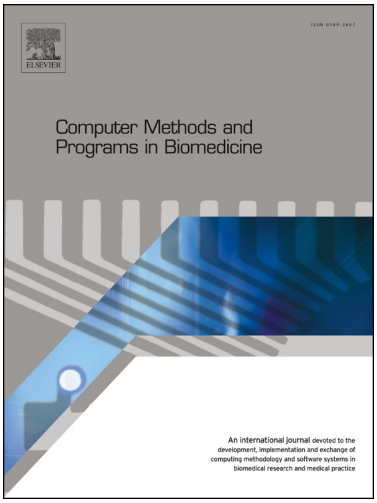
Active Learning for Left Ventricle Segmentation in Echocardiography

Journal Pre-proof



Active Learning for Left Ventricle Segmentation in Echocardiography

Eman Alajrami, Tiffany Ng, Jevgeni Jevsikov, Preshen Naidoo, Patricia Fernandes et al.

PII: S0169-2607(24)00107-X
DOI: https://doi.org/10.1016/j.cmpb.2024.108111
Reference: COMM 108111

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 29 November 2023
Revised date: 21 February 2024
Accepted date: 1 March 2024

Please cite this article as: E. Alajrami, T. Ng, J. Jevsikov et al., Active Learning for Left Ventricle Segmentation in Echocardiography, *Computer Methods and Programs in Biomedicine*, 108111, doi: https://doi.org/10.1016/j.cmpb.2024.108111.

Highlights

- The study contributes a unique dataset of echocardiogram images annotated by accredited experts, which is made publicly available for further research and model development.
- The study proposes a novel active learning approach for efficient left ventricle segmentation in echocardiography.
- The proposed approach combines uncertainty-based and representativeness sampling methods to improve annotation efficiency.
- The authors evaluate their approach on two datasets and demonstrate that it can significantly reduce annotation costs by up to 80% while maintaining high segmentation performance.

Active Learning for Left Ventricle Segmentation in Echocardiography

Eman Alajrami[1], Tiffany Ng[2], Jevgeni Jevsikov[1,2], Preshen Naidoo[1], Patricia Fernandes[1], Neda Azarmehr[1], Fateme Dinmohammadi[1], Matthew J Shun-shin[2], Nasim Dadashi Serej[1], Darrel P Francis[2], and Massoud Zolgharni[1,2]

[1] Intelligent Sensing and Vision, University of West London, London, UK
[2] National Heart and Lung Institute, Imperial College London, London, UK
eman.alajrami@uwl.ac.uk

Abstract

Background and Objective: Training deep learning models for medical image segmentation requires large annotated datasets, which can be expensive and time-consuming to create. Active learning is a promising approach to reduce this burden by strategically selecting the most informative samples for segmentation. This study investigates the use of active learning for efficient left ventricle segmentation in echocardiography with sparse expert annotations.
Methods: We adapt and evaluate various sampling techniques, demonstrating their effectiveness in judiciously selecting samples for segmentation. Additionally, we introduce a novel strategy, Optimised Representativeness Sampling, which combines feature-based outliers with the most representative samples to enhance annotation efficiency.
Results: Our findings demonstrate a substantial reduction in annotation costs, achieving a remarkable 99% upper bound performance while utilizing only 20% of the labelled data. This equates to a reduction of 1680 images needing annotation within our dataset. When applied to a publicly available dataset, our approach yielded a remarkable 70% reduction in required annotation efforts, representing a significant advancement compared to baseline active learning strategies, which achieved only a 50% reduction. Our experiments highlight the nuanced performance of diverse sampling strategies across datasets within the same domain.
Conclusions: The study provides a cost-effective approach to tackle the challenges of limited expert annotations in echocardiography. By introducing a distinct dataset, made publicly available for research purposes, our work contributes to the field's understanding of efficient annotation strategies in medical image segmentation.

Keywords: Echocardiography · Deep learning · Active learning · Image segmentation

1 Introduction

Cardiovascular diseases (CVDs) are a leading cause of global mortality [1]. Echocardiogram (Echo), or cardiac ultrasound, examinations are widely used for non-invasive and safe diagnosis of CVDs [2]. However, manual interpretation of Echo images by trained clinicians can be prone to intra- and inter-observer variability, potentially resulting in diagnostic errors [3]. Therefore, there is a strong need for automated Echo interpretation for various tasks [4–10], including left ventricle (LV) segmentation.

Accurate LV segmentation is crucial for measuring clinical parameters such as LV ejection fraction, which is a key indicator of cardiac function [11, 12]. Deep learning (DL) methods, including the U-Net architecture, have shown significant performance in medical image segmentation [13, 14]. The U-Net model is widely applied for LV segmentation due to its efficiency and performance compared to alternative networks [15].

However, DL models typically require large annotated datasets for effective training, which, in medical imaging, can be costly, time-consuming and laborious, requiring highly skilled experts. Limited annotations in medical imaging hinder effective DL model training, leading to inadequate segmentation results [16].

Therefore, automated methods are required to minimise annotation efforts in medical imaging. This is where weakly supervised learning emerges as a valuable tool, able to leverage diverse sources of information with minimal manual annotations. Two weakly supervised approaches, active learning (AL) and semi-supervised learning can enhance model performance with less annotated data [17]. AL, combined with DL, focuses on selecting the most informative samples for annotation and training [18, 19], while semi-supervised learning leverages both labelled and unlabelled samples to refine data representation [20, 21]. Nonetheless, the challenge of selecting samples for labelling underscores the significance of AL.

1.1 Related work

AL selection methods can be categorised into uncertainty-based, diversity-based, and hybrid approaches [22]. Uncertainty sampling, a commonly used strategy in AL, measures the model's uncertainty on unseen instances to select the most uncertain samples for annotation [19, 23–25]. However, it may overlook the distribution of data points and choose redundant samples with similar features [26]. Representativeness sampling selects the diverse samples that highly represent the unlabelled dataset to reduce annotation costs [26, 27]. Hybrid methods combine uncertainty and diversity sampling to choose representative and uncertain samples [28].

Numerous studies have highlighted the advantages of using AL in various areas of medical imaging, including histopathology, breast cancer, and digital pathology [29–33]. This technique holds immense promise for both improving accuracy and reducing the need for extensive data annotation, a crucial aspect in areas reliant on expert analysis.

Budd, Robinson, and Kainz [22] conducted a comprehensive survey discussing active learning and human-in-the-loop approaches in the field of medical imaging. Gal and Ghahramani [34] introduced Monte Carlo dropout (MCD) as a Bayesian approximation for modelling uncertainty in CNNs, primarily for image classification. This technique involves generating multiple predictions for each image and using various metrics such as variational ratios, maximum entropy, mean of standard deviation, Bayesian Active Learning with Disagreement (BALD), and random selection to calculate the uncertainty score [35].

Gorriz et al. [23] applied Cost-effective Active Learning (CEAL) with MCD for melanoma segmentation, where MCD was used to estimate pixel-wise uncertainty. They employed the CEAL approach proposed by Wang et al. [36] to select uncertain images for subsequent training iterations and generate pseudo labels for confident samples.

