Enhancing Model Generalisability through Sampling Diverse and Balanced Retinal Images

Tianfeng Zhou[1] and Yukun Zhou[2,3]

[1] School of Electronic Information, Central South University, China
[2] UCL Institute of Ophthalmology, University College London, UK
[3] Centre for Medical Image Computing, University College London, UK
yukun.zhou.19@ucl.ac.uk

Abstract. Model generalisability, i.e. performance on multiple unseen datasets, can be improved by training on large volumes of annotated data, from which models can learn diverse representations. However, annotated medical data is limited due to the scarcity of expertise. In this work, we present an efficient data sampling pipeline to select DIVerse and bAlanced images (DataDIVA) from image pools to maximise model generalisability in retinal imaging. Specifically, we first extract image feature embeddings using off-the-shelf foundation models and generate embedding clusters. We then evenly sample images from those diverse clusters and train a model. We run the trained model on the whole unlabelled image pool and sample the remaining images from those classified as rare categories. This pipeline aims to sample the retinal images with diverse representations and mitigate the unbalanced distribution. We show that DataDIVA consistently improved the model performance in both internal and external evaluation, on six public datasets, with clinically meaningful tasks of referable diabetic retinopathy and glaucoma detection. The code is available at https://doi.org/10.5281/zenodo.12674694.

Keywords: Model generalisability · Foundation model · Diverse representation · Unbalanced distribution · Data sampling.

1 Introduction

In recent years, there has been an explosion of interest in the application of deep learning models to various medical tasks including disease diagnosis, lesion localisation and segmentation, and biomarker discovery. Several studies have demonstrated that medical models can achieve comparable or superior performance to humans across multiple tasks [9,28]. However, the real-world utility of these models is limited by poor generalisability outside of their original development environment [10,16]. Generalisability on unseen sets can be improved by training on large-scale annotated data, however often only a small proportion of data can be selected for labelling due to limited annotation resources. It has therefore been a long-standing challenge to select a limited subset of informative data for labelling and model training which maximise model generalisability.

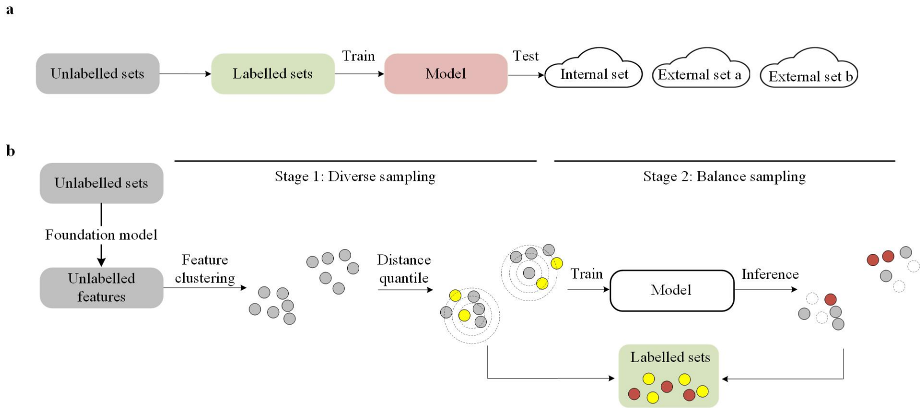

Fig. 1. Schematic of a. model training and evaluation on multiple sets and b. DataDIVA. DataDIVA consists of diverse sampling and balance sampling to maximise the model generalizability. After feature extraction via foundation models, diverse sampling clusters the features and selects data points in quantile distances to cluster centroids (yellow points). Balance sampling chooses the data points that are classified as rare categories (red points). Concentric circles show the distance quantile to centroids.

