ACTIVE LEARNING WITH DATA AUGMENTATION FOR ORGAN CLASSIFICATION IN MEDICAL IMAGES

Candidate Number: **REDACTED**

ABSTRACT

Medical image processing, and more specifically classification, is a crucially important step in both diagnosis and treatment, and consequently deep learning approaches have been explored to accelerate the process. However, such approaches naturally struggle at this task for two key reasons: suitable data is hard to come by due to privacy concerns and other real-world factors; and image labelling requires significant expertise and time. The scarcity of suitable, well-labelled medical images thus motivates approaches that maximise the information gained from each and every label. In this paper, we assess the merits of active learning (using the model's predictions on unlabelled data to choose which new labels would be the most informative) and data augmentation (applying small changes to some images to increase dataset diversity) for this purpose on the task of organ classification. We demonstrate that training using active learning gives the most reliable improvements on the base training method. We also show that a combination of active learning and data augmentation is not as effective as either individually, and we propose an explanation as to why this might be the case.

***Index Terms*—** active learning, data augmentation, organ identification

1. INTRODUCTION

Medical image processing captures a broad range of tasks that are vitally important in multiple areas of medicine, including diagnostics [1] and treatment [2]. Consequently, there is a large body of research dedicated to applying machine learning (ML) techniques to such tasks [3, 4], but these approaches are dependent on having high-quality data readily available.

Availability of data is inherently an issue for medical research: data privacy norms make access to large amounts of data difficult to attain; moreover, the data that is available may lack variety for medically relevant reasons (e.g. medical imaging costs money, so is reserved for cases where practitioners suspect the existence of an abnormality); finally, complex medical research datasets require a large amount of expert labelling time. Medical data can thus be relatively scarce, limited, and of varying quality.

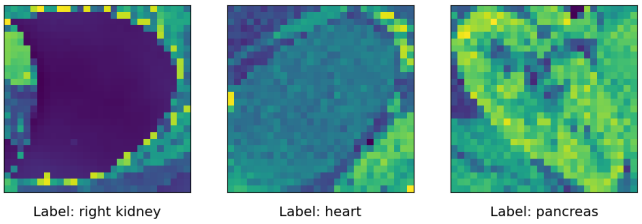


Fig. 1: Examples from the MedMNIST dataset [5], with the associated label shown.

This problem is exacerbated by the complexity of the data - radiological data is often 3D (e.g. CT or MRI scans [5]), and related tasks may require pixel-level precision (e.g. tumour segmentation [6]). This motivates research into data efficiency techniques which aim to maximise model accuracy even when trained on limited data.

Data augmentation and active learning are two existing approaches for doing this.

Data augmentation involves diversifying or expanding a dataset by applying small transformations which importantly don't change the label assigned to the data. Example augmentations for images include affine transformations, colour changes, and noising. Previous work has shown how even simple augmentations can improve accuracy and sample efficiency [7, 8].

Active learning is a modification of the training setup inspired by how humans learn [9, 10]. A model is trained on a small initial dataset, and then a query is generated which contains unlabelled samples that would be informative to the model if labelled. The query is then labelled and added to the dataset for the next training iteration. Effective query selection can, much like DA, lead to better efficiency [11].

Previous work has explored the uses of each individually for medical image processing [12, 13]. Our contribution to the literature is exploring the combination of the two: we examine whether applying active learning to an augmented dataset, and whether applying augmentations after the active learning step, yield further improvements in a data-scarce training setup. We investigate the merits of these two approaches on the task of classifying images from the MedMNIST v2 dataset [5] which contains 2D CT scans of various organs.

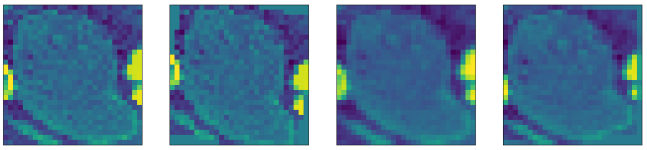


Fig. 2: Examples of augmented images from the MedMNIST dataset, from L to R: *unaugmented*; *rotated*; *blurred*; and *deformed*.

2. BACKGROUND

In this section, we introduce important concepts in data augmentation (DA) and active learning (AL).

2.1. Data Augmentation

DA is an example of a regularisation technique; it seeks to improve the generalisability and robustness of the model without the need for additional labelled samples. DA techniques can vary significantly in complexity, from traditional label-preserving transformations, to methods based on learned augmentation algorithms or data synthesis with generative models [14]. The traditional transformations used include geometric transformations (scaling, resizing, cropping, etc.) and photometric transformations (motion blur, optical noise, colour jittering, etc.). The methods included in this paper are:

- Rotation by $\pm 30°$
- Gaussian blur
- Elastic deformation

Augmentations should be chosen carefully, both because different image modalities respond better to different augmentations [15], and because the task being learned may preclude certain augmentations: for example, we avoid using region cutout (masking portions of the image with zero values) as it could occlude the features of the image that identify the organ, particularly in such a small image.