Other studies have explored representativeness sampling, such as the Core-set method, which minimises the Euclidean distance between sampled and remaining points in the feature space [37]. Nguyen et al. [27] and Xu et al. [38] utilised clustering techniques to identify representative and diverse samples for querying.

Hybrid techniques that combine uncertainty and diversity sampling have also been proposed [28, 39, 40]. For example, Kim et al. [41] introduced a selection strategy that involves constructing a small representative core-set from the unlabeled data and subsequently selecting the most uncertain images from the core-set.

While AL techniques have been extensively studied in classification tasks [35, 36, 42], their application to image segmentation is relatively limited. To our knowledge, no approach utilises current state-of-the-art AL methods in echocardiography. Therefore, this study focuses on applying AL to improve image segmentation tasks in echocardiography.

1.2 Main Contributions

The study makes several significant contributions to the field of active learning for left ventricle segmentation in echocardiography:

- **Comprehensive evaluation of active learning approaches, when applied in echocardiography:** Our study conducts a thorough evaluation of existing active learning approaches tailored specifically for medical image segmentation. By systematically assessing the performance of various active learning strategies, we establish a valuable baseline for comparing the efficacy of our proposed framework.
  This comprehensive evaluation not only highlights the strengths and weaknesses of different active learning techniques but also provides invaluable insights to guide future enhancements in this field. Our findings contribute to advancing the understanding of active learning methodologies tailored to the complexities of echocardiographic data.
- **Optimised surrogate metric for representativeness:** We introduce an innovative surrogate metric for representativeness, which serves as a simple

yet highly effective addition to current active learning approaches. This metric offers several advantages, including a notable reduction in the amount of labeled data required to train deep learning models.
By minimising the issue of uncertainty-based methods commonly querying samples with redundant information, our approach achieves substantial gains in annotation cost reduction, up to 80%, while maintaining high segmentation performance. This novel contribution addresses a crucial aspect of active learning and presents a promising avenue for improving efficiency in echocardiographic image annotation processes.

– **Facilitation of data accessibility and standardisation:** As part of our study, we curate a dataset comprising echocardiography images annotated by accredited and experienced echocardiography experts. By making this dataset publicly available through our report, we aim to standardise echocardiographic analysis practices and facilitate advancements in automated model development.

These contributions represent not only technical advancements but also significant strides in advancing scientific knowledge, fostering reproducibility, and facilitating collaboration within the realm of active learning-driven approaches for left ventricle segmentation in echocardiography.

2 Methods

2.1 Patient datasets and expert annotations

– **Unity:** A large random sample of 1224 echocardiographic studies from different patients performed was extracted from Imperial College Healthcare NHS Trust's echocardiogram database. Ethical approval was obtained from the Health Regulatory Agency for the anonymised export of large quantities of imaging data. It was not necessary to approach patients individually for consent of data originally acquired for clinical purposes.
The images were acquired during examinations performed by experienced echocardiographers, according to the standard protocols for using ultrasound equipment from the corresponding manufacturers. Automated anonymisation was performed to remove the patient-identifiable information. A detailed description, including patient characteristics, can be found in Table 1.
A CNN model, previously developed in our research group to detect different echocardiographic views [43], was then used to identify and separate the apical 4-chamber (A4C) views. From these videos, a total of 2800 images were subsequently automatically extracted from different time-points in the cardiac cycle.
Each image underwent expert labelling by one individual from a pool of 6 experts using our web-based real-time platform (https://unityimaging.net). This platform enables experts to accurately label the endocardial border; they labelled the LV endocardial curve including specific points for the apex, and for the two ends, namely the septal and lateral mitral hinge points.

The dataset was then split into three parts to generate training, validation, and testing sets (70%, 15%, 15%, respectively). We ensured that a single study's images did not appear in more than one set.
This dataset (images and labels) are available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license at https://intsav.github.io/efficient_annotations.html. The release of associated dataset received a Favourable Opinion from the South Central − Oxford C Research Ethics Committee (Integrated Research Application System identifier 279328, 20/SC/0386).

Table 1. A summary of the patient datasets used in this study.

	Unity	**CAMUS**
Size	1224 A4C videos 2800 frames were randomly selected	450 A4C videos 450 end-diastolic frames
Source and year enrolled	Random selection of echo studies from 7 UK laboratories during 2015-16	Sequential echo studies from University Hospital St Etienne (France) in 2019
Sex and Age	M: 401 (33%); F: 753 (62%) Unspecified : 70 (5%)	Unkown
Original size (pixels)	(400×300) to (1024×768) resized to 512×512	Unkown; resized to 512×512
Manufacturer /Model	Philips Healthcare (iE33, Affinity 70C, Epic 7C, Affinity 50G, CX50) and GE Healthcare (Vivid I, Vivid q, Vivid S70, Vivid S6, Vivid E9, Vivid 7)	GE Healthcare (Vivid E95)
Format	DICOM	MHD

– **CAMUS:** The second dataset is a publicly available dataset for which detailed information can be found elsewhere [44]. We used 450 end-diastolic images from 450 distinct patients. This dataset was divided into training, validation, and testing sets in a ratio of 70%, 15%, and 15%, respectively.

2.2 Network architecture

Fig. 1 presents the MCD U-Net architecture with a depth of 5 designed for Bayesian AL in LV segmentation. Dropout layers are integrated after each encoder and decoder block to enable the MCD for uncertainty estimation. The encoder blocks consist of Conv2D layers with a 3×3 kernel size, followed by

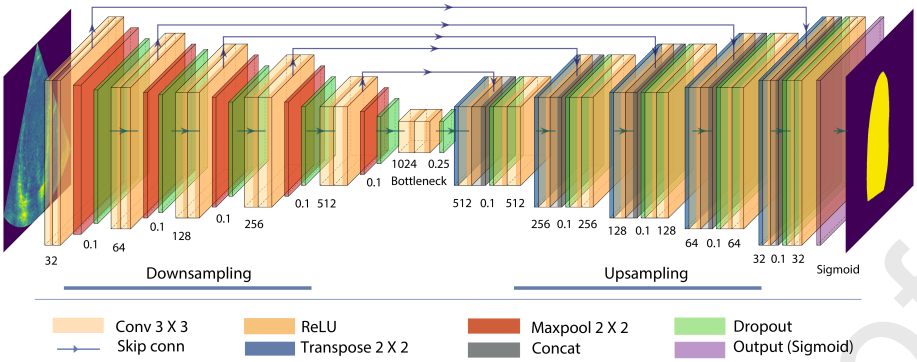


Fig. 1. Adapted U-Net model of depth 5 with dropout for uncertainty: features Conv2D layers with 3×3 kernels, batch normalization, and ReLU in encoder and decoder blocks. Includes 2×2 maxpooling and dropout (0.1 probability) in each block, with a dropout (0.25) at the bottleneck. Decoder employs Conv2DTranspose with concatenation and batch normalization. Output is a 1×1 Conv2D layer with Sigmoid activation.

batch normalization, ReLU activation, and 2×2 Maxpooling. A Dropout layer (probability of 0.1) is applied. The network bottleneck contains two Conv2D layers, and a Dropout layer (probability of 0.25) is employed. In the decoder blocks, Conv2DTranspose with a 2×2 kernel size is used, followed by concatenation, Dropout (probability of 0.1), batch normalization, and ReLU activation. Two Conv2D layers with a 3×3 kernel size, batch normalization, and ReLU activation are included. The output layer consists of a Conv2D layer with a 1×1 kernel size and the Sigmoid activation function. The architecture visualization in Fig. 1 was created using the PlotNeuralNet tool.

