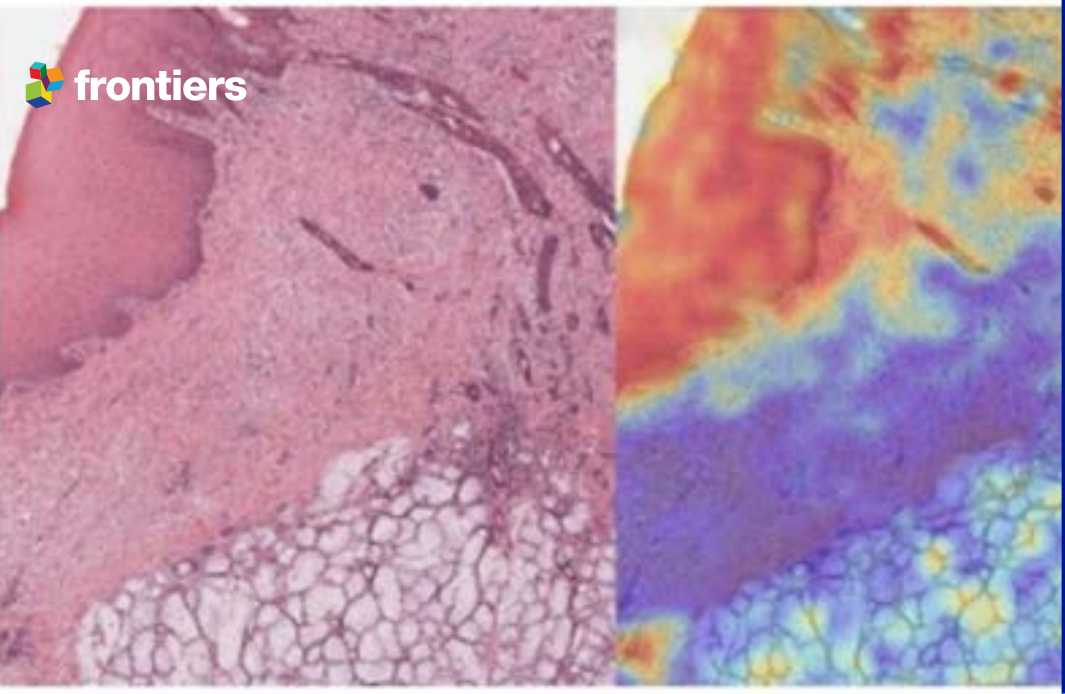
frontiers



Medical Image Understanding and Analysis

Manchester, UK

Medical Image Understanding and Analysis, Manchester, UK

ISBN
9782832512449

DOI
10.3389/978-2-8325-1244-9

Citation
Thomas, C., Kendrick, C., Cootes, T., Reeves, N., Yap, M. H., Zwiggelaar, R. (2024).

Manchester Metropolitan University, Manchester, UK

Table of contents

Welcome to Medical Image Understanding and Analysis

MIUA is a UK-based international conference for the communication of image processing and analysis research and its application to medical imaging and biomedicine. MIUA 2024 is organized by a combined team at Manchester Metropolitan University, University of Manchester and Aberystwyth University.

LIST OF ORGANIZERS

Chairs:
Moi Hoon Yap, Manchester Metropolitan University
Timothy Cootes, University of Manchester
Reyer Zwiggelaar, Aberystwyth University
Neil Reeves, Manchester Metropolitan University

Set 1: Brain Imaging, Medical Images and Computational Models

Analysis on multi-ethnic retinal fundus image datasets of debiasing techniques in deep learning for diabetic macular edema recognition

Author
Megha Hegde – School of Computer Science and Mathematics, Kingston University London
Farzana Rahman – School of Computer Science and Mathematics, Kingston University London
Roshan A. Welikala – School of Computer Science and Mathematics, Kingston University London
Jiri Fajtl – School of Computer Science and Mathematics, Kingston University London
Christopher G. Owen – Population Health Research Institute, St George's, University of London
Alicja R. Rudnicka – Population Health Research Institute, St George's, University of London
Sarah A. Barman – School of Computer Science and Mathematics, Kingston University London

Citation
Hegde, M., Rahman, F., Welikala, R.A., Fajtl, J., Owen, C.G., Rudnicka, A.R., Barman, S.A. Analysis on multi-ethnic retinal fundus image datasets of debiasing techniques in deep learning for diabetic macular edema recognition.

Abstract
Diabetes-related hyperglycaemia can lead to sight-threatening conditions such as diabetic retinopathy (DR) and diabetic macular edema (DME). Manual diagnosis is time-consuming and requires a high level of expertise, hence automation is helpful. Deep learning models have achieved high accuracies but can be biased against groups that are underrepresented within the data. This study extends existing work on debiasing DR detection models to DME, for which there is less available data and less reported algorithm development. The best method (image augmentation) reduced the gap in accuracy between the best and worst-performing classes within the dataset from 18.5% to 4.5%.

Introduction
Diabetes mellitus, known colloquially as "diabetes", is a condition characterised by issues with insulin production or response, leading to hyperglycaemia (elevated blood sugar levels) [1]. When left untreated, diabetes-related hyperglycaemia can damage the blood vessels and nerve tissue in the retina, leading to conditions such as diabetic retinopathy (DR) and diabetic macular edema (DME), which can lead to vision impairment or loss [2]. Manual detection of DR and DME is time-consuming and requires a high level of expertise, making it difficult for people living in remote or rural areas to access diagnosis and treatment [3,4]. This is a significant issue for diabetic individuals who reside in low and middle-income countries, where there is a significant shortage of ophthalmologists [5, 6]. Therefore, automation is a promising solution to this problem.