2.2. Active Learning

Current AL methods can broadly be categorised into two groups: query-acquiring (also known as pool-based) methods, and query-synthesising methods. Both share the same core loop: a query of samples is generated; labels for those samples are obtained from an oracle (an information source, e.g. a human evaluator); these new labelled data points are added to the training dataset; and the loop repeats. Both methods aim to increase how informative the query samples are, thus maximising the performance of the learner while minimising the number of labels required from the oracle. In this paper, we focus solely on pool-based AL.

The pool-based method is implemented by selecting from a pool of unlabelled samples according to some acquisition function, which, given a set of samples, returns a smaller set using a chosen metric. One of the most common acquisition functions used is uncertainty [16], which involves selecting the samples about which the classifier is most uncertain according to some uncertainty metric. In this paper, we use the Shannon entropy of the model's outputs:

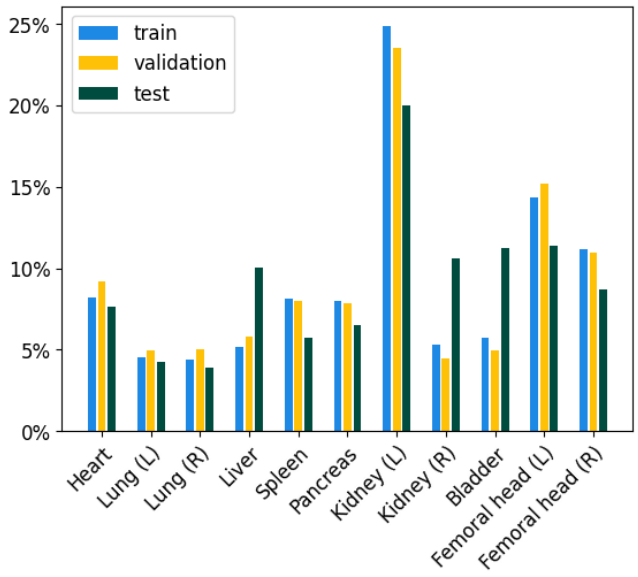
$$H(\boldsymbol{y}) = -\sum_{i=1}^{n} y_i \log_2 y_i \tag{1}$$

The oracle is often a human 'in-the-loop', labelling query samples as they are generated. However, in the case of medical data, labelling requires significant expertise. A common alternative is to use an already-labelled dataset as the pool from which queries are generated; the oracle is the assigned labels from the dataset. This is the approach we take in this paper.

3. DATA

The MedMNIST dataset [5] contains single-channel images of size 28x28 of CT scans for 11 different organs (examples in Fig. 1). The dataset is split into train (13,932 samples), validation (2452), and test (8827) sets, and the class distributions within these splits (found in Fig. 3) are quite closely aligned. However, the test set does contain relatively more liver, right kidney, and bladder scans, and fewer left kidney and left femoral head scans: thus, this may be a source of minor distributional shift. We applied limited processing to the image data, only centring it and converting the labels into one-hot encoding.

Fig. 3: Class distributions compared across the train, validation, and test splits.



4. METHOD

We apply various training regimes to a convolutional model, assessing their efficacy with accuracy on the test set. In this section, we outline the training regimes used and how they

differ from each other. We also explain the model architecture and detail how we implement each training setup.

4.1. Training Regimes

In order to assess the efficacy of combining AL with DA, we compare the following six regimes.

{1} *Training on the full dataset* of 13,932 samples: this provides an upper bound on what our model architecture is capable of given sufficient data.

{2} *Training on a limited dataset* containing just 1% (139 samples) of the full dataset: this provides a baseline to which the various methods can be compared.

{3} *Training with DA*: we randomly apply the augmentations listed in section 2.1 to the training dataset.

{4} *Training with AL*: we train a small model on an initial dataset containing just 0.05% (7 samples) of the full dataset; we then get model predictions for the 'unlabelled' pool dataset (the remaining 99.95%) and calculate the uncertainty (entropy) for each prediction; we take the four samples with the highest uncertainties and add them to the training dataset, along with their labels; we repeat this process until the dataset contains 139 samples, and then train a model from scratch on this data.

{5} *Training with DA applied before AL steps*: identical to method {4}, but a random augmentation is applied to all data (training and pool data) from the offset. We call this method Active Learning with Data Augmentation (ALDA).

{6} *Training with DA applied after AL steps*: identical to method {5}, but the random augmentations are applied only when the model trains on the data; thus, training is carried out on augmented data, but the pool dataset contains unaugmented data. We call this method Active Learning with Post-query Augmentation (ALPA).