2.3 Sampling strategies

This study follows a standard pool-based AL methodology, illustrated in Fig. 2. The AL process comprises four steps:

1. Initial training of the model using labelled data (L).
2. computation of model uncertainty scores and/or representativeness scores (depending on the sampling approach adopted) for the unlabeled data pool (U).
3. Selection of a batch of highest-ranked images (K), followed by expert annotation and addition to L.
4. Iterative model retraining using the updated labelled data (L).

These steps are repeated until the desired number of AL iterations is reached, or a desired level of model performance is achieved. This approach enables the active selection of the most informative samples for annotation, achieving high segmentation performance with a limited number of labelled images.

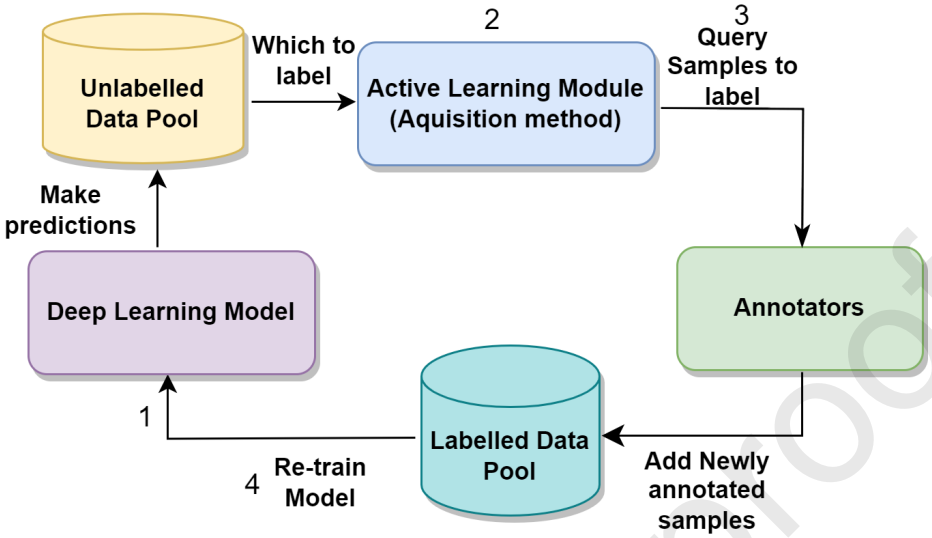


Fig. 2. A typical active learning cycle: Begins with initial training on labelled data, followed by scoring unlabeled data for uncertainty or representativeness. Selects and annotates top-ranked images for addition to the labelled pool and concludes with iterative retraining using the expanded labelled dataset

Random sampling and a variety of different selective sampling approaches were used for selecting the next batch of images from the unlabelled pool:

- **Random** is the baseline technique for randomly acquiring the next batch of images to be labelled. The random uniform distribution allocates random scores over the interval [0,1] for each image in U.
- **Uncertainty scoring** We explored various uncertainty techniques, including classification uncertainty and Entropy, and adapted BALD and MCD Entropy to improve segmentation performance [35, 42, 45, 46].
  - **Pixel-wise**, known as the least confident[22]. Since our model's output layer is sigmoid, it gives a probability P for each pixel in an image between 0 and 1, where 0 is for the background, and 1 is for the foreground (the mask), which can be described in Equation 1 as follows:

$$f(P) = \begin{cases} 1 & \text{if } P \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The closer the pixel probabilities are to 0.5, the more uncertain the model's prediction is. By calculating the absolute value of the difference between each pixel probability and 0.5, one can measure the uncertainty of each prediction. Summing up these values for all pixels in the image gives us an overall uncertainty score for that image.

Images with low uncertainty scores are the ones where the model is relatively confident. On the other hand, images with high uncertainty scores are the ones where the model is less certain about its predictions, and their labels will be acquired for the next training iteration.

- **Predictive Entropy**, also called Maximum Entropy (Max-entropy), is a commonly used metric for quantifying uncertainty, utilising Shannon Entropy [47]. This measure focuses on epistemic uncertainty, which represents the level of information required to encode a given distribution [24]. The predictive entropy can be computed using the following equation:

$$H(X) = -\sum_{c=0}^{C} P(c) \log P(c) \tag{2}$$

where P is the probability estimated for a class C given an image X belongs to U. In the context of image segmentation, predictive entropy can be applied to assess the uncertainty of the model's predictions on a per-image basis. We adapted the predictive entropy for segmentation by deriving probabilities for foreground and background pixel classes. The probability for the foreground class is obtained directly from the sigmoid output, while the probability for the background class is derived as 1 - P. Then, we used the probabilities vector for each pixel within an image to calculate an Entropy score per pixel using Equation 2.

In this modified approach, we replaced the original probability value P in Equation 2 with P_i, a list containing both P and 1 - P, representing the probabilities for foreground and background classes, respectively. The updated Equation 3 is as follows:

$$H(X) = -\sum_{i=0}^{N} \sum_{c=0}^{C} P_i\,(c) * \log P_i\,(c) \tag{3}$$

where P_i denotes the pixel probabilities for each pixel i belonging to an image X, and N represents the number of pixels of an image X belonging to U. In the context of active learning, images with high predictive entropy are considered more uncertain and informative because the model is less confident about its predictions for those images.

- **Ensembles-based methods**, where we incorporated dropout layers into the U-Net model (Fig. 1), use the MCD technique for Bayesian approximation [34]. The concept behind MCD is the activation of dropped neurons during inference, leading to multiple predictions for each image and allowing us to compute uncertainty scores for the unlabeled images.

In our study, we performed T forward passes for each image in U, employing $T = 100$ to achieve satisfactory performance. Next, uncertainty measures are used, including:

Variance (Var), where this measure calculates the uncertainty of a pixel by considering the variance of its predicted probabilities across the multiple predictions generated by the MCD technique [23]. We averaged the pixel

variances to get an uncertainty score per image, and the images with high scores were queried.