Previous work samples the informative subset that approximates the distribution of the whole dataset and includes maximised information. The commonly used strategies can be categorised into geometry based [23,27,4], uncertainty based [11,25,2], decision boundary based [8,19], and their combination [29,26,1]. Geometry-based methods assume that data points close to each other in the feature space tend to have similar properties, thus selecting subsets that mimic full dataset distribution. Uncertainty-based methods assume that samples with high uncertainty (e.g. high entropy) contain the information not yet learnt by models, and therefore include them in model training. Decision boundary-based methods find data points distributed around the decision boundary to improve the model performance in vague cases. Combined strategies aim to achieve a trade-off between properties in data selection [29,26,1]. These sampling methods have been widely used in data-efficient research, including active learning [24,22] and continual learning [18,5]. Despite progress in this area, most of these methods are usually designed with the target of maximising the performance on a single set which has the same data distribution as the unlabelled data pool. The performance on multiple unseen sets is more relevant for real-world applications but remains under investigation. Moreover, the data unbalanced issue, commonly observed in medical AI which decays the model performance, has not been thoroughly considered in the data sampling methods. Finally, several powerful medical AI resources such as the foundation models [31,14] have been proposed and validated to be capable of extracting good feature representation. The efficacy of using the foundation model in facilitating the data sampling has not been studied.

In this work, we formalise the data selection problem in clinical scenarios, aiming to select a data subset from large-scale unlabelled and unbalanced data for labelling and model training, in a manner that maximises model performance in internal and external evaluation, as shown in Figure 1a. We present an efficient sampling pipeline to select data with diverse representation and relatively balanced distribution in two steps, depicted in Figure 1b. The first step selects diverse data in feature clusters and the second step samples the data predicted as rare categories. The focus of this work is on the data sampling strategy, which is compatible with other data diversity-based techniques, such as data augmentation and generation.

We summarise our contributions. 1) We present DataDIVA, a pipeline for sampling retinal images with diverse and balanced distribution, for real-world clinical scenarios where labelling resources are limited and generalisation on unseen data is necessary for reliable and robust applications. 2) We incorporate the open-source medical AI resource, i.e. foundation model, in the data sampling pipeline and demonstrate its potential in guiding data sampling. 3) We show that DataDIVA achieves improved performance compared to several competitive baselines on multiple datasets, with different network backbones, in various clinically meaningful tasks with retinal images.

2 Methods

2.1 Problem definition

We define an unlabelled and unbalanced set $\mathcal{D}_u = \{x_i\}_{i=1}^{N}$ with N data points, an internal evaluation set $\mathcal{D}_{in}$ drawn from the same distribution as $\mathcal{D}_u$, and a list of external sets $\mathcal{D}_{out}$ drawn from separate distribution and unseen at training time. We sample and label a subset with M data points, $\mathcal{D}_s = \{x_j, y_j\}_{j=1}^{M}$ from $\mathcal{D}_u$ using selection strategy S, i.e. $\mathcal{D}_u \overset{S}{\to} \mathcal{D}_s$, where $M \ll N$. A model f is trained on $\mathcal{D}_s$ to maximise the performance in internal and external evaluations,

$$\mathcal{D}_s = argmax_{(x,y)\in(\mathcal{D}_{in}\cup\mathcal{D}_{out})} argmax_{(x,y)\in\mathcal{D}_s} T(f(x), y) \tag{1}$$

where $T(\cdot)$ indicates target function. The selection strategy S can be either a fixed selection criteria (e.g., entropy ranking [24]) or a network [27]. The $\mathcal{D}_s$ can be sampled in multiple steps, each selecting a proportion of data to successively update f, such as active learning. However, such iterative update significantly increases computation complexity and the selection strategy S may not represent well the unseen external sets $\mathcal{D}_{out}$. In this paper, we present DataDIVA, an efficient data selection pipeline which finds an optimised solution for Eq.1 by sampling a $\mathcal{D}_s$ with diverse representation and balanced distribution.

2.2 Sampling data with diverse representation

Unlike natural images (e.g. ImageNet-1k) where there is significant image variation according to object category and localisation, medical images in a specific modality (e.g. retinal fundus photographs) show much less variation due to the standardisation of imaging protocols. The variation across medical images highlights the difference in populations and imaging devices [21,3], which are highly relevant to domain shift and model generalisability. Following the assumption of geometry-based methods [23,27,4], i.e. data points close to each other in the feature space tend to have similar properties, we sample the data $\mathcal{D}_{diverse}$ with diverse distribution in feature space to capture such variation.