4.2. Model Architecture

The convolutional model **C** is initialised with a parameter f which adjusts the number of features learned in the convolutional layers, as well as the number of nodes in the fully connected layers.

The model consists of three convolutional blocks, each of which contains two iterations of the following: 3x3 convolutional layer with stride 1; ReLU activation function; and batch normalisation. Between the convolutional blocks are 2x2 max-pooling layers with stride 3. These convolutional layers convert the 1x28x28 image into a $36f$x3x3 feature map, which is flattened and passed to three fully connected (FC) layers, with $8f$, $2f$, and 11 nodes respectively; a ReLU function is applied before layers FC-$2f$ and FC-11, and the Softmax function is applied to the outputs of the final layer, producing probabilities for each class as an output.

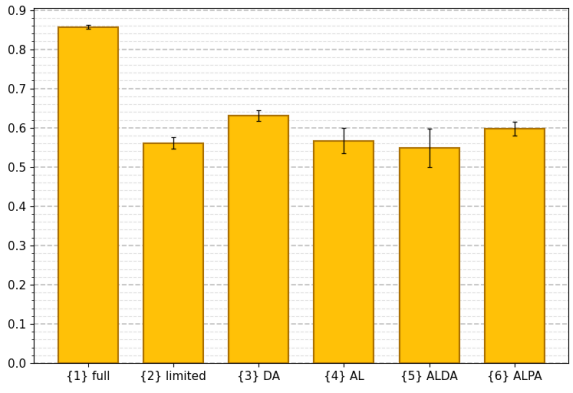
4.3. Experimental Setup

Most of the models are trained with the input parameter f=32; exceptions are in method {1}, where too many features decreases model performance, so f=16 is used, and the smaller models in methods {4}-{6}, which all use f=8. We use the cross-entropy loss function and optimise using AdamW with a learning rate of 0.001.

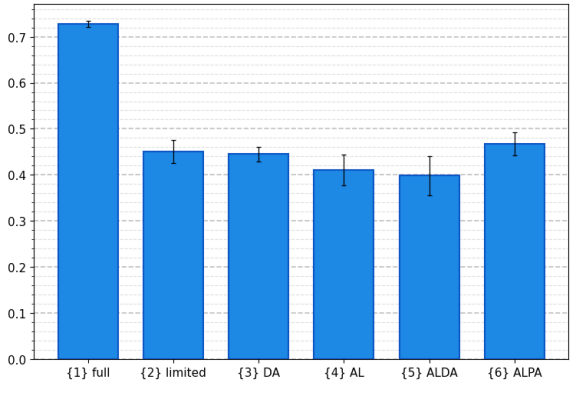
Full size models ($f \geq 16$) are trained for 40 epochs. The small models within the AL loop are trained in a different way: the model generates a query every n epochs, where n is initialised to 6; we reinitialise the model from scratch every three epochs, to avoid training the model excessively on the initial dataset; for each model reset, n is incremented by 1, to allow sufficient training time on the expanding dataset.

5. RESULTS AND DISCUSSION

Fig. 4 below charts the accuracies of the various training approaches listed in section 4.1. Models are trained as described in section 4.3 and evaluated on both the validation set (during training, maximum accuracy used) and the test set (using final model after training).



(a) Validation accuracies



(b) Test accuracies

Fig. 4: Mean accuracies with deviations on validation and test sets, taken across five different models per training regime.

All training approaches decline in performance from the validation to test set: this is likely due to overfitting towards the end of the training process, which doesn't affect the maximum accuracy attained on validation but does impact the final model, and slight distributional shift identified in section 3.

The full training of $\{1\}$ drastically outperforms all other methods; this reflects the importance of having adequate training data for model efficacy. Moreover, none of methods $\{3\}$-$\{6\}$ offers significant improvements on the baseline regime tested in $\{2\}$. The remainder of this section analyses each of these methods in turn and compares them to the baseline.

Method $\{3\}$ is more effective on the validation dataset than $\{2\}$, but offers no improvements in accuracy on the test set. However, the deviation of the models is decreased; this might suggest that using DA does improve model robustness to initial conditions, as models initialised differently reach a similar accuracy more consistently. The improvement seen on only the validation dataset indicates that the efficacy of DA is limited by the distributional similarities between training and validation. However, time and compute limitations led us to test only a narrow range of augmentations, with minimal tuning; future work could implement more sophisticated augmentation techniques in this setting (e.g. learned augmentations or data synthesis [14]).