MCD-entropy, where this measure calculates the uncertainty of a pixel by considering the distribution of its predicted probabilities. We calculated the entropy of the mean predictions for each image in U over T times as below, and images with high scores are selected for annotations.

$$H(x) = -\sum_{i=0}^{N}\sum_{c=0}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}P_i(c)\log\left(\frac{1}{T}\sum_{t=1}^{T}P_i(c)\right)\right) \tag{4}$$

BALD, which aims to select the unlabelled images with the highest disagreement between the predictions of different models for annotation, operates by maximising Mutual Information (MI). MI measures the degree of dependency between the model's parameters and its output (i.e., predictions) [48]. Essentially, it quantifies how much the uncertainty in the model's predictions can be reduced by knowing the true label for a particular image. We used MCD and sampled many networks, so when they disagreed on an image, some of them were wrong. We adapted BALD for binary segmentation using Entropy as described in Equation 4. To calculate the MI and have an uncertainty score for an image, we used the following equation:

$$MI(x) = H(x) - E[H(x)] \tag{5}$$

To plug BALD with MCD, we calculate Entropy1 H and Entropy2 E over T predictions; the first is H, similar to MCD-Entropy; the second is E, computed by calculating the Entropy for each prediction and averaging these Entropies as follows:

$$E[H(x)] = -\sum_{i=0}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{c=0}^{C}P_i(c)\log P_i(c)\right) \tag{6}$$

By substituting H and E in Equation 5, we get the following equation for MI (i.e. BALD) scoring:

$$MI(x) = -\sum_{i=0}^{N}\sum_{c=0}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}P_i(c)\log\left(\frac{1}{T}\sum_{t=1}^{T}P_i(c)\right)\right) + \sum_{i=0}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{c=0}^{C}P_i(c)\log P_i(c)\right) \tag{7}$$

Images with high MI are selected for annotations.

– **Representativeness sampling**

Methods that solely consider uncertainty have a drawback; they may concentrate only on limited regions of the data distribution because the models tend to be uncertain for similar types of images. If these methods train on samples from the same region, this could lead to redundancy or bias. Introducing a representativeness measure addresses this issue by promoting selection strategies that sample from diverse regions of the distribution. We first applied

simple representativeness sampling and later optimised it using feature-based outliers:

- **Simple representativeness sampling (SRS)** This approach aims to select images that represent the remaining data but are unseen by the model. To achieve this, we initially extracted image features from both the labelled (L) and unlabeled (U) datasets using two different methods: VGG16 [49] and GLRM [50], to compare the performance of the two models. Then, we used cosine similarity to measure the similarity between samples [51]:

$$\text{Cos\_Sim}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}, \quad \text{where} \quad x, y \in U \tag{8}$$

$$\text{Cos\_Sim}(x, z) = \frac{x \cdot z}{\|x\|\|z\|}, \quad \text{where } x \in U \ , \ z \in L \tag{9}$$

  For each image x in the unlabeled set, we calculated the maximum similarity (max_sim) score between x and the other images in the unlabeled set U and the maximum similarity score between x and all the images in the labelled set L. We then calculated the representativeness score (Rep_score) for each image x in the unlabeled set by subtracting the two scores:

$$\text{Rep\_score}(x) = \text{max\_sim}(x, U) - \text{max\_sim}(x, L) \tag{10}$$

  We sorted the unlabeled images in descending order based on their representativeness scores, applying Equation 10, and selected the K highest-ranked samples for annotation.
- **Optimised representativeness sampling (ORS)** One limitation of the SRS is that it tends to select images from the unlabeled pool that are highly representative of the remaining images, thus excluding images with significantly different characteristics. This can lead to a lack of diversity in the training data, hence hindering the performance of the model.
  Therefore, we propose ORS, a density-based sampling approach to address the limitation of SRS by identifying regions of high data density in the feature space, and selecting samples from both high- and low-density areas to ensure coverage of different data distributions. To achieve this, we select some samples for annotation that are feature-based outliers, i.e., images with lower similarities to other unlabeled images that are also unseen by the model. By including feature-based outliers in the selected images, ORS improves the diversity of the training data.
  To implement ORS, we first divided the unlabeled set U into two lists: list-1, most representative images, and list-2, feature-based outliers. We followed the same steps as in the previous section to sort the most representative images in descending order.
  For feature-based outliers, we pick the images with similarities to other unlabeled images below a threshold which can be tuned; here, it was the mean similarity of the unlabeled set. We then calculated their representativeness scores and ranked them in ascending order.

Finally, to get the K samples, we used both lists as follows: K1 = 0.75% of K from list-1, K2 = 0.25% of K from list-2, and then K = K1 + K2. This split can be tuned; we tried different splits and achieved optimal results with this approach.

– **Combined uncertainty with ORS sampling** To leverage the benefits of both uncertainty-based and representativeness-based active learning, we combined both approaches as follows: (1) Initially, we trained the model using the available labelled images. (2) Next, we employed the ORS technique to select the top 50 representative images from the pool of unlabeled data. (3) We then applied the predictive entropy method to measure uncertainty on these 50 images. (4) We selected K samples for labelling from the subset of images with the highest uncertainty, which were used to retrain the model.

2.4 Implementation and training settings

Tensorflow [52] and Keras [53] frameworks are used to develop the DL models. The training was conducted using an Nvidia RTX3090 GPU. The U-Net was trained using binary cross-entropy loss and ADAM optimiser, with a learning rate of 0.0001 for 200 epochs. In order to avoid overfitting, early stopping was applied with a patience of 10. Variable-sized images (and corresponding ground truth) were zero-padded to a uniform size and then resized to 512×512 pixels. A fixed batch size of 8 was used.

For the CAMUS dataset, we selected 10% (35 images) of the initial training data as L (labelled pool), and U (unlabelled pool) was the remaining 90%. For the Unity dataset, we chose 4% (82 images) as the initial L, and the remaining was U. For each dataset, all experiments using different approaches for selective sampling started with the same model, trained on the initial labels in L. The size of the K samples was 5% and 1% of the total CAMUS and Unity datasets, respectively.

2.5 Evaluation metrics

The Dice-Coefficient (DC), also known as Sorensen-Dice or Dice Similarity Coefficient, is a similarity measure over sets [54]. The original formula of the Dice coefficient formula is defined as follows:

$$DC = \frac{2|A \cap B|}{|A| + |B|} \tag{11}$$

where A represents the ground truth, and B represents the predicted mask. This equation calculates the degree of overlap between the predicted mask and its corresponding ground truth.

The models were evaluated after every AL iteration using the DC metric, widely used for evaluating image segmentation accuracy, using Equation 11. DC was computed between the ground truth and the inferred prediction for each image in the testing dataset. Then, the mean of DC scores of all images is calculated to present the model's accuracy.