To generate a good feature space, we use the powerful open-source foundation model (e.g. RETFound [31] for retinal fundus photographs) to extract the features $\mathcal{F}_u = \{F_i\}_{i=1}^{N}$ from $\mathcal{D}_u$. Features $\mathcal{F}$ are then grouped into K clusters via clustering techniques like k-means [15], each cluster encompassing the feature points with similar representation. Within each cluster, we calculate the distance $\left\{\left|F_i^j - C_i\right|^2\right\}_{j=1}^{|C_i|}$ between the cluster centroid $\{C_i\}_{i=1}^{K}$ and feature points belonging to that cluster. A large distance $\left|F_i^j - C_i\right|^2$ corresponds to samples different from the centroid, while small distances indicate similar samples. After removing potential outliers, the range between maximum and minimum distance describes the data diversity within each cluster.

A diverse sampling strategy is generally preferred to increase diverse representation. The balance of $\mathcal{D}_{diverse}$ matching $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ can be achieved by enhancing inter-cluster and intra-cluster diversity. DataDIVA includes two steps for sampling in diverse representation. First, for each cluster, we will evenly sample the data for inter-cluster diversity considering clusters represent separate representations. Each cluster will contain M/K data points. Second, within each cluster, we rank the distance list from small to large and split them into five subgroups in quantile and then sample $\mathcal{D}_{diverse}$ from the five subgroups evenly. With such a strategy, DataDIVA is able to sample $\mathcal{D}_{diverse}$ with highly diverse feature representation.

2.3 Sampling data with balanced distribution

The unbalanced distribution commonly exists in medical AI and biases the model performance. In large-scale clinical datasets, some categories significantly outnumber others, e.g. health category in EyePACS and AIROGS. This affects model generalisability by causing biased predictions towards the majority class and decreasing the performance in detecting disease cases. To mitigate the unbalanced challenge, we aim to sample a subset $\mathcal{D}_s$ better representative of rare categories.

It is infeasible to understand the real label distribution of unlabelled $\mathcal{D}_u$, so we design DataDIVA as a two-step sampling pipeline. After we sample $D_{diverse}$, we first train a model f_{inital} and run it on all remaining $\mathcal{D}_u$. Due to the unbalanced issue, the prediction easily tends to the major categories. A straightforward solution is to select the data points that are classified as rare categories by f_{inital} in the second step. Although there are certain false positive cases, i.e. major categorical data that are wrongly classified as rare categories, this operation may secure a higher proportion of rare categories in sampled data $D_{balance}$. We finally combine the $D_{diverse}$ and $D_{balance}$ as D_s for labelling and model training to improve the model generalisability.

3 Experiments

3.1 Experiment setting

Data. We evaluate the efficacy of DataDIVA using clinically meaningful tasks, including referable diabetic retinopathy (DR) detection and glaucoma detection. **In referable DR detection**, we use three publicly available DR datasets, the EyePACS [12], Kaggle APTOS-2019[1], and IDRiD [20] for model training and evaluation. EyePACS includes 88,702 colour fundus photographs collected in the US and by multiple imaging devices. 35,126 images are set for training and 53,576 for testing. We regard the training set as the unlabelled pool $\mathcal{D}_u$ and keep the test set for internal evaluation. The objective is to select a subset $\mathcal{D}_s$ from $\mathcal{D}_u$ for data labelling and model training, so as to maximise the model performance in internal and external evaluation. The models are externally evaluated on Kaggle APTOS-2019 and IDRiD which have clear differences from EyePACS in both demographics and imaging devices. Following clinical definition, the first two categories (no DR and mild DR) are grouped as non-referable DR and the other three are grouped as referable DR. **For glaucoma detection**, we include AIROGS [6], REFUGE [17], and ORIGA [30]. We split the available AIROGS data (101,442 images) into 70%:30%, where 70% works as the unlabelled pool $\mathcal{D}_u$ and 30% for internal evaluation. The REFUGE and ORIGA, two benchmarks commonly used for glaucoma detection, are used for external evaluation. For data labelling and model training, we sample 600 images (about 1.7% of the EyePACS train set) for referable DR detection and 1200 images (about 1.7% of the AIROGS train set) for glaucoma detection. The sampled images are split into 80%:20% as training and validation sets for model training.

