided discriminative localization. Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data p. 135–144 (2021). https://doi.org/10.1007/978-3-030-71827-5_18
9. Munjal, P., Hayat, N., Hayat, M., Sourati, J., Khan, S.: Towards robust and reproducible active learning using neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 223–232 (June 2022)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011), http://jmlr.org/papers/v12/pedregosa11a.html
11. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
12. Shen, X., Ouyang, X., Liu, T., Shen, D.: Cascaded networks for thyroid nodule diagnosis from ultrasound images. Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data p. 145–154 (2021). https://doi.org/10.1007/978-3-030-71827-5_19
13. Shui, C., Zhou, F., Gagné, C., Wang, B.: Deep active learning: Unified and principled method for query and training. In: Chiappa, S., Calandra, R. (eds.) Proceedings of the Twenty Third International Conference on Artificial Intelligence and

Statistics. Proceedings of Machine Learning Research, vol. 108, pp. 1308–1318. PMLR (26–28 Aug 2020), https://proceedings.mlr.press/v108/shui20a.html
14. Smailagic, A., Costa, P., Noh, H.Y., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P., et al.: Medal: Accurate and robust deep active learning for medical image analysis. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (2018). https://doi.org/10.1109/icmla.2018.00078
15. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering **63**(7), 1455–1462 (2016). https://doi.org/10.1109/TBME.2015.2496264
16. Wang, D., Shang, Y.: A new active labeling method for deep learning. In: 2014 International Joint Conference on Neural Networks (IJCNN). pp. 112–119 (2014). https://doi.org/10.1109/IJCNN.2014.6889457
17. Wei, Q., Zeng, S.E., Wang, L.P., Yan, Y.J., Wang, T., Xu, J.W., Zhang, M.Y., Lv, W.Z., Cui, X.W., Dietrich, C.F., et al.: The value of s-detect in improving the diagnostic performance of radiologists for the differential diagnosis of thyroid nodules. Medical Ultrasonography **22**(4), 415–423 (2020). https://doi.org/10.11152/mu-2501
18. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 Lecture Notes in Computer Science p. 399–407 (2017). https://doi.org/10.1007/978-3-319-66179-7_46
19. Zhan, X., Wang, Q., Huang, K.h., Xiong, H., Dou, D., Chan, A.B.: A comparative survey of deep active learning. arXiv preprint arXiv:2203.13450 (2022)
20. Zhang, Y., Lai, H., Yang, W.: Cascade unet and ch-unet for thyroid nodule segmentation and benign and malignant classification. Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data p. 129–134 (2021). https://doi.org/10.1007/978-3-030-71827-5_17
21. Zhou, Z., Shin, J., Feng, R., Hurst, R.T., Kendall, C.B., Liang, J.: Integrating active learning and transfer learning for carotid intima-media thickness video interpretation. Journal of Digital Imaging **32**(2), 290–299 (2019). https://doi.org/10.1007/s10278-018-0143-2