IBS: Interactive Batch Selection for Medical Image Segmentation

Bernhard Föllmer[1][0000−0003−1377−3371], Kenrick Schulze[2], Marc Dewey[3], and Hervé Delingette[1]

[1] Inria, Université Côte d'Azur, Epione team, Sophia-Antipolis, France
[2] Department of Radiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, 10117, Germany
[3] Department of Radiology, University of Cambridge, British Heart Foundation (BHF) Centre of Research Excellence, Victor Phillip Dahdaleh Heart and Lung Research Institute (HLRI), Cambridge University Hospitals, Royal Papworth Hospital, Cambridge, UK
bernhard.follmer@inria.fr

Abstract. Voxel-wise annotation for medical image segmentation is time-consuming and requires substantial expert effort. Active learning (AL) reduces annotation costs by iteratively selecting informative samples from an unlabeled pool for annotation and model training. Existing batch-selection strategies typically rely on automated acquisition criteria that balance uncertainty, diversity, and representativeness. However, their effectiveness varies across datasets, annotation budgets, model architectures, training procedures, and AL rounds, making it difficult to design a generalizable acquisition criterion. Moreover, clinical priorities and expert-defined segmentation goals are often not explicitly incorporated, while fully automated strategies provide limited transparency to the annotator. We propose **IBS** - **I**nteractive **B**atch **S**election, an expert-guided AL framework that enables interactive adaptation of the acquisition strategy. IBS constructs a gradient-based similarity matrix between samples and allows experts to weight complementary informativeness criteria, including uncertainty, affinity to positive reference samples, repulsion from negative reference samples, and subgroup stratification. The similarity matrix and expert-weighted informativeness scores are combined in a fixed-size determinantal point process to select a diverse and informative annotation batch. Across three medical image segmentation tasks, IBS achieves competitive or improved labeling efficiency compared with conventional AL baselines while enabling transparent, expert-controllable batch selection.

Keywords: Active Learning · Medical Image Segmentation · Human-AI Interaction

1 Introduction

Medical image segmentation requires large amounts of pixel-wise expert annotation, which is time-consuming and costly, particularly for complex segmentation

tasks. Active learning (AL) aims to reduce this burden by iteratively training a model on a small labeled set and selecting informative samples from the unlabeled pool for annotation. By prioritizing samples expected to improve model performance, AL can reduce the amount of voxel-wise annotation required for model development. Most AL methods estimate sample informativeness using heuristics based on uncertainty, diversity, or representativeness [19]. However, no single strategy consistently performs best across datasets and experimental settings [16], and random or stratified random sampling often remain strong baselines [3]. A central challenge is therefore to balance uncertainty, diversity, representativeness, and task-specific priorities for each dataset, model, training procedure, AL round, and initialization. Uncertainty-based methods select samples for which the model is least confident, using predictive entropy [20] or approximate Bayesian estimates from Monte Carlo dropout [1], Bayesian neural networks [9], and deep ensembles [4]. However, purely uncertainty-driven selection can produce redundant batches or emphasize outliers. Diversity- and representativeness-based methods address this by selecting samples that better cover the data distribution [18], while hybrid methods combine multiple criteria. For example, USIM [7] uses submodular subset selection to promote uncertainty, diversity, and representativeness, whereas stochastic batch acquisition combines uncertainty with randomized batch selection [13]. Targeted AL further incorporates task-specific objectives into sample selection [6, 8, 14]. For 3D biomedical segmentation, *nnActive* provides a systematic evaluation framework [16], but focuses on fully automated acquisition strategies and does not support interactive expert adaptation. In this work, we investigate whether interactive batch selection can improve AL efficiency while enabling expert guidance and increasing the interpretability of sample selection.

We propose **IBS**-**I**nteractive **B**atch **S**election, an expert-guided AL framework for medical image segmentation. IBS allows experts to steer a batch-selection process through uncertainty weighting, reference-set affinity and repulsion, and subgroup-level priorities. In contrast to prior AL methods with predefined acquisition criteria, IBS exposes the components of the acquisition function during batch selection, enabling the annotator to adapt the selection strategy to current model failure modes, anatomical targets, and dataset-specific challenges.