Machine learning (ML) and deep learning (DL) methods have proven effective in detecting DR and DME from retinal images, with convolutional neural network (CNN) methods achieving accuracies as high as 97.3% and 86.4% respectively. However, these models can become biased against groups that

are underrepresented in the available training data. Ethnicity and iris colour are associated with retinal pigmentation, which impacts on the appearance of retinal images and disease markers within them [7]. Hence, it is crucial that models for DR and DME detection are exposed to a diverse, balanced range of images during training. This is challenging, as capturing large new datasets is expensive and time consuming, while obtaining current patient data raises privacy and ethical obstacles. Despite this, computational methods have been developed to produce unbiased, or "fair" models, with promising results. However, most of the existing research has been performed with respect to DR and looks at debiasing with respect to only "lighter-skin" and "darker-skin" groups, rather than testing across a spectrum of ethnically diverse datasets [8, 9]. This paper enhances the current debiasing methods that were developed for DR, and applies them to DME, for which there is less available data, and less reported algorithm development. Furthermore, it explores debiasing models for a diverse dataset comprising more than two ethnic groups.

Data

Three open-access datasets of retinal fundus images were combined: the Hamilton Eye Institute Macular Edema Dataset (HEI-MED) [17], the Indian Diabetic Retinopathy Image Dataset (IDRiD) [18, 19], and MESSIDOR-2 [20]. All datasets graded DME according to severity. For this study, all DME grades were converted to a binary label of 0 (no DME) or 1 (DME). HEI-MED reported the patient ethnicity for each sample, and the other two datasets reported the data collection location, making it possible to infer the ethnic group(s) mainly represented. Following the approach of Paul *et al.* [9], individual typology angle (ITA) was investigated as a proxy for patient ethnicity, however it did not show any clear relationship with the ethnicity labels provided in the HEI-MED. Instead, a CNN-based method was implemented to classify the ethnicity of the retinal images.

Methods

The study followed a similar methodology to that in Paul *et al.* [9], shown in Figure 1. Experiments were conducted to test model bias due to domain generalisation, by training on samples derived from one ethnic group only (White), and testing on a diverse dataset consisting of White, Indian, and African samples. Then, the following debiasing methods were trialled: data augmentation, adversarial

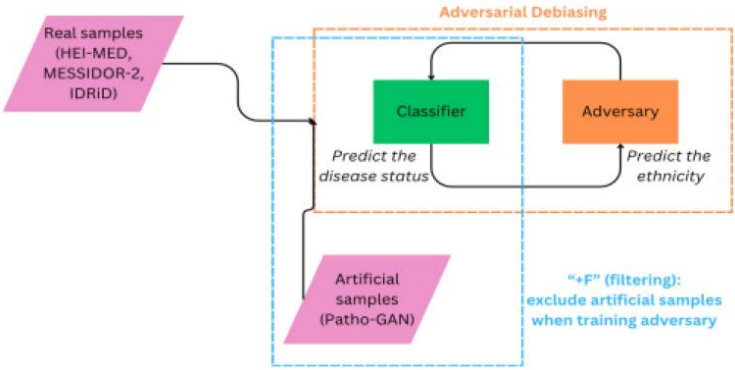


FIGURE 1
Methodology for adversarial debiasing, TARA, and TARA+F, based on Paul *et al.* [9].

debiasing, TARA [9], and TARA+F [9]. This study aimed to extend the original work by testing whether debiasing with respect to the ethnicities represented in the training data could improve performance on other, unseen ethnicities. Additionally, this study used a much smaller dataset, with 200 training samples compared with the 20,000 in the original study. The baseline and adversarial debiasing models were trained on White samples only, while the augmentation and TARA+F models were trained on a combination of White samples and artificial Indian samples. All models were tested on the same dataset to enable comparison. The models were trained to perform a binary classification task on each sample, to categorise it as "DME" or "non-DME". In the data, all DME samples have DR, as DME develops as a complication of DR. The non-DME samples contain both samples with and without DR. ResNet-50, the baseline classifier used in [9], resulted in overfitting on our dataset, even after applying dropout and regularisation techniques. Instead, DenseNet121 [10] was used as the baseline classifier, as it has fewer parameters but has demonstrated high performance on similar tasks [11-13].

Patho-GAN [14] was used to generate artificial samples of Indian individuals with DME to augment the dataset. A version of the model pre-trained on

IDRiD was available, hence this method had a higher speed and lower computational cost compared with alternatives. Adversarial debiasing, TARA, and TARA+F were implemented following the methodology in [9].

Results

All methods were tested on White, Indian, and African samples. The test dataset consisted of 132 images, 44 of each ethnicity (22 with DME, and 22 without). Table 1 shows a comparison of accuracy across the models. The study proved that the novel debiasing methods used in Paul et al. [9] were applicable to a different disease, and when using a smaller dataset. However, as the code for the original study was not available, the implementation will have differed from the original implementation. This may explain why, in this case, the novel debiasing method was outperformed by simple dataset augmentation using Patho-GAN [51]. The results showed that, for data augmentation and TARA+F, debiasing with respect to White and Indian samples improved accuracy on the African samples, of which no real or artificial samples were included in the training or validation datasets. This suggests that these techniques help the model generalize well to new samples. In the case of adversarial debiasing, the acc_{gap} was reduced, but the accuracy on the African samples was not improved. Dataset augmentation and TARA+F improved the overall AUC, but increased the AUC_{gap}, while adversarial debiasing decreased the overall AUC, but reduced the AUC_{gap}. The overall accuracy and AUC of the models appears to have been limited by the low number of training samples. Hence, further work is required to optimise the models and bring the model performance to an acceptable standard for clinical use. Additionally, given that the AUC values prefer different debiasing methods to the accuracy values, it is crucial to examine the true positive and false positive rates.

TABLE 1: Comparison of accuracy (*acc*) and AUC across the baseline model and the best model for each of the debiasing methods. The acc_{gap} *and* AUC_{gap} are the difference in accuracy and AUC, respectively, between the best-performing and worst-performing classes.

Model	Overall		White		Indian		African			
	acc	AUC	*acc*	AUC	*acc*	AUC	*acc*	AUC	acc_{gap}	AUC_{gap}
Baseline	77.3%	0.6995	86.4%	0.8595	77.3%	0.9184	68.2%	0.7211	18.2%	0.1973
Augmentation	72.7%	0.7142	75.0%	0.8678	72.7%	0.9638	70.5%	0.7531	4.5%	0.2107
Adversarial Debiasing	75%	0.6930	77.3%	0.8533	81.8%	0.9132	65.9%	0.7180	15.9%	0.1952
TARA+F	76.5%	0.7165	77.3%	0.8760	81.8%	0.9514	70.5%	0.7500	11.3%	0.2014

This study explores debiasing with respect to ethnicity, however, this is only one axis on which bias may exist; it is necessary to ensure that models are also not biased with respect to other protected characteristics, such as age and sex. Additionally, as explored by Li et al. [15] and Mehta and Waheed [16], there are variations in retinal characteristics even within ethnic groups, so an ethnically diverse dataset still may not adequately capture a set of characteristics that represents biological diversity across the human race. This reinforces the need to produce models which generalise well to unseen samples that are not represented within the training or validation data.