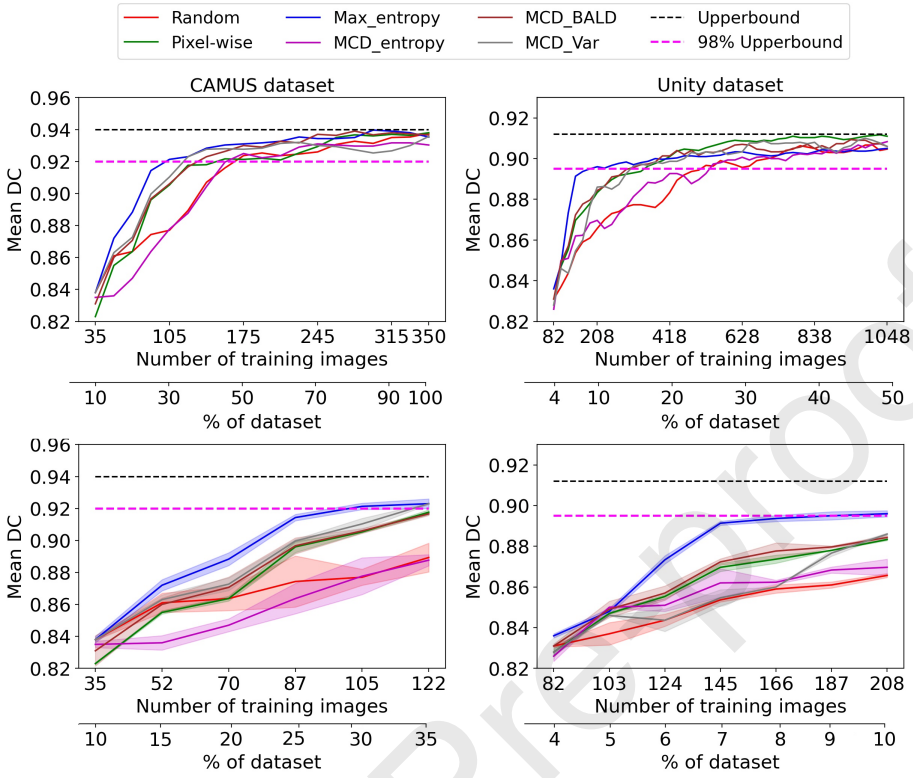


Fig. 3. Performance profiles for various uncertainty-based sample selection strategies at each active learning iteration; lower panel shows a magnified version of early stages presented at the upper panel. Black and pink horizontal broken lines indicate the upper bound (training on the full dataset) and 98% of the upper bound, respectively. Shaded areas denote ± half of the standard deviation. As it is evident, the predictive entropy was the best uncertainty strategy for the two echo datasets used in this study, especially at the initial active learning iteration stages.

Each AL selection strategy was trained three times, and the average *DC* at each AL iteration was computed and used to plot the results.

3 Results

All results are reported for the testing dataset to evaluate different AL approaches. Each acquisition method is trained three times, and the average of their Dice Coefficient scores is used.

3.1 Uncertainty sampling

Fig. 3 shows that all uncertainty methods outperformed random selection except MCD-entropy. Predictive entropy was the best uncertainty strategy for the two echo datasets used in this study, especially at the initial AL iteration stages. It achieved 97.7% of the upper bound (maximum accuracy achievable) using the entire CAMUS dataset with only 25% of the annotations, while other approaches required approximately 35% or more to achieve similar performance. This would reduce the cost of labelling images by 10%.

For the Unity dataset, the predictive entropy significantly outperformed all other methods from the early stages of AL, achieving 98.3% and 98.6% of the maximum achievable performance. After using 30% of the annotations, the pixel-wise selection approach almost converged to the best performance, outperforming all other uncertainty techniques.

Additionally, the shaded areas indicate that results from all selection approaches, except random and MCD-entropy, are highly reproducible.

3.2 Representativeness sampling

Fig. 4 illustrates performance plots for representativeness sampling strategies. Evidently, our proposed optimised method (ORS) outperformed the existing approaches for both datasets, improving the performance, particularly at the early stages of AL interactions. This is likely due to ORS selecting images with various distributions in the dataset.

ORS method achieved 98% of the maximum achievable performance (upper bound) using only 25% of labels in the CAMUS dataset, outperforming the SRS method, which required 35% of the annotations to approach the same performance level.

With GLRM, ORS achieved almost the same accuracy as the full dataset using 50% of annotations, while the SRS method needed 90% of the labelled data to reach that performance. With VGG16, however, the ORS method outperformed the SRS at the early AL stages until 30% of labelled data was added to the training dataset; after that, both methods performed similarly.

For the Unity dataset, ORS with GLRM converged to upper bound performance using only 15% of the labelled data, significantly reducing the labelling effort. SRS required 33% of the annotations to approach that performance. With VGG16, the improved performance of ORS over SRS was less pronounced in later stages, following the pattern observed in the CAMUS dataset.

3.3 Combined uncertainty with ORS results

The results of integrating the ORS technique with the Max-Entropy uncertainty-based method are illustrated in Fig.5. The combination of ORS with Max-Entropy enhanced the performance of AL on the CAMUS dataset compared to using each method individually. It achieved 98.6% of the maximum performance using only 25% of the annotations, while utilising Max-entropy alone required 35% of annotations to reach the same performance level.

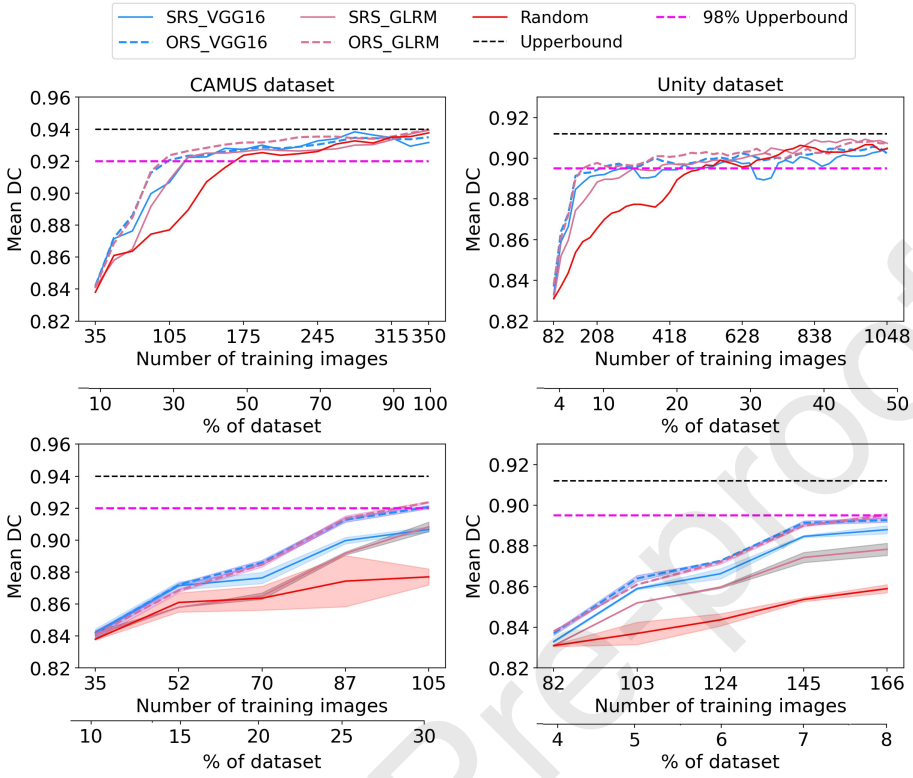


Fig. 4. Performance profiles for various representativeness sampling strategies versus the proposed optimised approach. Black and pink horizontal broken lines indicate the upper bound (training on the full dataset) and 98% of the upper bound, respectively. haded areas denote ± half of the standard deviation. Evidently, our proposed optimised method (ORS) outperformed the existing approaches for both datasets, improving the performance, particularly at the early stages of active learning interactions.

However, on the Unity dataset, the combined approach achieves a more substantial boost, mainly at the early stages of active learning, when compared with Max-entropy. The proposed ORS method performed as well as the early stages combined approach. Afterwards, all three methods exhibited similar performance.