2 Method

Let $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ denote the full dataset of image patches extracted from the training images, where each patch $\boldsymbol{x}_i \in \mathbb{R}^{h \times w \times d}$ corresponds to a 2D or 3D image patch. The unlabeled pool is denoted by $\mathcal{X}_U \subseteq \mathcal{X}$ with $N_U = |\mathcal{X}_U|$ patches. The labeled set is given by $\mathcal{X}_L = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_L}$, where $\boldsymbol{y}_i$ denotes the annotation mask for patch $\boldsymbol{x}_i$. In each active learning round, the model trained in the previous round predicts pseudo-labels for the unlabeled patches. These pseudo-labels are used as targets to compute loss gradients, a subset of which forms a gradient embedding for each candidate patch. Pairwise candidate similarities are then calculated using the cosine similarity between these embeddings. For each patch, an

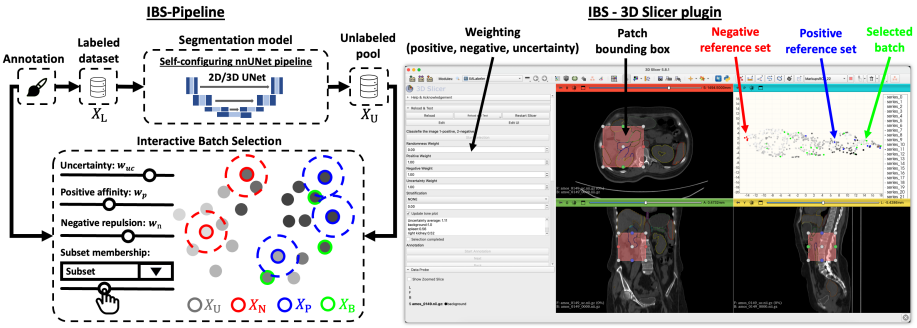

Fig. 1. Interactive batch selection pipeline. An initial nnU-Net-based segmentation model is trained on a small labeled dataset $\mathcal{X}_L$. From the unlabeled pool $\mathcal{X}_U$, the pipeline proposes a candidate annotation batch $\mathcal{X}_B \subset \mathcal{X}_U$. The expert can iteratively refine the batch by adjusting uncertainty weights w_{uc}, selecting positive and negative reference sets $\mathcal{X}_P$ and $\mathcal{X}_N$, weighting affinity to positive references w_p and repulsion from negative references w_n, and controlling subgroup membership weights $w_{sub,m}$.

expert-weighted informativeness score combines predictive uncertainty, affinity to expert-selected positive reference examples, repulsion from negative reference examples, and optional subset-specific priorities. The gradient-based similarities and informativeness scores are combined to construct the kernel of a fixed-size determinantal point process (k-DPP) [15], which favors batches that are both informative and diverse while reducing redundancy. Through an interactive 3D Slicer-based interface, annotators can inspect candidate patches, adjust the selection weights, define positive and negative reference sets, and iteratively refine the proposed batch before annotation.

2.1 Batch Selection with Fixed-Size Determinantal Point Processes

A fixed-size determinantal point process (k-DPP), introduced by Kulesza et al. [15], is a subset-selection method that uses a similarity kernel to select exactly k samples while favoring informative and mutually diverse subsets. We use k-DPP sampling to select an annotation batch $\mathcal{X}_B \subset \mathcal{X}_U$ with $|\mathcal{X}_B| = k$. The k-DPP kernel $\boldsymbol{L}$ is constructed as an L-ensemble by reweighting a positive semidefinite similarity matrix $\boldsymbol{S}_{\text{PSD}}$ using the vector of patch-wise informativeness scores $\boldsymbol{s} \in \mathbb{R}^{N_U}$. The score vector $\boldsymbol{s}$ contains expert-weighted informativeness scores for all candidates in the unlabeled pool. The kernel is defined as

$$\boldsymbol{L} = \text{diag}(\boldsymbol{s})\, \boldsymbol{S}_{\text{PSD}}\, \text{diag}(\boldsymbol{s}). \tag{1}$$

Here, $\boldsymbol{S}_{\text{PSD}}$ is obtained by symmetrizing the gradient-based cosine-similarity matrix and projecting it onto the PSD cone. Positive semidefiniteness ensures that $\boldsymbol{L}$ defines a valid DPP kernel. For a candidate batch B, the k-DPP assigns a probability proportional to $\det(\boldsymbol{L}_B)$, where $\boldsymbol{L}_B$ is the submatrix of $\boldsymbol{L}$ restricted

to the selected patches. The determinant captures both informativeness and diversity: high sample-wise scores increase the diagonal scaling of $\boldsymbol{L}_B$, whereas high pairwise similarity between selected patches reduces the determinant by making the corresponding feature vectors more redundant. Therefore, subsets with large determinant values contain patches that are individually informative while remaining mutually diverse in gradient-similarity space.

