





# Exploring Optimal Configurations in Active Learning for Medical Imaging

Alec Parise<sup>(✉)</sup>  and Brian Mac Namee 

School of Computer Science, University College Dublin, Dublin, Ireland  
alec.parise@ucdconnect.ie, brian.macnamee@ucd.ie

**Abstract.** Medical imaging is a critical component of clinical decision-making, patient diagnosis, treatment planning, intervention, and therapy. However, due to the shortage of qualified radiologists, there is an increasing burden on healthcare practitioners, which underscores the need to develop reliable automated methods. Despite the development of novel computational techniques, interpreting medical images remains challenging due to noise and varying acquisition conditions. One promising solution to improve the reliability and accuracy of automated medical image analysis is Interactive Machine Learning (IML), which integrates human expertise into the model training process. Active learning (AL) is an important IML technique that can iteratively query for informative samples to be labeled by humans, leading to more data-efficient learning. To fully leverage the potential of active learning, however, it is crucial to understand the optimal setup for different components of an AL system. This paper presents an evaluation of the effectiveness of different combinations of data representation, model capacity, and query strategy for active learning systems designed for medical image classification tasks. The results of this evaluation show that employing raw image representations as input, in conjunction with a ResNet50 model and margin-based queries, yields more reliable and accurate automated methods for medical image analysis.

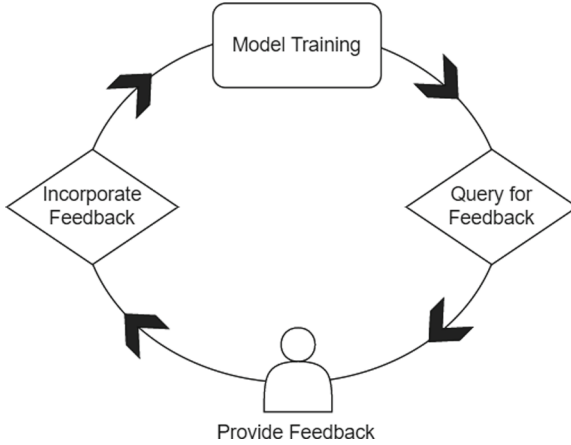
**Keywords:** Human-in-the-loop (HITL) · Interactive Machine Learning (IML) · Active Learning (AL)

## 1 Introduction

Medical imaging plays a crucial role in clinical decision-making, patient diagnosis, treatment planning, intervention, and therapy. However, the shortage of skilled radiologists poses a significant challenge, placing a growing burden on healthcare practitioners [10]. Consequently, there is an urgent requirement to

---

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.



**Fig. 1.** The basic Interactive Machine Learning (IML) framework

devise dependable automated techniques that can alleviate this strain and provide reliable solutions [42]. The field of Interactive Machine Learning (IML) has gained significant attention in the medical field in recent years [2, 8, 22, 25, 41]. Training image-based machine learning models typically relies solely on automated processes to learn patterns and make predictions, offering no ability for interaction from clinicians during this process [41]. In contrast, IML incorporates human input and feedback into the modelling process, resulting in more effective models [2, 25]. A typical IML workflow is presented in Fig. 1, where the automated model training process is periodically interrupted by user interaction. The user provides feedback to queries posed by the training process, and this feedback is then incorporated into another round of automated model training.

Active learning (AL) is a powerful IML technique for data-efficient model training [38]. By actively selecting informative instances to label, AL reduces the annotation burden while maintaining, or even improving, model performance. In the context of medical imaging, where labeled data is often scarce and costly to obtain, AL holds significant potential to enhance the accuracy and efficiency of diagnostic systems [38]. In this study, we aim to address the challenge of designing optimal AL systems for medical image classification, specifically focusing on three crucial aspects: data representation, model capacity, and query strategies. Our research aims to establish reliable default options for building active learning solutions.

To investigate the impact of model capacity, we consider three distinct models: a high-capacity Resnet50 model, a medium-capacity shallow convolutional neural network (CNN), and a low-capacity random forest model. Through the exploration of these diverse models, we seek to discern the influence of model capacity on the active learning process. Additionally, we examine the utility of image embeddings obtained from pre-trained models as an alternative to raw image inputs for the low-capacity models. This offers potential improvements in computational efficiency but also effectiveness in the presence of limited labeled samples.