Network architecture and implementation. We use ResNet-50 [13] and ViT-large [7] as network backbones. We initially load the ImageNet weights for ResNet-50 and use RETFound weights [31] for ViT-large. The sampled images are used to fine-tune the models. We use RETFound to extract features from colour fundus photographs for DataDIVA. We compare the proposed DataDIVA with random sampling, as well as highly relevant and competitive baselines, including CoreSet sampling [23] and ALFA-Mix [19]). The images sampled by CoreSet and ALFA-Mix are obtained via publicly available code repository[2], using two-step image sampling similar to DataDIVA. The data quantity, model architecture, and hyper-parameters were standardised in all cases to achieve a fair comparison, as detailed in shared codes. Within each cluster, we calculate the distance between data points and the cluster centroid. To remove data points

[1] https://www.kaggle.com/competitions/aptos2019-blindness-detection/data

[2] https://github.com/AminParvaneh/alpha_mix_active_learning

Table 1. Ablation study for DataDIVA. The left side includes results on diabetic retinopathy detection while the right side shows results on glaucoma detection. $D_{diverse}$ samples all images with diverse sampling. $D_{balance}$ randomly selects half images for initial model training and selects the second half in balance. The proposed DataDIVA combines $D_{diverse}$ and $D_{balance}$ strategies. The sample quantity equals for each method.

Internal-EyePACS-ViT				Internal-AIROGS-ViT		
Method	F1-score	AUROC	AUPR	F1-score	AUROC	AUPR
$D_{diverse}$	0.49±0.01	0.76±0.02	0.73±0.02	0.22±0.04	0.82±0.01	0.60±0.02
$D_{balance}$	0.33±0.15	0.75±0.02	0.70±0.03	0.44±0.04	0.88±0.01	0.70±0.03
DataDIVA	**0.50±0.02**	**0.77±0.01**	**0.74±0.01**	**0.46±0.04**	**0.90±0.02**	**0.72±0.02**
External-APTOS2019-ViT				External-REFUGE-ViT		
$D_{diverse}$	0.75±0.09	0.88±0.04	0.88±0.03	0.46±0.14	0.81±0.17	0.79±0.09
$D_{balance}$	0.56±0.11	0.73±0.05	0.74±0.04	0.60±0.07	0.89±0.01	0.87±0.02
DataDIVA	**0.83±0.04**	**0.92±0.02**	**0.92±0.02**	**0.64±0.12**	**0.93±0.01**	**0.91±0.01**
External-IDRiD-ViT				External-ORIGA-ViT		
$D_{diverse}$	0.84±0.03	0.90±0.02	0.88±0.02	**0.30±0.17**	0.76±0.09	0.70±0.07
$D_{balance}$	0.57±0.26	0.86±0.02	0.84±0.03	0.28±0.12	**0.77±0.02**	0.70±0.03
DataDIVA	**0.86±0.02**	**0.92±0.01**	**0.90±0.02**	0.29±0.15	0.75±0.04	**0.71±0.03**

that are extremely far away from the centroid (defined as outliers in this paper), we set a threshold of 95% and removed the data points that were distributed farther than 95% of the distance distribution. We use Tesla T4 GPUs (16GB) for model training and evaluation in all experiments. The model uses the colour fundus photographs as input and outputs the probability for each category. All training images were preprocessed with AutoMorph [32] and resized to $(256, 256)$. We run model training and evaluation with four random seeds to calculate the mean performance and standard deviation. **Evaluation metrics.** Model performance is reported using Area Under the Receiver Operating Curve (AUROC), Area Under the Precision-Recall curve (AUPR), and F1-score.