Method $\{4\}$ meanwhile offers marginal improvements in validation but decreases model efficacy in testing. This, alongside the increased deviation, suggests that the training environment is inadequate for carrying out AL. Possible explanations include: using too rudimentary an uncertainty measure; poor selection of AL hyperparameters; and insufficient data. Similar experiments should be carried out using more advanced acquisition functions (e.g. algorithms that incorporate both uncertainty and diversity, like BADGE [17] or BatchBALD [18]).

The least effective of all methods is $\{5\}$, the naive combination of AL with DA. Applying augmentations before the AL step causes a decrease in accuracy (validation and testing) and an increase in variability. The biggest concern levelled at this method, and the most likely reason for its low efficacy, is that uncertainties are calculated on augmented data; higher uncertainty values are likely to be assigned to data which has been more heavily augmented, which means the query-selection process favours less recognisable images.

This motivates the post-query augmentation approach of $\{6\}$, which avoids calculating uncertainties on augmented data. This is the only method which averages a higher accuracy than the baseline $\{2\}$ on both validation and test data. The initial conclusion that we draw from the relative success of this method is that there is promise in the combination of AL and DA techniques.

The results laid out above indicate that the combination of AL and DA methods could improve data efficiency, if implemented correctly. We recommend that future research investigates the combination of more advanced AL and DA techniques, as described above. Further evaluation of this approach could extend to other datasets, examine the amount of training data needed to make this approach feasible, or explore the impact of distributional shift on DA and AL techniques.

6. CONCLUSION

In this work, we propose a method for combining techniques from AL and DA to improve data efficiency in the task of classifying organs in CT scan image data. We find that the naive approach of applying augmentations before the query-generation step decreases model accuracy; this motivates an alternative combination in which the AL steps are applied on an unaugmented dataset, with the augmentations applied only when the model is training. This approach shows some promise, marginally outperforming all other baseline methods on the test set, but requires further investigation using more sophisticated DA and AL methods.

7. REFERENCES

[1] Baidaa Mutasher Rashed and Nirvana Popescu, "Critical Analysis of the Current Medical Image-Based Processing Techniques for Automatic Disease Evaluation: Systematic Literature Review," *Sensors*, vol. 22, no. 18, pp. 7065, Sept. 2022.

[2] Mohammed Yusuf Ansari, Alhusain Abdalla, Mohammed Yaqoob Ansari, Mohammed Ishaq Ansari, Byanne Malluhi, Snigdha Mohanty, Subhashree Mishra, Sudhansu Sekhar Singh, Julien Abinahed, Abdulla Al-Ansari, Shidin Balakrishnan, and Sarada Prasad Dakua, "Practical utility of liver segmentation methods in clinical surgeries and interventions," *BMC Medical Imaging*, vol. 22, no. 1, pp. 97, May 2022.

[3] Rakesh Kumar, Pooja Kumbharkar, Sandeep Vanam, and Sanjeev Sharma, "Medical images classification using deep learning: a survey," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 19683–19728, July 2023.

[4] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha, "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities," *Healthcare Analytics*, vol. 4, pp. 100216, Dec. 2023.

[5] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni, "MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, no. 1, pp. 41, Jan. 2023, arXiv:2110.14795 [cs, eess].

[6] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, pp. 654, Jan. 2024.

[7] Luke Taylor and Geoff Nitschke, "Improving Deep Learning using Generic Data Augmentation," 2017, Publisher: [object Object] Version Number: 1.

[8] Cherry Khosla and Baljit Singh Saini, "Enhancing Performance of Deep Learning Models with different Data Augmentation Techniques: A Survey," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, June 2020, pp. 79–85, IEEE.

[9] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active Learning with Statistical Models," Feb. 1996, arXiv:cs/9603104.

[10] Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8410–8415, June 2014.

[11] Xiongquan Li, Xukang Wang, Xuhesheng Chen, Yao Lu, Hongpeng Fu, and Ying Cheng Wu, "Unlabeled data selection for active learning in image classification," *Scientific Reports*, vol. 14, no. 1, pp. 424, Jan. 2024.

[12] Jingwen Wang, Yuguang Yan, Yubing Zhang, Guiping Cao, Ming Yang, and Michael K. Ng, "Deep Reinforcement Active Learning for Medical Image Classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, Eds., vol. 12261, pp. 33–42. Springer International Publishing, Cham, 2020, Series Title: Lecture Notes in Computer Science.

[13] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, pp. 106391, Jan. 2023.

[14] Alhassan Mumuni and Fuseini Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, pp. 100258, Dec. 2022.

[15] Evgin Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12561–12605, Nov. 2023.

[16] Samuel Budd, Emma C. Robinson, and Bernhard Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, pp. 102062, July 2021.

[17] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal, "Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds," 2019, Publisher: [object Object] Version Number: 2.

[18] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal, "BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning," 2019, Publisher: [object Object] Version Number: 2.