2.2 Gradient-based Similarity Matrix

Constructing a k-DPP kernel requires a pairwise similarity matrix over the candidate samples. Previous work has proposed using similarities between gradient embeddings for active learning [7]. Following this approach, we construct the gradient-based similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N_U \times N_U}$ by computing a gradient embedding for each candidate patch in the unlabeled pool $\mathcal{X}_U$. For each patch $\boldsymbol{x}_i$, a pseudo-label $\bar{\boldsymbol{y}}_i$ is obtained from the nnU-Net segmentation model trained in the current active learning round. Since nnU-Net-like models contain a large number of trainable network weights, we follow [7] and compute gradients only with respect to an informative subset ($\bar{\boldsymbol{\theta}} \subset \boldsymbol{\theta}$) of the segmentation model parameters, where $\boldsymbol{\theta}$ denotes all trainable weights of the network. The pairwise similarity between patches is then defined as the cosine similarity between their gradient embeddings:

$$S_{ij} = \frac{\boldsymbol{g}_i^\top \boldsymbol{g}_j}{\|\boldsymbol{g}_i\|_2 \, \|\boldsymbol{g}_j\|_2}, \qquad \boldsymbol{g}_i = \nabla_{\bar{\boldsymbol{\theta}}} \log p_{\boldsymbol{\theta}}(\bar{\boldsymbol{y}}_i \mid \boldsymbol{x}_i), \qquad i, j = 1, \ldots, N_U. \tag{2}$$

Thus, S_{ij} captures similarity in gradient direction, where similar gradients indicate patches expected to induce similar model updates, while dissimilar gradients may provide complementary information for active learning. Gradient embeddings and k-DPP selection build on prior work; our contribution is their integration into an interactive framework for active-learning batch selection.

2.3 Expert-weighted Informativeness Score

For each candidate $\boldsymbol{x}_i$, the informativeness score combines predictive uncertainty, expert feedback, and subset membership. Expert feedback is encoded through affinity to a positive reference set and repulsion from a negative reference set, while subset membership allows prioritization of predefined candidate groups. The expert adjusts the corresponding weights interactively during each AL round. To ensure positive scores for the L-ensemble kernel, we define:

$$s(\boldsymbol{x}_i) = \exp(w_{\mathrm{uc}} s_{\mathrm{uc}}(\boldsymbol{x}_i) + w_{\mathrm{p}} s_{\mathrm{p}}(\boldsymbol{x}_i) - w_{\mathrm{n}} s_{\mathrm{n}}(\boldsymbol{x}_i) + s_{\mathrm{sub}}(\boldsymbol{x}_i)) \tag{3}$$

Each component of the score is independently min–max normalized over the candidate set to prevent components with different numerical ranges from dominating the combined score.

Uncertainty score: We estimate $s_{\mathrm{uc}}(x_i)$ as the mean predictive standard deviation across classes using MC dropout with stochastic forward passes. Patches with higher prediction variability are considered more uncertain and are therefore assigned higher uncertainty scores.

Positive reference-set affinity and negative reference-set repulsion: To incorporate expert preferences into batch selection, the annotator may assign desirable patches to a positive reference set $\mathcal{P} \subseteq \mathcal{X}_U$. For each candidate patch $\boldsymbol{x}_i$, we define its positive affinity as the maximum similarity to any patch in $\mathcal{P}$. Conversely, patches that are undesirable, redundant, or should be deferred can be assigned to a negative reference set $\mathcal{N} \subseteq \mathcal{X}_U$. The corresponding negative repulsion score is defined as the maximum similarity to any patch in $\mathcal{N}$.

$$s_{\mathrm{p}}(\boldsymbol{x}_i) = \max_{\boldsymbol{x}_j \in \mathcal{P}} S_{\mathrm{PSD},ij}, \qquad s_{\mathrm{n}}(\boldsymbol{x}_i) = \max_{\boldsymbol{x}_j \in \mathcal{N}} S_{\mathrm{PSD},ij}. \tag{4}$$

If the positive or negative reference set is empty, we set the corresponding score to zero, i.e., $s_{\mathrm{p}}(\boldsymbol{x}_i) = 0$ for $\mathcal{P} = \emptyset$ and $s_{\mathrm{n}}(\boldsymbol{x}_i) = 0$ for $\mathcal{N} = \emptyset$. Thus, $s_{\mathrm{p}}(\boldsymbol{x}_i)$ encourages the selection of candidates similar to expert-preferred examples, whereas $s_{\mathrm{n}}(\boldsymbol{x}_i)$ discourages candidates similar to patches the annotator considers less informative or wishes to defer. Reference-set affinity and repulsion are computed using the same normalized similarity matrix used for batch selection.