The combination of multiple datasets in this study may have introduced confounding factors related to image characteristics, for example, due to the use of different cameras; however, it is anticipated that the comparative approach reported will still provide valuable insights for future studies. While debiasing methods have been successful in reducing the disparity in results in cases with limited data, the development of a large, diverse, open-access dataset reporting anonymised patient demographic labels would be invaluable in training models that perform well across a wide patient demographic.

Acknowledgement
The authors acknowledge support from the Wellcome Collaborative Award in Science (224390/Z/21/Z) and Women In STEM Hub (WISH) seed funding from Kingston University.

References

[1] Diabetes UK: Diabetes Mellitus, https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes/diabetes-mellitus, last accessed 2023/03/21.

[2] Mookiah, M.R.K. et al.: Application of different imaging modalities for diagnosis of diabetic macular edema: a review. Computers in biology and medicine, 66, pp.295-315 (2015). doi: 10.1016/j.compbiomed.2015.09.012

[3] Scotland, G.S. et al.: Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. British Journal of Ophthalmology, 94(6), pp.712-719 (2010). doi: 10.1136/bjo.2008.151126

[4] Silberman, N., Ahrlich, K., Fergus, R., Subramanian, L. Case for automated detection of diabetic retinopathy. In: AAAI spring symposium series (2010).

[5] Resnikoff, S., Felch, W., Gauthier, T.M., Spivey, B.: The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners. British Journal of Ophthalmology, 96(6), pp.783-787 (2012). doi: 10.1136/bjophthalmol-2011-301378

[6] Sivaprasad, S., Gupta, B., Crosby-Nwaobi, R., Evans, J.: Prevalence of diabetic retinopathy in various ethnic groups: a worldwide perspective. Survey of ophthalmology, 57(4), pp.347-370 (2012). doi: 10.1016/j.survophthal.2012.01.004

[7] Rajesh, A.E. et al.: Ethnicity is not biology: retinal pigment score to evaluate biological variability from ophthalmic imaging using machine learning. medRxiv, pp.2023-06 (2023). doi: 10.1101/2023.06.28.23291873

[8] Gronowski, A., Paul, W., Alajaji, F., Gharesifard, B., Burlina, P.: Renyi fair information bottleneck for image classification. In 2022 17th Canadian Workshop on Information Theory (CWIT), pp. 11-15, IEEE (2022). doi: 10.1109/CWIT55308.2022.9817669

[9] Paul, W., Hadzic, A., Joshi, N., Alajaji, F., Burlina, P.: Tara: training and representation alteration for ai fairness and domain generalization. Neural Computation, 34(3), pp.716-753 (2022). doi: 10.1162/necoa01468

[10] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708 (2017). doi: 10.48550/arXiv.1608.06993

[11] Mishra, S., Hanchate, S., Saquib, Z.: Diabetic retinopathy detection using deep learning. In: 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 515-520 (2020). IEEE. doi: 10.1109/ICSTCEE49637.2020.9277506

[12] Mohanty, C. et al.: Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. Sensors, 23(12), p.5726 (2023). doi: 10.3390/s23125726

[13] Kumar, A., Tewari, A.S., Singh, J.P.: Classification of diabetic macular edema severity using deep learning technique. Research on Biomedical Engineering, 38(3), pp.977-987 (2022). doi: 10.1007/s42600-022-00233-z

[14] Niu, Y., Gu, L., Zhao, Y., Lu, F.: Explainable diabetic retinopathy detection and retinal image generation. IEEE journal of biomedical and health informatics, 26(1), pp.44-55 (2021). doi: 10.1109/JBHI.2021.3110593

[15] Li, X. et al.: Racial differences in retinal vessel geometric characteristics: a multiethnic study in healthy Asians. Investigative ophthalmology & visual science, 54(5), pp.3650-3656 (2013). doi: 10.1167/iovs.12-11126

[16] Mehta, N. and Waheed, N.K.: Diversity in optical coherence tomography normative databases: moving beyond race. International Journal of Retina and Vitreous, 6(1), pp.1-4 (2020). doi: 10.1186/s40942-020-0208-5

[17] Giancardo, L.; Meriaudeau, F.; Karnowski, T. P.; Li, Y.; Garg, S.; Tobin, Jr, K. W.; Chaum, E., 'Exudate-based diabetic macular edema detection in fundus images using publicly available datasets.', Medical Image Analysis 16(1), pp. 216-226 (2012). doi: 10.1016/j.media.2011.07.004

[18] Abràmoff, M.D. et al.: Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA ophthalmology, 131(3), pp.351-357 (2013).

[19] Porwal, P. et al.: Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data, 3(3), p.25 (2018). doi: 10.3390/data3030025

[20] Decencière, E. et al.: Feedback on a publicly distributed image database: the Messidor database. Image Analysis & Stereology, 33(3), pp.231-234 (2014).

Automatic segmentation of pediatric brain tumours using diffusion-weighted MRI: Towards an early, in-vivo classification pipeline

Author

Daniel Griffiths-King – Aston Institute of Health and Neurodevelopment, College of Health and Life Sciences, Aston University, Birmingham, United Kingdom
Timothy Mulvany – Aston Institute of Health and Neurodevelopment, College of Health and Life Sciences, Aston University, Birmingham, United Kingdom
Andrew Peet – Department of Oncology, Birmingham Childrens Hospital, Birmingham, United Kingdom; Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, United Kingdom
John Apps – Department of Oncology, Birmingham Childrens Hospital, Birmingham, United Kingdom; Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, United Kingdom
Jan Novak – Aston Institute of Health and Neurodevelopment, College of Health and Life Sciences, Aston University, Birmingham, United Kingdom

Citation

Griffiths-King, D., Mulvany, T., Peet, A., Apps, J., Novak, J. Automatic segmentation of pediatric brain tumours using diffusion-weighted MRI: Towards an early, in-vivo classification pipeline.