Fig.6 shows the improvement of the model in segmenting the left ventricle at the early stages of active learning when different image selection strategies are adopted.

Table 2 outlines the percentage of full data required to attain various upper bound performance levels; upper bound denotes the maximum achievable accuracy when trained on the complete dataset for each utilised dataset in this study.

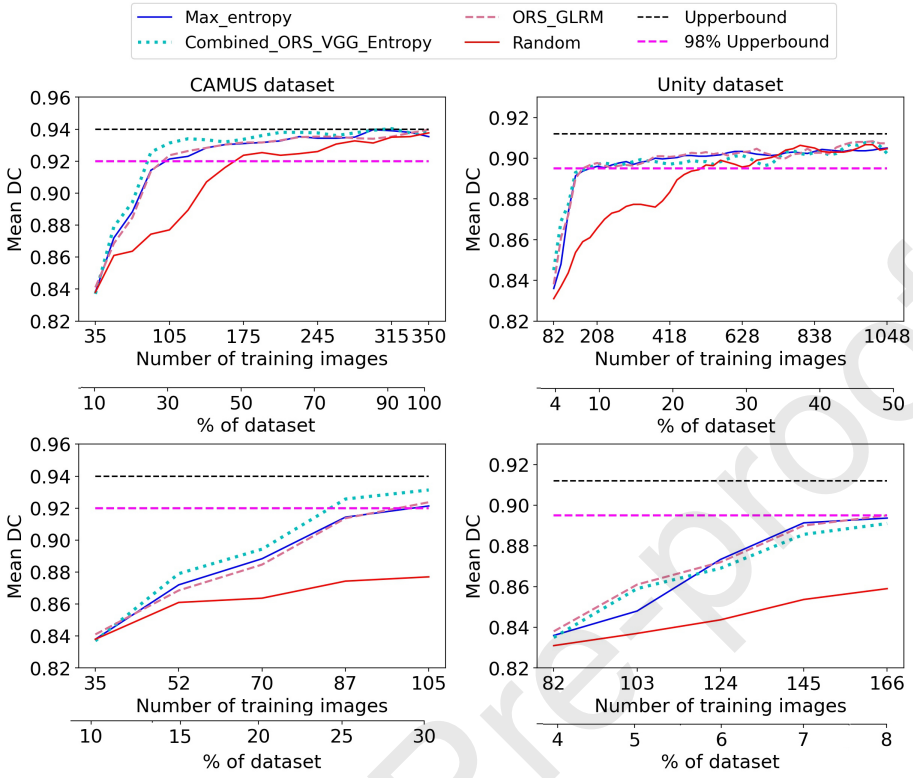


Fig. 5. Performance profiles comparing the best of each approach (uncertainty and representativeness) and also a combination of both (ORS and Max Entropy). The lower panel shows a magnified version of the early stages presented in the upper panel. Black and pink horizontal broken lines indicate the upper bound (training on the full dataset) and 98% of the upper bound, respectively.

To reach a 99% upper bound performance in CAMUS and Unity datasets, only 30% (combined method) and 20% (ORS_GLRMmethod) of the images need to be labelled, respectively. This translates to a remarkable reduction of 70% (245 fewer images) and 80% (1680 fewer images), resulting in substantial savings in annotation costs.

If we consider a 95% upper bound performance as acceptable, this implies that 280 and 1995 fewer images need to be annotated for the CAMUS and Unity datasets, respectively. This demonstrates a notable efficiency in achieving high accuracy with a reduced annotation burden.

Table 2. Percentage of labelled data needed to achieve various levels of performance in terms of Dice score for different sampling strategies; performance is given as the ratio of upper bound accuracy where upper bound is the maximum achievable accuracy when training on the full dataset.

Datasets	CAMUS				Unity			
Ratio of upper bound	**0.93**	**0.95**	**0.98**	**0.99**	**0.93**	**0.95**	**0.98**	**0.99**
Max_entropy	15%	23%	28%	50%	5.5%	6%	7.5%	20%
ORS_GLRM	17%	23%	28%	50%	4.5%	5%	7.5%	20%
Combined	13%	20%	24%	30%	4.5%	5%	8%	20%
Random	30%	45%	50%	75%	7%	15%	25%	40%

4 Discussions

The efficacy of AL methods may vary depending on the dataset, limiting their generalizability across different datasets [24]. As shown in Fig.3, uncertainty-based AL strategies have varied performance on two datasets from the same domain.

Our proposed method, ORS, outperforms SRS in the initial phases of active learning, possibly because ORS deliberately selects images with diverse distributions (see Fig.4. While SRS focuses on representative images for the entire dataset, it overlooks the consideration of diversity within those distributions. In contrast, ORS is designed to specifically choose images that encapsulate various distributions present in the dataset. This targeted approach proves advantageous, particularly in the early stages of active learning when the model is still acquainting itself with the dataset's diverse distributions.

To demonstrate the significance of our proposed ORS approach compared to SRS, we conducted a paired-sample t-test analysis. The results indicate significant differences between ORS and SRS for the Unity dataset, with p values of 0.005 and 0.0003 when utilising GLRM and VGG16, respectively. For the CAMUS dataset, the t-test revealed a significant difference ($p = 0.007$) between ORS and SRS when using GLRM. However, when using VGG16, no significant difference was found ($p = 0.094$).

The distinction in performance between GLRM and VGG16 can be attributed to their architectural differences. Despite VGG16's superiority with SRS, its performance levels align more closely with GLRM under the influence of ORS, emphasizing the efficacy of ORS in enhancing the performance of less complex models like GLRM, especially in comparison to the data demands of deep neural networks such as VGG16.

For the Unity dataset, the performance of ORS and combined methods was mostly similar throughout all stages of active learning. This could be attributed to the high similarity of image features within this large dataset, making selecting images from various distribution regions more crucial for labelling. Consequently, it is recommended to use ORS, as it provides similar performance while requiring less training time than the combined approach.