Subset membership: Subset membership enables group-level prioritization and is conceptually distinct from the positive and negative reference sets. Reference sets are example-based: during interaction, the expert selects individual patches, such as outliers, and candidate patches are promoted or suppressed according to their similarity to these examples. In contrast, subsets are defined at the group level based on predefined candidate characteristics or model-derived attributes, such as the predicted foreground content. This allows the expert to prioritize entire groups of patches rather than candidates resembling a particular reference example. We denote the predefined subsets by $\{\mathcal{A}_m\}_{m=1}^{M}$, where $\mathcal{A}_m \subseteq \mathcal{X}_U$ and M is the number of subsets. For each candidate patch $\boldsymbol{x}_i$, the subset-membership score is defined as

$$s_{\mathrm{sub}}(\boldsymbol{x}_i) = \sum_{m=1}^{M} w_{\mathrm{sub},m}\, \mathbb{I}(\boldsymbol{x}_i \in \mathcal{A}_m)\,, \tag{5}$$

where $w_{\mathrm{sub},m}$ is the expert-defined weight assigned to subset $\mathcal{A}_m$, and $\mathbb{I}(\cdot)$ denotes the indicator function. Foreground-class imbalance represents a major challenge in active learning for medical image segmentation. In our experiments, we therefore define class-specific subsets based on the predicted foreground classes. A patch is assigned to subset $\mathcal{A}_c$ if its pseudo-label contains at least one voxel belonging to class c. Because a patch may contain multiple foreground classes, it may belong to more than one subset. These subsets allow the expert to prioritize patches containing underrepresented classes or classes for which the current model exhibits poor segmentation performance. Future work will investigate subsets defined using additional information, such as scanner type, imaging protocol, and clinical variables, including the presence of specific pathologies.

2.4 Interactive Batch Selection Interface

Interactive batch selection is implemented as an in-house 3D Slicer plugin [5] (Figure 1). In each AL round, the annotator can inspect 3D candidate patches in axial, sagittal, and coronal views through an interactive t-SNE visualization. The highlighted batch $\mathcal{X}_B$ can be refined by adjusting uncertainty, positive-affinity, negative-repulsion, and subset-membership weights or by assigning candidates to positive and negative reference sets. After each interaction, the informativeness scores and selected batch are immediately updated. Optional voxel-wise uncertainty maps and subset highlighting support localized uncertainty assessment and balanced selection. The expert adjusts the interaction weights based on the current predictions, pseudo-labels, and annotation objective. Increasing w_{uc} prioritizes uncertain patches, whereas reducing it emphasizes diversity through the k-DPP kernel. Positive reference examples may represent underrepresented anatomy, clinically relevant structures, or recurring model failures; increasing w_p promotes candidates with similar gradient-based representations. Conversely, negative references may identify artifacts, uninformative background, or outliers; increasing w_n suppresses similar candidates. Subset-membership weights $w_{\text{sub},m}$ impose group-level priorities, such as coverage of rare foreground classes, scanner protocols, pathologies, or other predefined subgroups. The interface therefore translates visual inspection of candidate patches, pseudo-labels, uncertainty maps, the t-SNE layout, and subgroup composition into interpretable adaptations of the acquisition function.