Abstract

Diffusion-weighted MRI (DWI) can offer vital quantitative biomarkers to understand pediatric brain tumours. However, extraction of these markers

requires delineation of pathological tissue, which is time-consuming, requires expert-resource and can still be highly variable. This study develops an automated segmentation approach, leveraging transfer-learning and multi-modal ensembling to tackle the issues specific to this DWI-only approach. Using a 3D CNN (namely Deepmedic) to perform automatic segmentations, we demonstrate the benefit of transfer-learning and ensembling for this task. However, we find limited accuracy of these segmentations, with limited reliability of the radiomic features extracted from these regions of interest (compare to the ground truth tumour mask). The current study highlights the potential of this approach, indicating the biases and issues which need to be addressed before it can be implemented in future clinical decision support tools.

Introduction

Diffusion-weighted MRI (DWI) is a standard imaging modality for neuroradiological assessment of pediatric brain tumours, with derived DWI-biomarkers providing insights into tumour microenvironment and enabling diagnostic classification of tumour-type [1]. Regions of interest (ROIs), drawn around visible pathology, are needed to extract DWI metrics from the pathological tissue, which are subsequently incorporated into classifiers. Individuals with the requisite neuro-radiological expertise to draw these ROIs have limited capacity to conduct this additional, time-consuming workload. However, even expert-drawn ROIs are still subject to inter-/intra-rater variability [2], therefore, automated approaches are appealing.

Automated approaches for brain tumour segmentation exist [3], but are inappropriate in this use case. Typically, they do not segment the DWI image, instead using other imaging modalities (T1w, T2w, FLAIR) – however, this then requires co-registration across modalities which is challenging in the presence of gross pathology/abnormality and geometric distortions common to DWI acquisitions [4]. Given this, automated methods for segmentation based solely on DWI, from which we wish to extract quantitative biomarkers, is warranted.

The current study employed a 3D, multiscale convolutional neural network (CNN) to develop a model for the automated segmentation of pediatric brain tumours from DWI, with the specific goal of implementing these automatic segmentations/ROIs into a diagnostic classifier, using DWI imaging biomarkers to predict tumour diagnosis in-vivo.

Method

Dataset

This study utilised data for n=107 pediatric brain tumour patients. DWI was acquired during standard care at the Birmingham Childrens Hospital, prior to surgical intervention or adjunct therapy. Consent was given for the upload of patient data to the UK Children's Cancer and Leukaemia Group Functional Imaging (CCLG-FIG) Database. Multiple tumour types were included in the current cohort.

Region of Interest (ROI) Drawing

ROIs were manually drawn on B0 volumes, using 3D Slicer [5] (ver.5.2.1). Additional imaging modalities were used to ensure inclusion of tumour only, with oedema and large cystic regions excluded. ROIs were iteratively improved through drawing by TM and review/evaluation by JN & DGK.

Image Preprocessing

B0 and B1000 volumes were extracted from DWI, and parametric maps of apparent diffusion coefficient (ADC) calculated from observed diffusion signal. B0/B1000 volumes were corrected for bias field and Gibbs artefact. Subsequently, all volumes were resampled to 1mm isotropic, intensity-normalised and had non-brain tissue removed.

Segmentation Tool

Automated segmentation was performed using Deepmedic [6] (ver.0.8.4), a multi-scale 3D Convolutional Neural Network, 11-layers deep, with a normal- and low-resolution pathway, with the final 3 layers fully-connected.

Training and Evaluation
All experiments were evaluated using a 4-fold cross-validation (CV) approach (each fold consisting of Training (n=80/81) and Testing data (n=27/26)). All hyperparameters and model choices were decided upon performance on training data with final performance evaluated on testing data across all folds. Concordance between segmentations and ground truth ROIs was assessed using Dice scores.

Exp. 1 Transfer-learning and Fine-Tuning
The rarity of pediatric brain tumours, and relative limited availability of DWI data poses challenges for sample sizes. We use transfer-learning to exploit knowledge from a similar classification problem with an additional, larger dataset. Specifically, the model was pre-trained on existing, open-access MRI data of pediatric brain tumours (n=99) –from the 2023 BraTS-PEDs Challenge [7]. BraTS-PEDs does not include DWI, and so T2w MRI was used as the closest to expected DWI contrast. Learning rate and number of fully connected layers frozen for transfer-learning were optimised using systematic search. Once non-frozen layers were retrained on the CCLG-FIG data, all layers were unfrozen and fine-tuned using a smaller learning rate.

A baseline model was also trained for comparison – without pretraining on BraTS-PEDs, using only CCLG-FIG data for training (4-fold CV).

Exp. 2 Multimodal Ensembling
Despite being a single acquisition sequence, DWI offers 3 reconstructable, volumes; B0, B1000 and the ADC map. Three models, each using a different image as input channels, were trained and the output segmentations ensembled using a union-mask approach (i.e. a voxel is included if it is included in ANY of the three masks).

Exp. 3 Concordance of Radiomic Markers
Radiomics can be used to extract quantitative features from DWI for future analysis and processing. Radiomic features from the ADC map were extracted from within the predicted segmentation (using PyRadiomics [8]) and compared to ground truth ROI using Intraclass correlation coefficients (ICC).

Results

Exp. 1

For the B0 volume, transfer-learning alone did not significantly improve performance on test data (Mean Dice Score ($\overline{Dice}$ = 0.25, Standard Deviation (SD) = 0.24) compared to the baseline model ($\overline{Dice}$ = 0.38, SD = 0.28). However, a slight improvement to average Dice score was achieved using transfer-learning followed by fine-tuning ($\overline{Dice}$ = 0.39, SD = 0.27).

Exp. 2

Fine-tuned models using B1000 performed similarly ($\overline{Dice}$ = 0.34, SD = 0.29), however ADC maps gave greater improvement ($\overline{Dice}$ = 0.47, SD = 0.30). The greatest performance on test data was achieved through ensembling ($\overline{Dice}$ = 0.49, SD = 0.26). Examples are seen in Figure 1.