Finally, we assess different query strategies to select informative instances during the active learning process: random sampling, margin sampling, and least confidence sampling. Through a comparative analysis of their performance, we aim to identify the sampling strategy that optimally selects informative data points for annotation, maximizing the efficiency of the active learning framework.

The experiments presented in this paper offer valuable insights into the effectiveness of different combinations of AL scenarios for medical image classification problems. This will be particularly useful to practitioners due to the lack of labelled data, where typical hyper-parameter tuning techniques are not suitable. The remainder of the paper proceeds as follows: Section 2 discusses related work; Sect. 3 provides a detailed explanation of the experimental setup; Sect. 4 discusses the experimental results; and finally Sect. 5 concludes the paper.

## 2 Related Work

Unlike typical machine learning algorithms that rely solely on large datasets, Interactive Machine Learning (IML) [2, 25] systems enable human interaction, allowing users to provide feedback and guidance to improve the accuracy and relevance of the models. The literature on IML includes algorithm development [23], user interface design [17, 18], and applications in diverse fields [1, 16, 31, 47]. Important IML methods include visual pattern mining [32], interactive anomaly detection [12, 29], interactive information retrieval [26], and visual topic analysis [27].

Classification problems, in particular, can benefit from two pivotal IML approaches: co-active learning [9] and active learning [13, 38]. Co-active learning takes advantage of multiple perspectives and disagreement-based methods to train multiple classifiers on different feature sets, capturing diverse views. The classifiers' disagreements are then identified, and clustering techniques based on proximity to cluster centroids are employed for labeling instances [45]. Active learning (AL) is an interactive approach that focuses on selecting the most informative instances from an unlabeled dataset for labeling. Typically, it involves training a single classifier on a specific feature set. The goal is to minimize labeling effort by selectively querying an oracle for labels on instances that are expected to provide valuable information [38]. AL aims to improve model performance by iteratively incorporating labeled data into the training process. This reduces the "labelling bottleneck" [38] by allowing the model to actively query for the most informative examples. There are three main approaches to active learning [10]: stream-based sampling, membership query synthesis, and pool-based sampling.

Stream-based sampling [5, 13, 15, 35] is employed when data arrives in a continuous stream, necessitating assessment to determine whether annotation is necessary for each incoming data point. The main challenge in designing algorithms for stream-based sampling is that they cannot take into account the overall distribution of the population [10, 14]. Alternatively, membership query synthesis [3] (MQS) involves generating synthetic data points rather than obtaining them from a dataset [3, 37, 38]. Synthetic data can be strategically generated to maximize label informativeness from the oracle. Although MQS can be useful [28] it

has been shown to be ineffective for image-based tasks due to the challenges in generating synthetic query images [6,37].

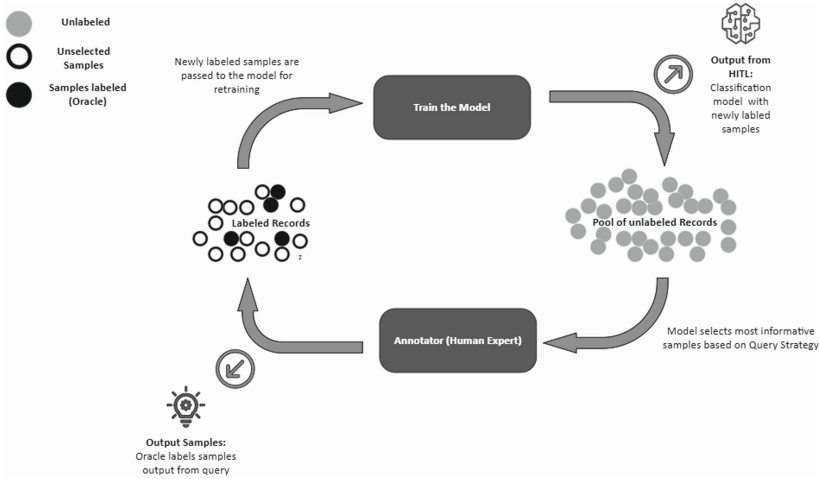


Fig. 2. Pool-based active learning framework.

Pool-based sampling [24,30,38] is the most common AL scenario and is the focus of this paper. Figure 2 shows a typical pool-based AL workflow that involves an *initial dataset* of labeled data points and a large collection of unlabeled data points, known as the *pool dataset*. First, the model is trained on the initial dataset, then a selection method ranks the most informative instances from the pool dataset for labeling by a human annotator, often called an *oracle*. The newly labeled samples are re-introduced into the AL model for re-training. This method is especially effective when used with batch-based training in deep learning approaches [11,30,33,38,46]. Pool-based AL can, however, be computationally expensive as it requires repeated re-training of a model and generating new predictions for the instances in the pool dataset at each iteration.