3 Experiments and Results

3.1 Experimental Settings and Baseline Methods

We evaluated IBS on three 3D medical image segmentation datasets: KiTS21 [11] for kidney and tumor segmentation, AMOS [12] for multi-organ abdominal segmentation, and ASOCA [10] for coronary artery segmentation. These datasets represent class-imbalanced, multi-organ, and thin-vascular-structure segmentation settings. We used the 3D full-resolution nnU-Net configuration and trained for five active learning (AL) rounds with 600 epochs per round. The initial labeled set was randomly sampled, and each round selected 25 patches, corresponding to approximately 60 minutes of annotation. Owing to the computational cost, we did not repeat the experiments with additional initial sets or random seeds. For IBS, one expert performed interactive batch selection for up to 10 minutes per round by inspecting candidate patches, pseudo-labels, uncertainty maps, and a t-SNE visualization and defining positive or negative reference sets. MC dropout uncertainty was estimated using $T = 5$ stochastic forward passes. To limit computational complexity, we randomly sampled 50,000 unlabeled patches and selected 500 candidates with probabilities proportional to their uncertainty. The k-DPP was applied to this reduced set. Larger candidate pools could be addressed using more restrictive prefiltering, low-rank or Nyström approximations, clustering-based reduction, or approximate k-DPP sampling. Unless modified by

the expert, weights were initialized as $w_{\mathrm{uc}} = w_{\mathrm{p}} = w_{\mathrm{n}} = 1.0$ and $w_{\mathrm{sub},m} = 0.0$. We report the area under the budget curve (AUBC) [21], area under the time curve (AUTC), final Dice score, and pairwise penalty matrix (PPM) [2]. AUBC measures labeling efficiency by selected patches, whereas AUTC accounts for estimated annotation time. For AUTC, we assumed 60 minutes of annotation per AL round and added 10 minutes per round for IBS to account for the additional interactive-selection time. We did not systematically evaluate the effect of different interaction-time budgets in this study; this will be addressed in future work. We compared IBS with random sampling (RANDOM), predictive entropy [19] (ENTROPY), MC-dropout mean standard deviation (MCD) [1], foreground-aware random sampling (RANDOM66) [16], and class-stratified predictive entropy (ClaSP PE) [17]. RANDOM66 allocated 66% of the patch-level annotation budget around randomly selected foreground regions, whereas ClaSP PE combines class-stratified predictive entropy with scheduled power noising.

3.2 Performance comparison

The performance improvement on all three datasets and active learning rounds as well as the pairwise penalty matrix (PPM) is shown in Figure 2. The pairwise penalty matrix summarizes the wins and losses between methods by aggregating the evaluation results over all three datasets and active learning rounds. IBS achieved the highest AUBC on AMOS and KiTS and the highest final Dice on AMOS, while achieving a similar final Dice to MCD on ASOCA, with slightly lower AUBC and AUTC. On KiTS, MCD achieved the highest final Dice, while IBS obtained the best budget- and time-aware performance. These results indicate that IBS is particularly beneficial when performance across the full AL trajectory is considered. Because annotation time was estimated rather than prospectively measured, AUTC should be interpreted as an approximate time-aware metric. MCD and ClaSP PE were the strongest automated baselines, with MCD performing particularly well on ASOCA and final KiTS Dice.

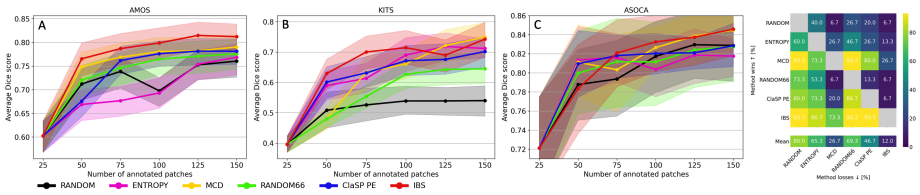

Fig. 2. Performance comparison of batch-selection methods across datasets. Panels (A–C) show the active learning performance curves for the AMOS, KiTS, and ASOCA datasets, respectively. Panel (D) shows the pairwise penalty matrix (PPM), which compares the proposed interactive batch-selection method, IBS, against baseline and recent active learning methods, including RANDOM, ENTROPY, MCD, RANDOM66, and ClaSP PE.

Table 1. Performance comparison of batch selection methods in terms of AUBC, AUTC, and final Dice after five active learning rounds across three datasets.

	AMOS			KiTS			ASOCA		
Query Method	AUBC	AUTC	Final Dice	AUBC	AUTC	Final Dice	AUBC	AUTC	Final Dice
RANDOM	0.6468	0.6468	0.7606	0.4593	0.4593	0.5401	0.7379	0.7379	0.8296
ENTROPY	0.6329	0.6329	0.7679	0.5403	0.5403	0.7184	0.7408	0.7408	0.8187
MCD	0.6761	0.6761	0.7908	0.5154	0.5154	**0.7495**	**0.7468**	**0.7468**	0.8449
RANDOM66	0.6611	0.6611	0.7742	0.4850	0.4850	0.6450	0.7407	0.7407	0.8275
ClaSP PE	0.6592	0.6592	0.7833	0.5393	0.5393	0.7020	0.7438	0.7438	0.8285
IBS	**0.6890**	**0.6859**	**0.8128**	**0.5670**	**0.5587**	0.7425	0.7448	0.7444	**0.8459**

Table 2. Ablation study of the proposed interactive batch-selection components on the AMOS dataset. Performance is reported in terms of AUBC, AUTC, and final Dice after five active learning rounds.