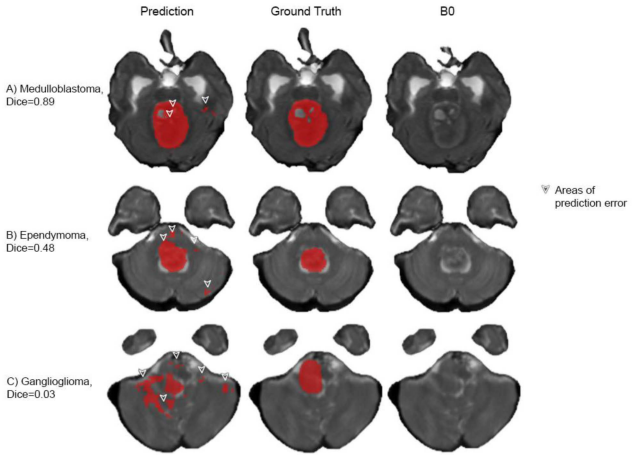


FIGURE 1
Examples of tumour masks for individual cases with high (A) to low (C) segmentation performance (using ensembled method), based upon Dice score. Arrows indicate areas of prediction error.

Exp. 3
19 radiomic features were extracted from the ADC, for both the ground truth ROI and the segmentations from the ensemble (B0, B1000,& ADC) model. Across features, agreement was poor with mean ICC = 0.37, SD = 0.23. Only 6 features performed with an ICC of 'moderate' or above (Entropy, Mean, RMS, Median, 10th Perc., 90th Perc. (in order)).

Exploratory Analysis
Considering DICE at the patient-level, no bias due to age at diagnosis was observed, but a positive correlation existed between tumour grade (where available) and Dice (Rho=0.23, p=0.038). Unsurprisingly, qualitative investigation suggested those cases with lowest Dice scores were typically rarer tumour types, with fewer cases in the dataset overall.

Discussion
Overall, based on Dice alone, the best model still performed poorly, for instance in comparison to results on the 2023 BraTS-PEDs Challenge (winning Dice score for tumour core was 0.80 using typical MR modalities). Further research should therefore focus on utilising prevailing state-of-the-art architectures such as U-Nets, alongside larger cohorts for both pre-training and fine-tuning segmentation models.

However, in developing automated segmentation approach for pediatric brain tumours from DWI, we demonstrated that transfer-learning and fine-tuning – even when pretrained with a different imaging modality (T2w vs DWI), improves segmentation performance, as does the multi-modal ensembling using additional volumes derivable from the DWI sequence alone.

In-vivo diagnosis and classification of pediatric brain tumours using MRI will help guide early disease management, potentially supporting decision making through data-driven, clinical decision support tools. Through developing ways to leverage additional imaging modalities, such as DWI, this work will further contribute to this goal.

References

[1] Novak, J., N. Zarinabad, H. Rose, T. Arvanitis, L. MacPherson, B. Pinkey, A. Oates, P. Hales, R. Grundy, D. Auer, D.R. Gutierrez, T. Jaspan, S. Avula, L. Abernethy, R. Kaur, D. Hargrave, D. Mitra, S. Bailey, N. Davies, C. Clark, and A. Peet, *Classification of paediatric brain tumours by diffusion weighted imaging and machine learning.* Sci Rep, 2021. **11**(1): p. 2987.

[2] Zhang, L., R. Tanno, M.-C. Xu, C. Jin, J. Jacob, O. Cicarrelli, F. Barkhof, and D. Alexander, *Disentangling human error from ground truth in segmentation of medical images.* J Advances in Neural Information Processing Systems, 2020. **33**: p. 15750-15762.

[3] Bhalodiya, J.M., S.N. Lim Choi Keung, and T.N. Arvanitis, *Magnetic resonance image-based brain tumour segmentation methods: A systematic review.* Digit Health, 2022. **8**: p. 20552076221074122.

[4] Crum, W.R., M. Modo, A.C. Vernon, G.J. Barker, and S.C. Williams, *Registration of challenging pre-clinical brain images.* Journal of Neuroscience Methods, 2013. **216**(1): p. 62-77.

[5] Pieper, S., M. Halle, and R. Kikinis. *3D Slicer*. in *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)*. 2004. IEEE.

[6] Kamnitsas, K., E. Ferrante, S. Parisot, C. Ledig, A.V. Nori, A. Criminisi, D. Rueckert, and B. Glocker. *DeepMedic for brain tumor segmentation*. in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Second International Workshop, BrainLes 2016, with the Challenges on BRATS, ISLES and mTOP 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 2*. 2016. Springer.

[7] Kazerooni, A.F., N. Khalili, X. Liu, D. Haldar, Z. Jiang, S.M. Anwar, J. Albrecht, M. Adewole, U. Anazodo, H. Anderson, S. Bagheri, U. Baid, T. Bergquist, A.J. Borja, E. Calabrese, V. Chung, G.M. Conte, F. Dako, J. Eddy, I. Ezhov, A. Familiar, K. Farahani, S. Haldar, J.E. Iglesias, A. Janas, E. Johansen, B.V. Jones, F. Kofler, D. LaBella, H.A. Lai, K. Van Leemput, H.B. Li, N. Maleki, A.S. McAllister, Z. Meier, B. Menze, A.W. Moawad, K.K. Nandolia, J. Pavaine, M. Piraud, T. Poussaint, S.P. Prabhu, Z. Reitman, A. Rodriguez, J.D. Rudie, M. Sanchez-Montano, I.S. Shaikh, L.M. Shah, N. Sheth, R.T. Shinohara, W. Tu, K. Viswanathan, C. Wang, J.B. Ware, B. Wiestler, W. Wiggins, A. Zapaishchykova, M. Aboian, M. Bornhorst, P. de Blank, M. Deutsch, M. Fouladi, L. Hoffman, B. Kann, M. Lazow, L. Mikael, A. Nabavizadeh, R. Packer, A. Resnick, B. Rood, A. Vossough, S. Bakas, and M.G. Linguraru, *The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs).* ArXiv, 2024.

[8] van Griethuysen, J.J.M., A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillion-Robin, S. Pieper, and H. Aerts, *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.