One way to assess the informativeness of instances in Active Learning is by quantifying the uncertainty associated with predictions made for those instance by a trained model. This is based on the assumption that data points with higher uncertainty in their predictions hold more valuable information, making them candidates for labeling and inclusion in the training set. To measure uncertainty in a classification task, the machine learning model assigns confidence scores to the predicted classes. Typically, the uncertainty in active learning is quantified by calculating the sum of probabilities assigned to the classes with the lowest confidence scores. This approach captures the uncertainty of the prediction, as the sum of probabilities for the lowest confidence classes represents the overall uncertainty [38]. When a model assigns high confidence scores to a specific class, the prediction is considered more certain. Conversely, a lower sum of minimum

class probabilities indicates higher uncertainty in the prediction [38]. Uncertainty can be summarized as:

$$x^*LC = \arg \max_x (1 - P\theta(\hat{y}|x)) \quad (1)$$

where,  $\hat{y}$  represents the most probable label determined by  $\arg \max_y P\theta(y|x)$ .

This formula depicts how uncertainty relates to the prediction made for a specific data point within the context of the data distribution. This method is commonly known as Least Confidence (LC) sampling and provides a way to rank the uncertainty of samples within a distribution [10]. However, LC sampling has a limitation as it solely focuses on the information associated with the most probable label, disregarding information about the remaining label distribution [38].

A margin-based query is a method that addresses the limitations of Least-Confidence queries and is particularly useful for AL for multi-class classification problems [34, 36, 39]. The main idea is that it considers the difference between the first and second most probable labels, resulting in a more accurate measure of the model's confidence in assigning a label. Margin based sampling is defined as:

$$x^*M = \arg \max_x (P\theta(\hat{y}_1|x) - P\theta(\hat{y}_2|x)) \quad (2)$$

where,  $\hat{y}_1$  and  $\hat{y}_2$  are the first and second most probable labels predicted by a model. Consequently, the larger the separation between these two labels, the more certain the model is in assigning a label.

Entropy [39] sampling is another query method that selects instances based on the uncertainty of the model's predictions. Entropy sampling can be defined as :

$$x^*EN = \arg \max_x \left( - \sum_y P_\theta(y|x) \log P_\theta(y|x) \right) \quad (3)$$

where  $y_i$  ranges across all possible annotations. Entropy is a measure of how much uncertainty is present in the predicted class distribution. This means the more uncertain a prediction is, the more information we can gain by including the ground truth for that sample in the training set [10].

Alternatively, query-by-committee is a query strategy used to measure uncertainty of unlabeled data by measuring the agreement between multiple models performing the same task [24, 38]. The premise of this method is that the more disagreement found between predictions on the same data point, the higher the level of uncertainty that data point has and is selected for labeling [24]. This method comes at the cost of computational resources, as multiple models need to be trained and maintained, and each of these needs to be updated in the presence of newly selected training samples [10].

In this study, we focus on pool-based AL and its application to medical images. We explore the effectiveness of different input representations, models of different capacities, and different selection strategies. This comprehensive analysis of AL methods in the context of medical image classification contributes to advancing the field of IML and has the potential to enhance diagnostic and predictive capabilities in healthcare.

### 3 Experimental Setup

This section provides an overview of the experimental design used in this paper, with a particular emphasis on the datasets procured from the MedMNIST collection.