Ablation setting	AUBC	AUTC	Final Dice
IBS without interaction components	0.671	0.668	0.802
IBS with uncertainty only	0.681	0.679	0.807
IBS with uncertainty and reference sets	0.671	0.667	0.806
Full IBS	0.689	0.686	0.813

3.3 Ablation Experiments

We performed ablation experiments on AMOS to assess the contribution of the individual interactive batch-selection components. The non-interactive IBS variant controls for the effect of the gradient-based k-DPP kernel without expert adaptation. Allowing the expert to adjust only the uncertainty weight improved active learning performance compared with the non-interactive variant. Adding positive and negative reference sets without subset-membership weighting did not further improve AUBC or AUTC, suggesting that reference examples alone may over-constrain batch selection unless complemented by class- or subgroup-level guidance. The full IBS approach, combining uncertainty weighting, reference-set guidance, and subset-membership weighting, achieved the highest AUBC, AUTC, and final Dice.

4 Conclusion

We introduced IBS, an interactive batch-selection strategy combining uncertainty, reference-set guidance, and subset membership within a k-DPP acquisition kernel. In the evaluated runs, IBS achieved competitive or improved performance compared with automated baselines while providing transparent, expert-controllable selection. The evaluation was limited to one expert, one initialization, and estimated annotation times. Future studies should assess robustness across users, initializations, and interaction and annotation budgets using

prospectively measured times.
The source code is publicly available at https://github.com/Berni1557/IBS.

Acknowledgments. This research is part of the MediTwin project, which has been funded by the French government as part of "France 2030" plan.

Disclosure of Interests. Professor Marc Dewey received grant support from the British Heart Foundation (BHF; CS/F/25/190105), the German Centre for Cardiovascular Research (DZHK), the German Research Foundation (DFG; DE 1361/41-1, project number: 567888473) via the Global Cardiovascular Research Funders Forum (GCRFF) mechanism (GCRFF/24/270169) for the INCHARGE trial. He also received grant support from the FP7 Program of the European Commission for the randomized multicenter DISCHARGE trial (EC-GA 603266 in HEALTH.2013.2.4.2-2). He also received funding from The Academy of Medical Sciences (AMS) for the AMS Professorship (APR13\1005). Professor Marc Dewey also received grant support from DFG in the Heisenberg Program (DE 1361/14-1, DFG project number: 213705389), the graduate program on quantitative biomedical imaging (BIOQIC, GRK 2260/1, DFG project number: 289347353), for fractal analysis of myocardial perfusion (DE 1361/18-1, DFG project number: 392304398), the DFG Priority Programme Radiomics (DFG project number: 402688427) for the investigation of coronary plaque and coronary flow (DE 1361/19-1 [DFG project number: 428222922] and DE 1361/20-1 [DFG project number: 428223139] in SPP 2177/1), the GUIDE-IT project on data sharing of medical imaging trials (DE 1361/24-1 [DFG project number: 495697118]), the Quantitative Cardiovascular Imaging (QCI) meetings (DE 1361/44-1), the Future of Medical Imaging meeting (DE 1361/28-1), for AI-supported imaging of the aortic valve (DE 1361/40-1, DFG project number: 567210027), the investigation of coronary atherosclerosis using CT (DE 1361/42-1, DFG project number: 567897117) and the extended follow-up of the DISCHARGE trial (DISCHARGE-Extend, DE 1361/35-1, DFG project number: 551186137). He also received funding from the Berlin University Alliance (GC_SC_PC 27), from the Innovationsfond of the G-BA for the IMPRO project (01NVF23002), and from the Digital Health Accelerator of the Berlin Institute of Health.
All other authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alex Kendall, V.B., Cipolla, R.: Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Tae-Kyun Kim, Stefanos Zafeiriou, G.B., Mikolajczyk, K. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 57.1–57.12. BMVA Press (Sep 2017). https://doi.org/10.5244/C.31.57
2. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. In: International Conference on Learning Representations (2020)
3. Burmeister, J.M., Rosas, M.F., Hagemann, J., Kordt, J., Blum, J., Shabo, S., Bergner, B., Lippert, C.: Less is more: A comparison of active learning strategies for 3D medical image segmentation (2022)