3.2 Experiment results

Ablation study on diversity and balance. We investigated the efficacy of diverse sampling and balance sampling of DataDIVA in Table 1. $D_{diverse}$ indicate all the samples were obtained in diverse sampling. $D_{balance}$ first randomly selected half images for initial model training and selected the second half with balanced sampling. DataDIVA ($D_{diverse} + D_{balance}$) outperformed the ablation methods on both referable DR detection and glaucoma detection, which demonstrated the efficacy of the two components.

Performance comparison in glaucoma detection. Table 2 compares the performance of DataDIVA and competitive baselines in referable glaucoma detection. The three sampling methods (i.e. CorSet, ALFA-Mix, and DataDIVA) in general all outperformed random sampling which demonstrated the efficacy of specifically designed methods. DataDIVA performed slightly better than the baseline ALFA-Mix in the internal set while clearly outperforming all baselines in external evaluations. This highlights the enhanced model generalisability

Table 2. Model performance on glaucoma detection with model ResNet-50 and ViT-large, including internal and external evaluation.

Internal-AIROGS-ResNet50				Internal-AIROGS-ViT		
Method	F1-score	AUROC	AUPR	F1-score	AUROC	AUPR
Random	0.22±0.05	0.82±0.03	0.62±0.02	0.29±0.04	0.85±0.03	0.63±0.03
CorSet	0.31±0.07	0.86±0.02	0.66±0.03	0.32±0.03	0.87±0.01	0.67±0.02
ALFA-Mix	0.39±0.05	0.88±0.01	0.71±0.01	0.44±0.03	0.90±0.02	0.71±0.01
DataDIVA	**0.42±0.04**	**0.90±0.01**	0.71±0.02	**0.46±0.04**	0.90±0.02	**0.72±0.02**
External-REFUGE-ResNet50				External-REFUGE-ViT		
Random	0.08±0.04	0.76±0.06	0.76±0.04	0.47±0.06	0.54±0.40	0.66±0.26
CorSet	0.38±0.09	0.83±0.03	0.80±0.02	0.60±0.22	0.87±0.01	0.86±0.01
ALFA-Mix	0.41±0.25	0.88±0.03	0.84±0.01	0.44±0.20	0.91±0.04	0.88±0.02
DataDIVA	**0.54±0.25**	**0.91±0.02**	**0.87±0.02**	**0.64±0.12**	**0.93±0.01**	**0.91±0.01**
External-ORIGA-ResNet50				External-ORIGA-ViT		
Random	0.15±0.10	0.68±0.02	0.65±0.02	0.21±0.18	0.64±0.03	0.62±0.03
CorSet	0.20±0.03	0.68±0.03	0.66±0.02	0.26±0.17	0.73±0.02	0.68±0.02
ALFA-Mix	0.20±0.03	0.70±0.03	0.66±0.02	0.21±0.07	0.70±0.02	0.65±0.01
DataDIVA	0.19±0.04	**0.73±0.02**	**0.69±0.02**	**0.29±0.15**	**0.75±0.04**	**0.71±0.03**

Table 3. Model performance on referable diabetic retinopathy detection with model ResNet-50 and ViT-large, including internal and external evaluation.

Internal-EyePACS-ResNet50				Internal-EyePACS-ViT		
Method	F1-score	AUROC	AUPR	F1-score	AUROC	AUPR
Random	0.43±0.01	0.74±0.01	0.70±0.01	0.32±0.21	0.71±0.05	0.65±0.06
CorSet	0.40±0.03	**0.76±0.01**	0.71±0.01	0.51±0.02	0.77±0.01	0.72±0.01
ALFA-Mix	0.43±0.03	0.75±0.02	**0.72±0.02**	**0.52±0.03**	0.75±0.01	0.74±0.01
DataDIVA	**0.45±0.01**	0.75±0.01	0.71±0.01	0.50±0.02	0.77±0.01	0.74±0.01
External-APTOS2019-ResNet50				External-APTOS2019-ViT		
Random	0.68±0.01	0.80±0.01	0.74±0.01	0.79±0.02	0.85±0.13	0.84±0.12
CorSet	0.70±0.03	0.85±0.01	0.84±0.01	0.78±0.09	0.90±0.03	0.91±0.03
ALFA-Mix	0.71±0.02	0.86±0.01	0.84±0.01	0.78±0.05	0.91±0.02	0.90±0.02
DataDIVA	0.71±0.03	**0.88±0.02**	**0.88±0.02**	**0.83±0.04**	**0.92±0.02**	**0.92±0.02**
External-IDRiD-ResNet50				External-IDRiD-ViT		
Random	0.78±0.03	0.81±0.02	0.80±0.02	0.83±0.01	0.84±0.05	0.83±0.04
CorSet	0.80±0.03	0.87±0.02	0.84±0.02	0.78±0.08	0.88±0.01	0.87±0.02
ALFA-Mix	0.81±0.01	0.87±0.01	0.84±0.03	0.84±0.03	0.91±0.01	0.88±0.01
DataDIVA	**0.83±0.03**	**0.90±0.01**	**0.87±0.02**	**0.86±0.02**	**0.92±0.01**	**0.90±0.02**