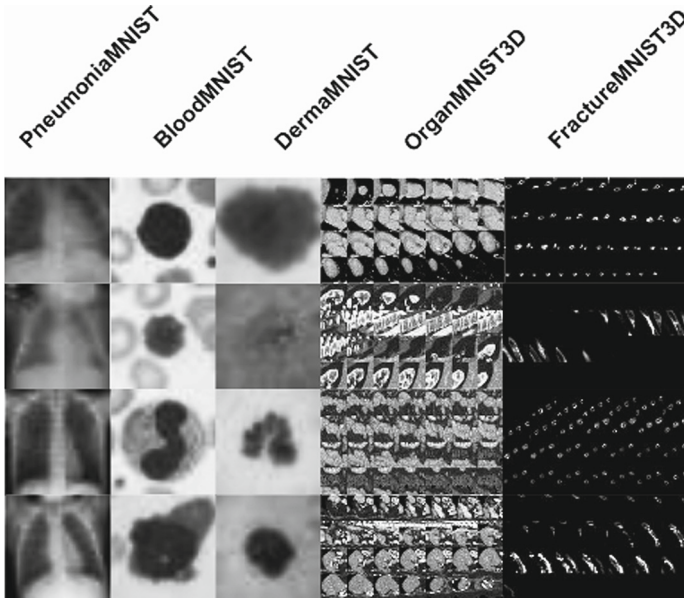
#### 3.1 Datasets and Pre-processing

One of the difficulties in studying machine learning applications in medical imaging is obtaining adequately annotated datasets [4]. In the experiments described in this paper we use publicly available, fully labelled datasets from the MedMNIST collection [44]:

- PneumoniaMNIST: 5,856 greyscale pediatric chest X-Ray images each with a size of  $28 \times 28$  pixels with two classes (pneumonia: 1 and normal cases: 0). Divided in a training set of 4,708 images and a test set of 624 images.
- BloodMNIST: 17,092 RGB images, each with a size of  $28 \times 28$  pixels. The images are categorized into eight classes, representing individual normal cells without any infection, hematologic, or oncologic disease. The dataset includes a training set with 11,959 images and a test set with 3,421 images.
- DermaMNIST: Contains images of common pigmented skin lesions categorized into seven different diseases. The images are represented as RGB and have a size of  $28 \times 28$  pixels. The dataset consists of a training set with 7,007 images and a test set with 2,005 images.
- Organ3DMIST: Consists of 185 CT scans categorized into 11 body organ classes. The images are  $28 \times 28 \times 28$  pixels and represented in greyscale. The training set contains 971 images, while the test set contains 610 images.
- FractureMNIST3D: Includes 1,267 images obtained from approximately 5,000 rib fractures found in 660 CT scans. The images are organized into four clinical categories and are represented as  $28 \times 28 \times 28$  greyscale images. The training set consists of 1,027 images, and the testing set contains 240 images (Fig. 3).

#### 3.2 Experimental Design

The experiment aimed to assess the performance of pool-based AL using various combinations of query strategies (random, margin, and Least-Confidence), model representations (raw image and bottleneck features), and model capacities (random forest, 5-layer CNN, and ResNet50). The AL workflow began by selecting an initial subset of labeled images consisting of 20 samples. Instead of relying on human agents for labeling, a simulated approach was used to iteratively label images queried from the AL model (the MedMNIST datasets are annotated, adhering to the standard practice in AL research). Over the course of 240 iterations, we employed the chosen query strategy to identify four unlabeled instances, add labels to them, and then integrated these instances into the labeled pool of data. At each iteration the pool of labeled data was evaluated



**Fig. 3.** Sample images from each of the five MedMNIST datasets used in these experiments in this paper.

against the MedMNIST-defined test set [44]. Performance evaluation was based on two metrics: the area under the learning curve (AULC) and accuracy (ACC) after 100 iterations. The above process was repeated for each combination of query strategy, model representation, and model capacity.

### 3.3 Image Representations

Image representations refer to the form of input data for machine learning models. In this experiment, two types of image representations were used. The first type, bottleneck feature representation, is a compressed version of intermediate outputs of a neural network. This representation offers a low dimensional representation of the data while preserving the necessary information for classification. In our experiments the bottleneck representations were extracted from the layer immediately prior to the final output layer of a ResNet50 model pre-trained on the Imagenet dataset<sup>1</sup> were used.

The second image representation, raw image representation, simply uses the original images as input to models. This representation contains all the information present in the image, but it may be computationally expensive and require a large amount of memory to process. For the medium and high capacity models the raw images were resized to  $224 \times 224$  pixels.

<sup>1</sup> ResNet50 model pre-trained on the Imagenet dataset. Available at: <https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html>.