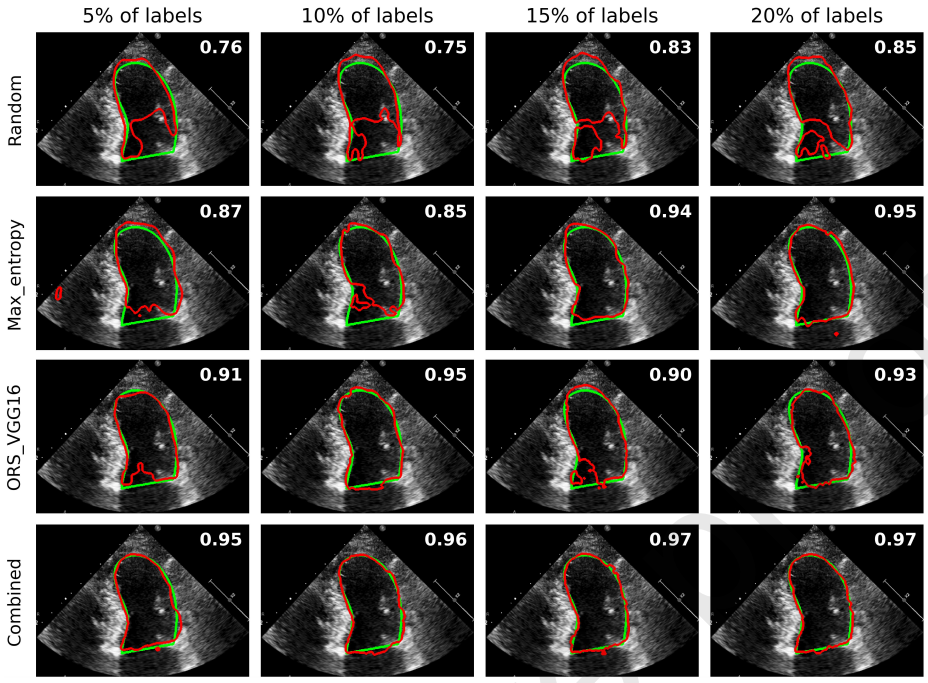


Fig. 6. Segmentation of left ventricle at different steps of active learning when using different sampling strategies: baseline (Random), uncertainty-based (max-entropy), proposed optimised representativeness sampling method (ORS), and the combined approach. Green and red contours represent manual and automated segmentations, respectively. The value of the *DC* metric has also been provided in the top right corner for each image.

4.1 Study limitations and future work

Our study showcases the effectiveness of our proposed ORS approach over other representativeness and uncertainty-based AL methods. However, it's crucial to acknowledge inherent limitations and areas for further enhancement within this domain, which warrant deeper discussion.

While our ORS approach, which emphasises dataset features, demonstrates improved generalisability compared to dataset-dependent uncertainty-based methods, its effectiveness remains subject to the feature distribution of the unlabeled pool. This highlights the importance of meticulous dataset analysis for future implementations.

An additional challenge in ORS lies in ensuring sample diversity during the selection of the most representative batches. Although our ORS technique is efficient, enhancing the sampling process by considering the dissimilarity of samples to previously chosen ones in each AL iteration could yield further improvements.

However, such enhancements may come at the cost of increased computational demands.

The iterative nature of AL renders it a resource-intensive framework. While our study conducted multiple experiments with consistent settings to compare competing methods fairly, the scope of our experimentation was limited, omitting certain potential setups. Therefore, extensive testing under varied conditions is imperative to comprehensively understand the applicability and performance of AL methods.

In future endeavours, we aim to address these limitations and expand upon our current method. This involves integrating self-supervised and semi-supervised learning techniques to merge high-quality image predictions with expert annotations as pseudo-labels. Furthermore, it's essential to examine the generalisability of the proposed method when applied to other medical imaging modalities and tasks, such as classification. This broader exploration will provide valuable insights into the versatility and efficacy of our approach across diverse medical imaging scenarios.

5 Conclusions

In this study, we address the challenge of limited annotations in medical image segmentation, specifically focusing on left ventricle segmentation in echocardiography. We introduce and adapt active learning techniques, including uncertainty-based and representativeness sampling methods, to improve model training efficiency.

Our findings demonstrate the effectiveness of these methods in reducing annotation costs while maintaining high segmentation performance. The integration of uncertainty-based methods leads to significant improvements in the early stages of active learning, allowing for the strategic selection of informative samples and optimisation of annotation efforts.

Additionally, our proposed optimised representativeness sampling method outperforms existing representativeness sampling techniques, providing a balanced and diverse set of samples for annotation. The combination of uncertainty-based and representativeness sampling can further enhance the efficiency of active learning, achieving close-to-maximum performance with reduced annotation efforts.

Our study also contributes a unique dataset of echocardiogram images annotated by accredited experts, which is made publicly available for further research and model development. Additionally, we introduce a majority-based consensus dataset, offering a robust benchmark for performance evaluations in left ventricle segmentation.

Acknowledgments

E Alajrami is supported by the Vice Chancellor's Scholarship at the University of West London. In addition, we would like to thank the Schlumberger Foundation for their funding and support for E Alajrami.

Statements of ethical approval

The release of the associated dataset received a Favourable Opinion from the South Central − Oxford C Research Ethics Committee (Integrated Research Application System identifier 279328, 20/SC/0386).

Funding

This work was supported in part by the British Heart Foundation, UK (Grants no. RG/F/22/110059 and PG/19/78/34733).

Declaration of Competing Interest

The authors declare no conflicts of interest.

References

[1] World Health Organisation. *Cardiovascular diseases (cvds)*. June 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Samuel Wang and Ping Hu. "Deep Learning for Automated Echocardiogram Analysis". In: *Journal of Student Research* 11.3 (2022).

[3] Graham D Cole et al. "Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation". In: *The international journal of cardiovascular imaging* 31 (2015), pp. 1303–1314.

[4] Elisabeth S Lane et al. "Automated multi-beat tissue Doppler echocardiography analysis using deep neural networks". In: *Medical & Biological Engineering & Computing* (2023), pp. 1–16.

[5] Elisabeth S. Lane et al. "Multibeat echocardiographic phase detection using deep neural networks". In: *Computers in Biology and Medicine* 133 (2021), p. 104373. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104373. URL: https://www.sciencedirect.com/science/article/pii/S0010482521001670.

[6] James P Howard et al. "Automated left ventricular dimension assessment using artificial intelligence developed and validated by a UK-wide collaborative". In: *Circulation: Cardiovascular Imaging* 14.5 (2021), e011951.

[7] Massoud Zolgharni et al. "Automated Aortic Doppler Flow Tracing for Reproducible Research and Clinical Measurements". In: *IEEE Transactions on Medical Imaging* 33.5 (2014), pp. 1071–1082. DOI: 10.1109/TMI.2014.2303782.

[8] Neda Azarmehr et al. "Neural architecture search of echocardiography view classifiers". In: *Journal of Medical Imaging* 8.3 (2021), p. 034002. DOI: 10.1117/1.JMI.8.3.034002. URL: https://doi.org/10.1117/1.JMI.8.3.034002.

[9] Eman Alajrami et al. "Deep Active Learning for Left Ventricle Segmentation in Echocardiography". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2023, pp. 283–291.

[10] Jevgeni Jevsikov et al. "Automated Analysis of Mitral Inflow Doppler Using Deep Neural Networks". In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2023, pp. 394–402.

[11] Preshen Naidoo et al. "Influence of Loss Function on Left Ventricular Volume and Ejection Fraction Estimation in Deep Neural Networks". In: *Medical Imaging with Deep Learning*. 2022.

[12] Zhisheng Yan et al. "SegNet-based left ventricular MRI segmentation for the diagnosis of cardiac hypertrophy and myocardial infarction". In: *Computer Methods and Programs in Biomedicine* 227 (2022), p. 107197. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.107197. URL: https://www.sciencedirect.com/science/article/pii/S0169260722005788.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *MICCAI 2015*. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[14] Gongping Chen, Yu Dai, and Jianxun Zhang. "C-Net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation". In: *Computer Methods and Programs in Biomedicine* 225 (2022), p. 107086. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.107086. URL: https://www.sciencedirect.com/science/article/pii/S0169260722004679.