brought by DataDIVA. We visualised the t-SNE maps based on REFUGE features extracted by fine-tuned models in Figure 2a, and observed that the model trained on DataDIVA samples indeed separated different categories better. We also studied the sample categorical balance for each method, where DataDIVA offered a more balanced distribution, as shown in Supplementary Table 1.

Performance comparison in referable DR detection. Table 3 compares the performance of DataDIVA and competitive baselines in referable DR detection. We observed that DataDIVA performed comparably to the baselines in the internal test while achieving the best performance in external evaluations. We also observed well-separated features extracted by the model fine-tuned on DataDIVA samples in Figure 2b. The sample balance distribution is introduced in Supplementary Table 2.

Effects of sampling hyperparameter. We studied the effects of cluster number K (default as 10) on referable DR detection in Supplementary Figure 1. The results showed that model performance is relatively stable while large cluster

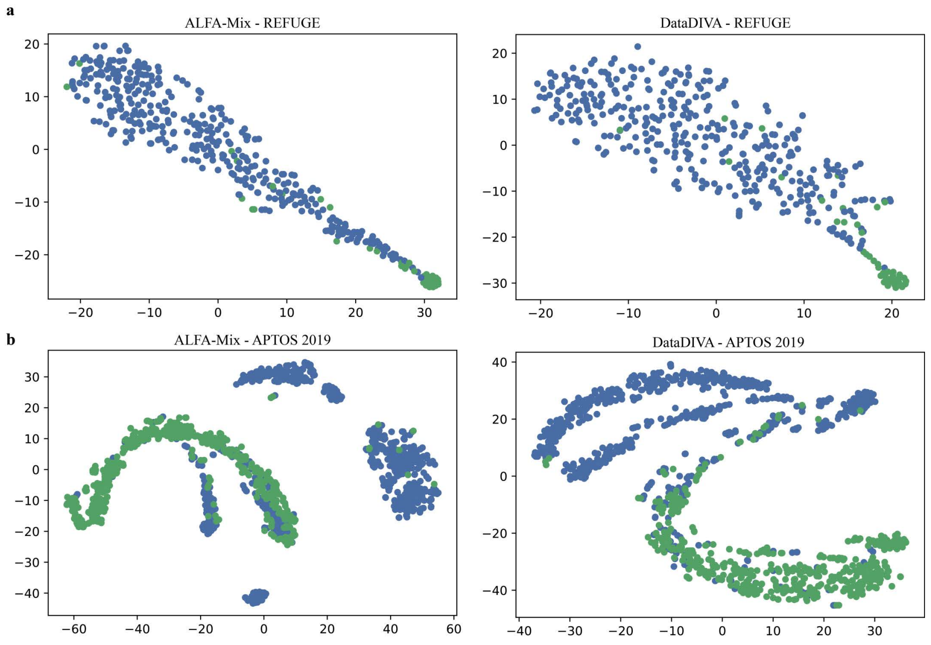

Fig. 2. t-SNE maps of a. REFUGE and b. APTOS 2019 test features extracted by models trained on ALFA-Mix (left) and DataDIVA (right) samples. The different colour indicates the different categories. DataDIVA separates categories relatively better.

numbers ($K = 10, 15$) provided slightly better performance. In Supplementary Figure 2, we showed a better performance using a specialised foundation model in extracting retinal features, compared to with a model trained on ImageNet-21k.