### 3.4 Model Capacities

The capacity of a machine learning model refers to its ability to learn and fit training data [7]. A model with low capacity may underfit the training data and fail to capture complex patterns, while a model with high capacity may overfit and perform poorly on unseen data. We use models representative of low, medium, and high capacity:

- **Random forest (low capacity):** This model was trained on the labelled data available at each iteration and optimized using a grid search and 10-fold cross-validation. The selected hyperparameters consisted of max depth = 16 and  $n_{estimators} = 256$ .
- **Shallow CNN (medium capacity):** The shallow CNN model used in the study consisted of two convolutional layers (with sizes of 32 and 64 respectively, kernel sizes of 3 and ReLU activations); two max-pooling layers; and, after flattening, a fully connected layer (of size 64 and ReLU activation) before the output layer. To train this model cross-entropy loss and an SGD optimizer were used with a learning rate of 0.01.
- **ResNet50 (high capacity model):** The ResNet50 model was pre-trained on over a million images from the ImageNet database, making it capable of generating rich feature representations for a wide range of images [20]. To fine-tune the ResNet50 model, cross-entropy loss was used [21] with the SGD optimizer [19, 40].

To leverage the pre-trained weights of a ResNet50 model it is common practice to freeze some of the layers during training to prevent overfitting when generalizing on a new dataset [43]. Freezing layers prevents the gradients from being computed and backpropagated through the layers [43].

### 3.5 Query Strategies

Query strategies constitute a fundamental aspect of AL, serving the purpose of selecting a subset of the most informative unlabeled data samples to improve the model’s accuracy and minimize the number of samples that need to be labeled. In this study three query strategies (random, margin and least-confidence) have been implemented to determine the most suitable combination in the AL framework. The random query method selects data points for labeling without considering their informativeness. In contrast, the margin query method prioritizes labeling data points that the model is most uncertain about, gauging uncertainty by the difference between the two most probable classes [38]. Conversely, the least-confidence method labels data points based on the model’s lowest confidence level, determined by assessing the probability of the most probable class [34].

## 4 Results and Discussion

The objective of this study was to investigate which combination of model representation, model capacity, and query strategy is the most effective when using



active learning for medical image classification. The experiment involved two types of image representations (bottleneck feature and raw image representations); three model architectures (Random Forrest (low capacity), a 5-layer CNN (medium capacity) and a ResNet50 (high capacity)); and three query strategies (random, margin, and least-confidence). The results of our experiments are shown in Tables 1 and 2.

**Table 1.** Experimental results measured using the Area Under Learning Curve (AULC). The best performing approach for each dataset is highlighted in bold.

Representation	Model	Query Strategy	Pneumonia MNIST	Blood MNIST	Derma MNIST	Organ MNIST3D	Fracture MNIST3D
Bottleneck Features	Random Forest	Random	0.8270	0.6399	0.6649	0.6529	0.4351
		Margin	0.8295	0.7144	0.6783	0.6456	0.4109
		Least-Confidence	0.8284	0.6341	0.6789	0.6531	0.4273
Raw Image	ResNet50	Random	0.7960	0.9194	0.6892	0.9027	0.4329
		Margin	<b>0.8614</b>	<b>0.9302</b>	0.6919	<b>0.9211</b>	0.4283
		Least-Confidence	0.8452	0.9262	<b>0.7134</b>	0.9180	<b>0.4542</b>
Raw Image	Shallow CNN	Random	0.8213	0.7421	0.6562	0.7121	0.4098
		Margin	0.7879	0.6571	0.6587	0.7558	0.3922
		Least-Confidence	0.8333	0.7488	0.6602	0.7307	0.4073

**Table 2.** Experimental results measured using the Accuracy metric (ACC %). The best performing approach for each dataset is highlighted in bold.

Representation	Model	Query Strategy	Pneumonia MNIST	Blood MNIST	Derma MNIST	Organ MNIST3D	Fracture MNIST3D
Bottleneck Features	Random Forrest	Random	83.02	62.76	65.97	68.03	40.83
		Margin	84.13	73.66	68.07	68.52	39.58
		Least-Confidence	83.33	61.36	67.98	68.53	42.08
Raw Image	ResNet50	Random	79.81	93.74	69.02	87.51	43.75
		Margin	<b>87.82</b>	<b>96.66</b>	72.15	<b>93.12</b>	40.41
		Least-Confidence	86.70	96.14	<b>72.76</b>	92.89	<b>45.00</b>
Raw Image	Shallow CNN	Random	83.81	77.05	63.48	75.78	37.08
		Margin	80.81	67.52	63.91	81.76	40.83
		Least-Confidence	86.86	79.63	64.66	82.13	40.41
Benchmark	ResNet50		85.70	95.60	73.1	85.70	49.40
	Shallow CNN		83.20	79.42	68.54	81.93	40.12

These results show that the high capacity ResNet50 model using a raw image representation and the margin or least-confidence query strategy achieved the highest accuracy across all experiments. The AL model using bottleneck representation and the low capacity RF model did not exceed the baseline accuracy achieved by Yang et al. [44] (shown in the final rows of Table 2). The convergence of this method is illustrated in Fig. 4.