4. Chitta, K., Alvarez, J.M., Lesnikowski, A.: Large-Scale Visual Active Learning with Deep Probabilistic Ensembles (Feb 2019). https://doi.org/10.48550/arXiv.1811.03575
5. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3D slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging **30**(9), 1323–1341 (2012)
6. Filstroff, L., Sundin, I., Mikkola, P., Tiulpin, A., Kylmäoja, J., Kaski, S.: Targeted active learning for bayesian decision-making. Transactions on Machine Learning Research (2024)
7. Föllmer, B., Schulze, K., Wald, C., Stober, S., Samek, W., Dewey, M.: Active learning with the nnUNet and sample selection with uncertainty-aware submodular mutual information measure. In: Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., Mateus, D. (eds.) Proceedings of the 7nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 250, pp. 480–503. PMLR (Jul 2024)
8. Föllmer, B., Serafimoski, V., Schulze, K., Biavati, F., Stober, S., Samek, W., Dewey, M.: Guided active learning for medical image segmentation. In: Guo, X., Jin, Y., Lamdouar, H., Men, Q., Ouyang, C., Sahu, M., Vedula, S.S. (eds.) Human-AI Collaboration. pp. 58–67. Springer Nature Switzerland, Cham (2026)
9. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1183–1192. PMLR (Aug 2017)
10. Gharleghi, R., Adikari, D., Ellenberger, K., Ooi, S.Y., Ellis, C., Chen, C.M., Gao, R., He, Y., Hussain, R., Lee, C.Y., Li, J., Ma, J., Nie, Z., Oliveira, B., Qi, Y., Skandarani, Y., Vilaça, J.L., Wang, X., Yang, S., Sowmya, A., Beier, S.: Automated segmentation of normal and diseased coronary arteries – The ASOCA challenge. Computerized Medical Imaging and Graphics **97**, 102049 (Apr 2022). https://doi.org/10.1016/j.compmedimag.2022.102049
11. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT (Jul 2023). https://doi.org/10.48550/arXiv.2307.01984
12. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
13. Kirsch, A., Farquhar, S., Atighehchian, P., Jesson, A., Branchaud-Charron, F., Gal, Y.: Stochastic Batch Acquisition: A Simple Baseline for Deep Active Learning (Sep 2023). https://doi.org/10.48550/arXiv.2106.12059
14. Kothawade, S., Savarkar, A., Iyer, V., Ramakrishnan, G., Iyer, R.: CLINICAL: Targeted active learning for imbalanced medical image classification. In: Zamzmi,

G., Antani, S., Bagci, U., Linguraru, M.G., Rajaraman, S., Xue, Z. (eds.) Medical Image Learning with Limited and Noisy Data. pp. 119–129. Springer Nature Switzerland, Cham (2022)
15. Kulesza, A., Taskar, B.: K-dpps: Fixed-size determinantal point processes. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 1193–1200 (2011)
16. Lüth, C.T., Traub, J., Kahl, K.C., Bungert, T.J., Klein, L., Krämer, L., Jaeger, P.F., Isensee, F., Maier-Hein, K.: nnActive: A Framework for Evaluation of Active Learning in 3D Biomedical Segmentation (Nov 2025). https://doi.org/10.48550/arXiv.2511.19183
17. Lüth, C.T., Traub, J., Kahl, K.C., Bungert, T.J., Klein, L., Krämer, L., Jäger, P.F., Maier-Hein, K., Isensee, F.: Finally Outshining the Random Baseline: A Simple and Effective Solution for Active Learning in 3D Biomedical Imaging (Jan 2026). https://doi.org/10.48550/arXiv.2601.13677
18. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018)
19. Settles, B.: Active learning literature survey (2009)
20. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal **27**(3), 379–423 (Jul 1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
21. Zhan, X., Liu, H., Li, Q., Chan, A.B.: A Comparative Survey: Benchmarking for Pool-based Active Learning. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. pp. 4679–4686. International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada (Aug 2021). https://doi.org/10.24963/ijcai.2021/634