Computational fluid dynamics modelling of blood flow performance for a stented patient specified peripheral arteries

Author

J Feng – Department of Engineering, Manchester Metropolitan University, Manchester, M1 5GD, UK

D Wang – Department of Engineering, Manchester Metropolitan University, Manchester, M1 5GD, UK

J Kendall – Department of Engineering, Manchester Metropolitan University, Manchester, M1 5GD, UK

F. Serracino-Inglott – Manchester Royal Infirmary, Manchester University NHS Foundation Trust, Manchester, M13, 9Wl, UK

Citation

Feng, J., Wang, D., Kendall, J., Serracino-Inglott, F. Computational fluid dynamics modelling of blood flow performance for a stented patient specified peripheral arteries.

Background

Plaque accumulation is the primary cause of peripheral artery disease (PAD), which strikes millions of lives worldwide (Martin et al., 2024). Stenting is widely accepted as a treatment for PAD. However, hemodynamic changes induced by the implementation of peripheral arterial stent can lead to in-stent restenosis

(ISR) and stent thrombosis (ST) (Hirschhorn et al., 2020). It is reported that ISR result from neointimal hyperplasia is linked with low and oscillating wall shear stress (WSS). Whereas high WSS induced platelet activation contributes to the formation of ST (Wu et al., 2024; Ng et al., 2017). Therefore, it is important to investigate hemodynamic factors, such as wall shear stress (WSS), time-averaged wall shear stress (TAWSS) and oscillating shear index (OSI), when studying stented peripheral arteries.

Aims and Objectives

In this study, we aim to rigorously investigate the hemodynamic quantities of a patient specific stented femoral artery model based on computational fluid dynamics (CFD) analysis. This patient-specific model is generated based on computed tomography (CT) image. To this end, a finite element (FE) analysis is first performed to deploy the stent in a patient-specific femoral artery. Subsequently, a CFD analysis is carried out based on the stented arterial model obtained from the FE analysis to investigate the hemodynamic factors of this model. Pulsatile boundary conditions are applied in the CFD simulations.

Methods

CFD Model Preparation (see (Wang et al., 2021) *for details):* Femoral artery computed tomography (CT) images of a 73-old male patient are used to construct a patient-specific 3D model. The data contains 3918 slices of CT images with 0.625 mm slice thickness, each slices contain 512×512 pixels (see for details). The CT images were identified and selected by our clinical collaborator. Ethical approval was obtained from HRA and Health and Care Research Wales (HCRW). A 25-mm-long left femoral artery model is reconstructed based on these CT images. The internal diameter of the arterial wall ranges from 8.3 mm to 9 mm. The femoral artery wall is assumed to be nonlinear hyperplastic material, and the strain energy function Mooney-Rivlin constitutive model. The stented peripheral arterial geometry was prepared through FE simulation. A six-ring stent is created using SolidWorks (Dassault Systems, SolidWorks Corp., MA). The stent is 6 mm long and the strut thickness is 1 mm. A FEM simulation is established

for fully expanded stent configuration with a diameter of 9 mm. The stent is considered as stainless steel.

CFD Simulation

A structure simulation is carried out first to obtain the stented artery geometry, which can then be used in the CFD simulation to study the hemodynamic behaviour in the stented femoral artery. Time-dependent velocity and pressure wave-forms are applied as the inlet and outlet boundary conditions, respectively. No-slip boundary condition is applied on all the lumen surface. ANSYS MESH is used to discretize the models with tetrahedron elements.

Results

To investigate the hemodynamic characteristics of the patient-specific stented model, we demonstrate CFD results including blood flow pattern, time-averaged WSS and oscillating shear index in this section. For example, Fig. 1 demonstrates the flow velocity streamlines and contours at two-time instants during cardiac cycle at the time of 0.16s and 0.48s.

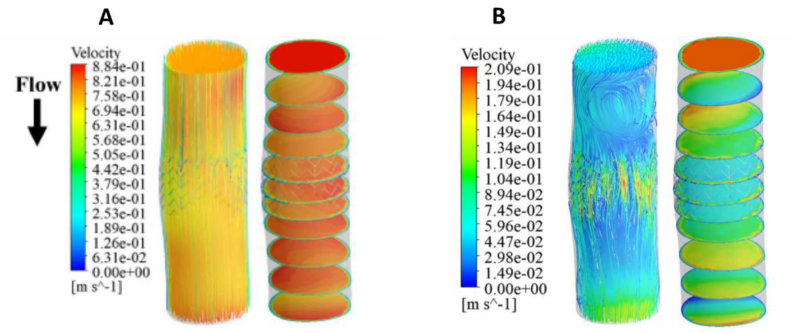


FIGURE 1
Velocity streamlines and contours at different time instants.

Conclusions

Hemodynamic features including blood velocity, time-averaged wall shear stress and oscillating shear index are carefully considered in this study. Notably, regions of low and oscillatory wall shear stress are observed around the stent strut surface, where in-stent restenosis is likely to happen. Whereas high wall shear stress which could lead to stent thrombosis is mainly found within the stent cell area.

References

Hirschhorn, M., Tchantchaleishvili, V., Stevens, R., Rossano, J. and Throckmorton, A. (2020) *Med Eng Phys*, 78, Apr, 20200217, pp. 1-13.

Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Barone Gibbs, B., et al. (2024). *Circulation*, 149(8), Feb 20, 20240124, pp. e347-e913.

Ng, J., Bourantas, C. V., Torii, R., Ang, H. Y., Tenekecioglu, E., Serruys, P. W. and Foin, N. (2017) *Arterioscler Thromb Vasc Biol*, 37(12), Dec, 20171109, pp. 2231-2242.

Wang, D., Serracino-Inglott, F. and Feng, J. (2021) ' *Biomech Model Mechanobiol*, 20(1), Feb, 20200911, pp. 255-265.