Figure 5, depicts the high-capacity ResNet50 model using raw image representations alongside three query strategies. The margin query strategy notably converged faster at the 100<sup>th</sup> iteration. However, regardless of the strategy, each

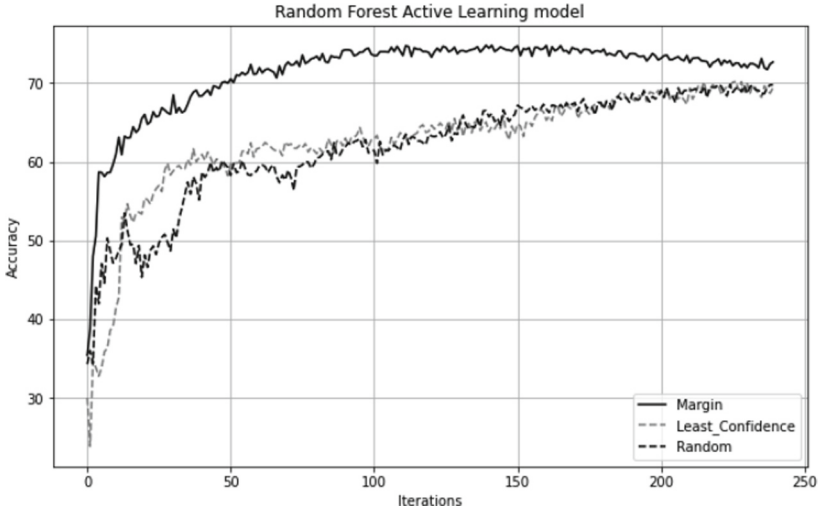


Fig. 4. Learning curves using bottleneck features and RF classifier applied to the BloodMNIST dataset.

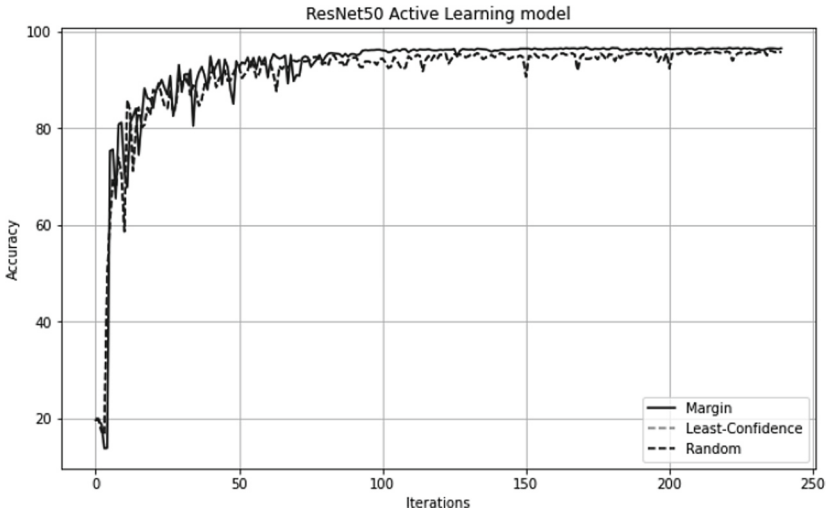
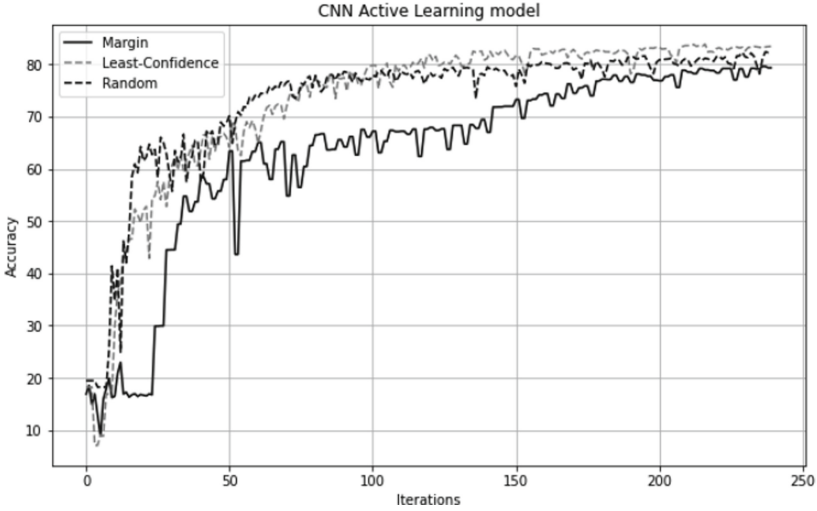