4 Conclusion

We present DataDIVA, an efficient data sampling pipeline for retinal images, aiming to select a diverse and balanced subset for model generalisability. Experimental results show that DataDIVA significantly improves model performance compared to competitive methods in clinically meaningful tasks. We validated the strength of the foundation model in facilitating feature extraction and data sampling. We also discovered that all specifically designed sampling strategies (CoreSet, ALFA-Mix, and DataDIVA) alleviated the unbalanced issue, partially explaining the solution principles. This research can provide insights into several healthcare AI applications, such as large-scale clinical database curation and data selection for annotation.

DataDIVA works on increasing data diversity and balance in logistic and simple ways. A huge room exists to further improve DataDIVA in its two components by combining various sampling strategies (e.g. incorporating ALFA-Mix into balance sampling). Future work will compare DataDIVA with the model trained on the full labelled data, and investigate the combination of DataDIVA with other generalisation techniques, including image augmentation and generation, which are well-recognised strategies to increase data diversity thus improving model generalisation. Additionally, it is worth exploring whether such a sampling pipeline has synergy with other generalisable training frameworks, such as federated learning and active learning. Finally, the proposed selection strategy will be evaluated in other clinical tasks, as well as extra medical fields, including radiology and dermatology. Large-scale clinical datasets with diverse health conditions will be used to evaluate the efficacy of DataDIVA in real-world healthcare scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019)
2. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9368–9377 (2018)
3. Blumberg, S.B., Palombo, M., Khoo, C.S., Tax, C.M., Tanno, R., Alexander, D.C.: Multi-stage prediction networks for data harmonization. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. pp. 411–419. Springer (2019)
4. Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., Kumar, S.: Batch active learning at scale. Advances in Neural Information Processing Systems **34**, 11933–11944 (2021)
5. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3366–3385 (2021)
6. De Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.D., et al.: Airogs: artificial intelligence for robust glaucoma screening challenge. IEEE transactions on medical imaging (2023)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841 (2018)
9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nature medicine **25**(1), 24–29 (2019)

10. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., Celi, L.A.: The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health **2**(9), e489–e492 (2020)
11. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International conference on machine learning. pp. 1183–1192. PMLR (2017)
12. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama **316**(22), 2402–2410 (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
15. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)
16. Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M.G., et al.: The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. Medical Image Analysis **66**, 101714 (2020)
17. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
18. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural networks **113**, 54–71 (2019)
19. Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G.R., Van Den Hengel, A., Shi, J.Q.: Active learning by feature mixing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12237–12246 (2022)
20. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: Idrid: Diabetic retinopathy–segmentation and grading challenge. Medical image analysis **59**, 101561 (2020)
21. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)
22. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
23. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations
24. Settles, B.: Active learning literature survey (2009)
25. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review **5**(1), 3–55 (2001)
26. Shui, C., Zhou, F., Gagné, C., Wang, B.: Deep active learning: Unified and principled method for query and training. In: International Conference on Artificial Intelligence and Statistics. pp. 1308–1318. PMLR (2020)
27. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5972–5981 (2019)

28. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine **25**(1), 44–56 (2019)
29. Yin, C., Qian, B., Cao, S., Li, X., Wei, J., Zheng, Q., Davidson, I.: Deep similarity-based batch mode active learning with exploration-exploitation. In: 2017 IEEE International Conference on Data Mining (ICDM). pp. 575–584. IEEE (2017)
30. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual international conference of the IEEE engineering in medicine and biology. pp. 3065–3068. IEEE (2010)
31. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)
32. Zhou, Y., Wagner, S.K., Chia, M.A., Zhao, A., Xu, M., Struyven, R., Alexander, D.C., Keane, P.A., et al.: Automorph: Automated retinal vascular morphology quantification via a deep learning pipeline. Translational Vision Science & Technology **11**(7), 12–12 (2022)