Deep learning with 3D convolutional neural networks for prediction of germline BRCA gene mutation in high-risk breast cancer patients

Author

Yongwon Cho – Department of Computer Science and Engineering, Soonchunhyang University, South Korea, Republic of Korea

Sung Eun Song – Department of Radiology, Korea University Anam Hospital, 73, Goryeodae-ro, Seoungbukgu, Seoul 02841, Republic of Korea

Kyu Ran Cho – Department of Radiology, Korea University Anam Hospital, 73, Goryeodae-ro, Seoungbukgu, Seoul 02841, Republic of Korea

Citation

Cho, Y., Song, S.E., Cho, K.R. Deep learning with 3D convolutional neural networks for prediction of germline BRCA gene mutation in high-risk breast cancer patients.

Abstract

We determine the feasibility of using a deep learning (DL) approach for predicting the gBRCA mutation by using the whole three-dimensional (3D) MR images of breast cancer and clinical data in high-risk cancer patients. A total of 324 breast cancer patients who had high-risk factors (such as 1) triple negative breast cancer; 2) primary bilateral breast cancer; 3) young breast cancer diagnosed at age less than or equal to 40 years; 4) at least one first- and/or second-degree relative with BRCA-related cancer and underwent gBRCA tests were retrospectively collected. A 3D convolutional

neural networks (CNNs)-based transformer architecture was trained on 80% of the data set and tested on the remaining 20%. Clinical data obtaining from biopsy (such as hormonal receptor, human epidermal growth factor receptor 2 and Ki-67 proliferation index) were also used. The models' performances with or without clinical data were analyzed in terms of accuracy, sensitivity, specificity, and areas under the receiver operating characteristic curve (AUCs). Among 324 patients, 100 (30.9%) had gBRCA 1/2 pathogenic mutation. For prediction of gBRCA mutation in the test set. a 3D CNNs-based transformer architecture was evaluated using subtraction and T2WI MRIs with Logistic Regression (LGR) classifier. The best performance (only deep learning) among our models achieved an AUC of 0.82 (95% confidence interval [CI]: 0.64, 0.93), 80% sensitivity, 83% specificity, 80% PPV, 83% NPV, and 81% accuracy. In conclusion, DL models can effectively predict the gBRCA mutation without clinical data. Artificial intelligence may provide an early detection of gBRCA mutation for breast cancer patients who had high-risk factors.

Introduction

The purpose of this study was to determine the feasibility of using a deep learning (DL) approach for predicting the gBRCA mutation by using the three-dimensional (3D) MR images of breast cancer and clinical data in high-risk cancer patients.

Materials and Methods

Datasets

A total of 324 breast cancer patients who had high-risk factors and underwent gBRCA tests were retrospectively collected. A data set of 3D MR images of breast cancer from 324 subjects were manually segmented on contrast-enhanced T1-weighted subtraction images by an experienced breast radiologist. The ITK-SNAP software version 3.6.0 was used for segmentation (www.itksnap.org) of breast cancer. After segmentation, the reviewer crosschecked and revised the segmentation results in consensus. Clinical data obtaining from biopsy were also used. They were randomly split into training, tuning, and test (n = 253, 28, 43, respectively) after breast cancer annotation, with summary shown in Table 1.

TABLE 1: Number of MRI and clinical datasets in the training and test sets

Heading level	Number of images in the training set (with tuning set)	Number of images in the test set
BRCA no	181 (18)	23
BRCA yes	100 (10)	20
Total	**281**	**43**

Development to Classify BRCA

We proposed an integrated architecture, named 3D CNNs[1]-based transformer in Figure 1. The lesion volumes extracted from MRIwere first entered to train this model for high-level feature representations through layers including convolutional layers and max-pooling. The clinical data and deep features were concatenated using machine learning like Logistic Regression (LGR) classifier connected to last layer to classify non-BRCA and BRCA. The validation of the classifier models was performed with 20 repeated 10-fold stratified cross-validations using test dataset. In addition, we conducted an ablation study on the prediction of BRCA with a test dataset to compare with radiomics[1,2,3] and clinical feature model.

Results

In figure 2, for prediction of gBRCA mutation in the test set, a 3D CNNs-based transformer architecture achieved an AUC of 0.83 and 0.6, respectively. The DL model using subtraction and T2WI MRI achieved 80%

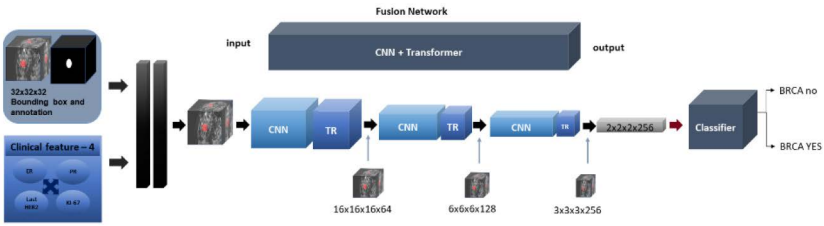


FIGURE 1
Architecture of 3D CNNs-based transformer to classify the patients including BRCA.

Image	Model	Diagnostic performance						
		AUC	Acc.	Sens.	Spec.	PPV	NPV	*P*- Value[a]
Subtraction	**Deep Learning**	**0.83 (0.71-0.95)**	**81%**	**80%**	**83%**	**80%**	**83%**	Reference
	Clinical feature	0.65 (0.48-0.82)	66%	56%	74%	63%	68%	0.14[a]
	Radiomic feature	0.79 (0.67-0.93)	79%	70%	87%	82%	77%	0.81[a]
T2WI	**Deep Learning**	0.65 (0.48-0.78)	60%	45%	74%	60%	61%	Reference
	Clinical feature	0.65 (0.47-0.82)	66%	56%	74%	63%	68%	0.53[a]
	Radiomic feature	0.53 (0.35-0.69)	53%	55%	52%	50%	57%	0.61[a]

a. P-value was acquired from comparison with the reference standard using the Delong method.

FIGURE 2
Diagnostic performances of the models in test set.

sensitivity, 83% specificity, 80% PPV, 83% NPV, and 81% accuracy in LGR classifier, and 45% sensitivity, 74% specificity, 60% PPV, 61% NPV, and 60% accuracy in LGR classifier, respectively.

Conclusions

We developed DL models to classify BRCA using MRI different from other studies[4,5]. In the future work, we will compare with another deep learning using our dataset. Using 3D MR images, DL models can effectively predict the gBRCA mutation without clinical data. Artificial intelligence may provide an early detection of gBRCA mutation for breast cancer patients who had high-risk factors.