Fig. 5. Learning curves using the ResNet50 model and raw image representations for the BloodMNIST dataset.

method performed well including the random query strategy. The results in Fig. 6 depicts the study using a 5-layer CNN with raw images. For this experiment the least-confidence query method proved to be the most robust method. The large discrepancy between each of the methods can be seen at the 100<sup>th</sup> iteration.



**Fig. 6.** Learning curves using the medium capacity CNN and raw image representations for the BloodMNIST dataset.

## 5 Conclusion

The studies presented in this paper was designed to explore the effectiveness of pool-based AL processes for medical image classification. The investigation considers input representations, model capacities, and selection strategies. The results of the experiments performed using 5 medical imaging datasets demonstrated that using raw image representations with a high capacity ResNet50 model was the most suitable approach across all experiments. Our findings reveal that this combination consistently delivered superior performance when employing the margin query strategy in the cases of PneumoniaMNIST, BloodMNIST, and OrganMNIST3D datasets. However, it's worth noting that the Least-Confidence query strategy exhibited stronger generalization capabilities on the DermaMNIST and FractureMNIST3D datasets. This suggests that the choice of query strategy should be contingent upon the specific image types and anatomical regions.

Overall, the best AL approach requires significantly fewer labeled samples compared to benchmark models and outperformed both ResNet50 and 5-layer CNN models trained over 200 epochs on the entire training dataset, seen in Table 2. Our findings emphasize the potential for efficient and precise classification tasks, particularly in scenarios where obtaining labeled data is limited or expensive. In future work, we intend to explore alternative query strategies and assess the generalizability across a broader spectrum of medical imaging modalities.

## References

1. Aghdam, H.H., et al.: Active learning for deep detection neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3672–3680 (2019)
2. Amershi, S., et al.: Power to the people: the role of humans in interactive machine learning. *AI Mag.* **350**(4), 105–120 (2014)
3. Angluin, D.: Queries and concept learning. *Mach. Learn.* **2**, 319–342 (1988)
4. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020)
5. Atlas, L., Cohn, D., Ladner, R.: Training connectionist networks with queries and selective sampling. In: *Advances in Neural Information Processing Systems*, vol. 2 (1989)
6. Baum, E.B.: Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Trans. Neural Netw.* **20**(1), 5–19 (1991)
7. Belkin, M., et al.: Reconciling modern machine-learning practice and the classical bias-variance trade-off. In: *Proceedings of the National Academy of Sciences*, vol. 1160. no. (32), pp. 15849–15854 (2019)
8. Berg, S., et al.: Ilastik: interactive machine learning for (bio) image analysis. *Nat. Methods* **160**(12), 1226–1232 (2019)
9. Branson, S., Perona, P., Belongie, S.: Strong supervision from weak annotation: Interactive training of deformable part models. In: *2011 International Conference on Computer Vision*, pp. 1832–1839. IEEE (2011)
10. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med. Image Anal.* **71**, 102062 (2021)
11. Cho, J.W., et al.: MCDAL: maximum classifier discrepancy for active learning. *IEEE Trans. Neural Netw. Learn. Syst.* (2022)
12. Chung, M.-H., et al.: Interactive machine learning for data exfiltration detection: Active learning with human expertise. In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 280–287. IEEE (2020)
13. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. *Mach. Learn.* **15**, 201–221 (1994)
14. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: *Machine Learning Proceedings 1995*, pp. 150–157. Elsevier (1995)
15. Dasgupta, S., Kalai, A.T., Tauman, A.: Analysis of perceptron-based active learning. *J. Mach. Learn. Res.* **100**(2) 2009
16. Du, X., Zhong, D., Shao, H.: Building an active palmprint recognition system. In: *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1685–1689. IEEE (2019)
17. Dudley, J.J., Kristensson, P.O.: A review of user interface design for interactive machine learning. *ACM Trans. Interact. Intell. Syst. (TiiS)* **80**(2), 1–37 (2018)
18. Fails, J.A., Olsen Jr, D.R.: Interactive machine learning. In: *Proceedings of the 8th International Conference on Intelligent User Interfaces*, pp. 39–45 (2003)
19. Goyal, P., et al.: Accurate, large minibatch SGD: training imageNet in 1 hour. *arXiv preprint [arXiv:1706.02677](https://arxiv.org/abs/1706.02677)* (2017)
20. He, K., et al.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