[15] Neda Azarmehr et al. "Segmentation of left ventricle in 2D echocardiography using deep learning". In: *Medical Image Understanding and Analysis: 23rd Conference, MIUA 2019, Liverpool, UK, July 24–26, 2019, Proceedings 23*. Springer. 2020, pp. 497–504.

[16] Xiangbin Liu et al. "A Review of Deep-Learning-Based Medical Image Segmentation Methods". In: *Sustainability* 13.3 (Jan. 2021), p. 1224. ISSN: 2071-1050. DOI: 10.3390/su13031224.

[17] Zeyu Ren, Shuihua Wang, and Yudong Zhang. "Weakly supervised machine learning". In: *CAAI Transactions on Intelligence Technology* 8.3 (Apr. 2023), pp. 549–580. DOI: 10.1049/cit2.12216.

[18] Burr Settles. "Active Learning Literature Survey". In: *University of Wisconsin, Madison* 52 (July 2010).

[19] Ren Pengzhen et al. "A Survey of Deep Active Learning". In: *arXiv* (Aug. 2020). DOI: 10.48550/arXiv.2009.00236.

[20] Zeyu Ren et al. "UKSSL: Underlying Knowledge based Semi-Supervised Learning for Medical Image Classification". In: *IEEE Open Journal of Engineering in Medicine and Biology* (2023), pp. 1–8. DOI: 10.1109/OJEMB.2023.3305190.

[21] Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. "Active learning for medical image segmentation with stochastic batches". In: *arXiv preprint arXiv:2301.07670* (2023).

[22] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. "A survey on active learning and human-in-the-loop deep learning for medical image analysis". In: *Medical Image Analysis* 71 (2021), p. 102062. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102062.

[23] Marc Gorriz et al. "Cost-Effective Active Learning for Melanoma Segmentation". In: *arXiv* (Nov. 2017). DOI: 10.48550/arXiv.1711.09168.

[24] Vishwesh Nath et al. "Diminishing uncertainty within the training pool: Active learning for medical image segmentation". In: *IEEE Transactions on Medical Imaging* 40.10 (2020), pp. 2534–2547.

[25] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. "Viewal: Active learning with viewpoint entropy for semantic segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9433–9443.

[26] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning, Aug. 2021, p. 424. ISBN: 9781638351030.

[27] Hieu Nguyen and Arnold Smeulders. "Active Learning Using Pre-clustering". In: *ICML* (2004). DOI: 10.1145/1015330.1015349.

[28] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. "Active Learning by Querying Informative and Representative Examples". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.10 (2014), pp. 1936–1949. DOI: 10.1109/TPAMI.2014.2307881.

[29] Lukasz Raczkowski et al. "ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning". In: *Scientific Reports* 9 (Oct. 2019). DOI: 10.1038/s41598-019-50587-1.

[30] André Meirelles et al. "Effective Active Learning in Digital Pathology: A Case Study in Tumor Infiltrating Lymphocytes". In: *Computer Methods and Programs in Biomedicine* 220 (Apr. 2022), p. 106828. DOI: 10.1016/j.cmpb.2022.106828.

[31] Jacob Carse and Stephen McKenna. "Active Learning for Patch-Based Digital Pathology Using Convolutional Neural Networks to Reduce Annotation Costs". In: July 2019, pp. 20–27. ISBN: 978-3-030-23936-7. DOI: 10.1007/978-3-030-23937-4_3.

[32] Mu Yang et al. "A CNN-based Active Learning Framework to Identify Mycobacteria in Digitized Ziehl-Neelsen Stained Human Tissues". In: *Com-*

puterized Medical Imaging and Graphics 84 (July 2020), p. 101752. DOI: 10.1016/j.compmedimag.2020.101752.

[33] Sanghoon Lee et al. "An Ensemble-based Active Learning for Breast Cancer Classification". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, pp. 2549–2553. DOI: 10.1109/BIBM47256.2019.8983317.

[34] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning* (June 2015).

[35] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. "Deep Bayesian Active Learning with Image Data". In: *arXiv* (Mar. 2017). DOI: 10.48550/arXiv.1703.02910.

[36] Keze Wang et al. "Cost-effective active learning for deep image classification". In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.12 (2016), pp. 2591–2600.

[37] Ozan Sener and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach". In: *arXiv preprint arXiv:1708.00489* (2017).

[38] Zuobing Xu, Ram Akella, and Yi Zhang. "Incorporating diversity and density in active learning for relevance feedback". In: *Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007. Proceedings 29*. Springer. 2007, pp. 246–257.

[39] Zheng Wang and Jieping Ye. "Querying discriminative and representative samples for batch mode active learning". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9.3 (2015), pp. 1–23.

[40] Firat Ozdemir et al. "Active learning for segmentation based on Bayesian sample queries". In: *Knowledge-Based Systems* 214 (2021), p. 106531. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2020.106531.

[41] Seong Tae Kim, Farrukh Mushtaq, and Nassir Navab. "Confident coreset for active learning in medical image analysis". In: *arXiv preprint arXiv:2004.02200* (2020).

[42] William H Beluch et al. "The power of ensembles for active learning in image classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9368–9377.

[43] Neda Azarmehr et al. "Neural architecture search of echocardiography view classifiers". In: *Journal of Medical Imaging* 8.3 (2021), p. 034002. DOI: 10.1117/1.JMI.8.3.034002. URL: https://doi.org/10.1117/1.JMI.8.3.034002.

[44] Sarah Leclerc et al. "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography". In: *IEEE transactions on medical imaging* 38.9 (2019), pp. 2198–2210.

[45] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning". In: *Advances in neural information processing systems* 32 (2019).

[46] Remus Pop and Patric Fulop. "Deep ensemble Bayesian active learning". In: *Bayesian Deep Learning Workshop At NeurIPS*. 2020.

[47] Claude E Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[48] Neil Houlsby et al. "Bayesian Active Learning for Classification and Preference Learning". In: *arXiv* (Dec. 2011). DOI: 10.48550/arXiv.1112.5745.

[49] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[50] Şaban Öztürk and Bayram Akdemir. "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA". In: *Procedia computer science* 132 (2018), pp. 40–46.

[51] Vagisha Gupta, Shelly Sachdeva, and Neha Dohare. "Chapter 8 - Deep similarity learning for disease prediction". In: *Trends in Deep Learning Methodologies*. Ed. by Vincenzo Piuri et al. Hybrid Computational Intelligence for Pattern Analysis. Academic Press, 2021, pp. 183–206. ISBN: 978-0-12-822226-3. DOI: 10.1016/B978-0-12-822226-3.00008-8.

[52] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[53] François Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015.

[54] Lee R Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

Declaration of Interest Statement

Eman Alajrami, None declared

Tiffany Ng, None declared

Jevgeni Jevsikov, None declared

Preshen Naidoo, None declared

Patricia Fernandes, None declared

Neda Azarmehr, None declared

Fateme Dinmohammadi, None declared

Matthew J Shun-shin, None declared

Nasim Dadashi Serej, None declared

Darrel P Francis, None declared

Massoud Zolgharni, None declared