21. Ho, Y., Wookey, S.: The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. *IEEE Access* **8**, 4806–4813 (2019)
22. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inform.* **30**(2), 119–131 (2016)
23. Holzinger, A., Plass, M., Holzinger, K., Crişan, G.C., Pinteă, C.-M., Palade, V.: Towards interactive machine learning (iML): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In: Buccafurri, F., Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) *CD-ARES 2016*. LNCS, vol. 9817, pp. 81–95. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-45507-5\\_6](https://doi.org/10.1007/978-3-319-45507-5_6)
24. Hu, R., Namee, B.M., Delany, S.J.: Sweetening the dataset: using active learning to label unlabelled datasets. In: *Proceedings of AICS*, vol. 8, pp. 53–62 (2008)
25. Jiang, L., Liu, S., Chen, C., Recent research advances on interactive machine learning: Recent research advances on interactive machine learning. *J. Vis.* **22**, 401–417 (2019)
26. Kelly, D., et al.: Methods for evaluating interactive information retrieval systems with users. *Found. Trends® Inf. Retrieval*, **30**(1–2), 1–224 (2009)
27. Kim, M., et al.: Topiclens: efficient multi-level visual topic exploration of large-scale document collections. *IEEE Trans. Vis. Comput. Graph.* **230**(1), 151–160 (2016)
28. King, R.D., et al.: The automation of science. *Science* **3240**(5923), 85–89 (2009)
29. Kose, I., Gokturk, M., Kilic, K.: An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Appl. Soft Comput.* **36**, 283–299 (2015)
30. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: *Machine Learning Proceedings 1994*, pp. 148–156. Elsevier (1994)
31. Lin, X., Parikh, D.: Active learning for visual question answering: an empirical study. *arXiv preprint [arXiv:1711.01732](https://arxiv.org/abs/1711.01732)* (2017)
32. Liu, Z., et al.: Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE Trans. Vis. Comput. Graph.* **230**(1), 321–330 (2016)
33. McCallum, A., Nigam, K., et al.: Employing EM and pool-based active learning for text classification. In: *ICML*, vol. 98, pp. 350–358. Citeseer (1998)
34. Rawat, S., et al.: How useful is image-based active learning for plant organ segmentation? *Plant Phenomics*, 2022 (2022)
35. Sabato, S., Hess, T.: Interactive algorithms: pool, stream and precognitive stream. *J. Mach. Learn. Res.* **18**, 1–39 (2017)
36. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden Markov models for information extraction. In: Hoffmann, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (eds.) *IDA 2001*. LNCS, vol. 2189, pp. 309–318. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-44816-0\\_31](https://doi.org/10.1007/3-540-44816-0_31)
37. Schumann, R., Rehbein, I.: Active learning via membership query synthesis for semi-supervised sentence classification. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 472–481 (2019)
38. Settles, B.: *Active learning literature survey* (2009)
39. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **270**(3), 379–423 (1948)
40. Smith, S.L., et al.: Don’t decay the learning rate, increase the batch size. *arXiv preprint [arXiv:1711.00489](https://arxiv.org/abs/1711.00489)* (2017)
41. Teso, S., Kersting, K.: Explanatory interactive machine learning. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 239–245 (2019)

42. The Royal College of Radiologists. Rcr clinical radiology census report 2021 (2021). <https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-census-report-2021> . Accessed 27 Feb 2023
43. Wang, Y., et al.: Efficient DNN training with knowledge-guided layer freezing. arXiv preprint [arXiv:2201.06227](https://arxiv.org/abs/2201.06227) (2022)
44. Yang, J., et al.: MedMNIST v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **100**(1), 41 (2023)
45. Betül Yüce, A., Yaslan, Y.: A disagreement based co-active learning method for sleep stage classification. In: 2016 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–4. IEEE (2016)
46. Zhan, X., et al.: A comparative survey of deep active learning. arXiv preprint [arXiv:2203.13450](https://arxiv.org/abs/2203.13450) (2022)
47. Zhang, Y., Lease, M., Wallace, B.: Active discriminative text representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)