Acknowledgments

We would like to thank the Advanced Medical Imaging Institute in the department of radiology, the Korea University Anam Hospital and This work supported by the Soonchunhyang University Research Fund.

References

1. Dening Lu, Qian Xie, Linlin Xu, Jonathan Li, 3DCTN: 3D Convolution-Transformer Network for Point Cloud Classification, arXiv:2203.00828v. (2022)

2. Van Griethuysen JJM et.al, Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res (2017)

3. Pyradiomics community (2021) Welcome to pyradiomics documentation! https://pyradiomics.readthedocs.io/en/latest/index.html. Accessed 20 July 2021

4. Xiaoxiao Wang, et. al., Prediction of BRCA Gene Mutation in Breast Cancer Based on Deep Learning and Histopathology Images, frontiers in Genetics (2021)

5. Raphaël Bourgade, Deep Learning for Detecting BRCA Mutations in High-Grade Ovarian Cancer Based on an Innovative Tumor Segmentation Method From Whole Slide Images, Modern Pathology (2023)

Exploring the potential of MRI variables for predicting conversion to mild cognitive impairment

Author

Martina Billichová – Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia

Davide Bruno – School of Psychology, Liverpool John Moores University, Liverpool, United Kingdom

Fariba Sharifian – School of Computer Science and Mathematics, Liverpool John Moores University, Liverpool, United Kingdom

Silvester Czanner – Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia; School of Computer and Engineering Sciences, University of Chester, Chester, United Kingdom

Gabriela Czanner – Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia

Citation

Billichová, M., Bruno, D., Sharifian, F., Czanner, S., Czanner, G. Exploring the potential of MRI variables for predicting conversion to mild cognitive impairment.

Abstract

The prevalence of dementia, mainly Alzheimer's disease, emphasises the need for early detection, especially in people with normal cognitive function. To predict the progression of Mild Cognitive Impairment and Alzheimer's disease, statistical modelling and machine learning methods are used to analyse complex datasets. Our study demonstrates the efficacy of a statistical

Cox regression model for clinical application and shows strong predictive potential. While utilising imaging data has demonstrated significant predictive capability, including both imaging and clinical data notably enhances the predictive accuracy of the Cox model.

Introduction

The field of healthcare is advancing, leading to increased global life expectancy. For instance, from 2014 to 2016, the UK witnessed a modest increase in life expectancy for both men and women, reflecting this global trend (Yang, et al. 2019). With longer life comes a higher risk of age-related brain disorders, impacting daily life (Li, et al. 2021). Dementia, notably Alzheimer's disease (AD), is a leading cause of death. The most common early symptoms of AD pathology are poorer episodic memory and orientation, followed by a widespread decline in cognition and it likely leads to loss of independence (Arvanitakis, Shah and Bennett 2019). As no treatments to date can fully restore cognitive health, pharmacological and non-pharmacological interventions are applied to attempt to slow cognitive decline and improve the overall quality of life (Chen, Jiao and Zhang 2022). Once healthcare systems adopt treatments, there will be a need to enhance early AD diagnostics. Statistical and machine learning models can boost accurate and timely prediction of the time of conversion to MCI, enabling physicians to promptly apply appropriate treatment strategies for at-risk patients.

Related Work

Detecting early cognitive impairment is crucial but challenging. Zhao et al. underscore the importance of identifying specific biomarkers for early-stage AD to enable early intervention. They (Zhao, et al. 2019) assessed the risk of AD progression within two years, using the AD-Resemblance Atrophy Index (AD-RAI) to categorise patients by brain atrophy severity. Statistical analyses (ANCOVA, ANOVA), and logistic regression showed AD-RAI's potential in estimating MCI or AD progression risk. Similarly, Lin et al. (Lin, et al. 2018) predicted MCI occurrence over four years, differentiating individuals transitioning to MCI from those with normal cognitive abilities.

Using support vector machine (SVM), logistic regression, and random forest (RF), they achieved over 75% ROC AUC. Cognitive decline in AD or MCI can be detectable through MRI. AI is increasingly applied to MRI analysis, facilitating non-invasive identification of brain structural changes. Frizzell et al. reviewed 97 studies from 2009 to 2021, focusing on AI models using MRI for AD diagnosis (Frizzell, et al. 2022). They compared algorithms like SVM, RF, neural networks, and logistic regression, mainly for diagnosis classification or MCI to AD conversion. Reported algorithmic performance ranged from 75% to 95%, with some reaching up to 98% accuracy in AD and MCI classification. Convolutional neural networks were found to be the most efficient. Deep neural networks are used in survival analysis due to their powerful image-processing capabilities (Ocasio and Duong 2021). However, statistical techniques such as the Cox regression model are widely used due to their effectiveness. Studies have shown that the Cox model performs equally well as machine learning approaches (Billichová, et al. 2024). The Cox model lacks image processing capability but effectively handles preprocessed tabular data. Notably, it can outperform machine learning models even with smaller sample sizes.

To our knowledge, there is currently limited literature on predicting the development of MCI. Existing studies focus mostly on Alzheimer's disease using demographic and clinical data. MRI data is gaining popularity in predicting neurodegenerative diseases. This paper focuses on predicting MCI conversion using imaging data in tabular form obtained from preprocessed MRI images to investigate their potential. We apply an advanced statistical Cox regression method, which has demonstrated its strength and advantages in this novel medical field.

Materials and Methods

Data

In our research, we use data from the National Alzheimer's Coordinating Center. We carefully chose a subset of measures based on brain areas commonly associated with AD-related atrophy, such as areas in the medial temporal lobe and the whole brain, as shown by MRI (Liu, et al. 2021).

Selected MRI variables were: total intracranial volume, total brain grey and white matter volume, segmented left and right hippocampus volume, segmented total hippocampi volume, left and right entorhinal grey matter volume, left and right parahippocampal grey matter volume, left and right entorhinal mean cortical thickness, and left and right parahippocampal mean cortical thickness. Demographic and clinical data were chosen by a backward feature selection approach. Due to a large number of features, we narrowed them down by referring to current publications (Lin, et al. 2018, Billichová, et al. 2024), ultimately selecting the variables: gender, age, education, memory decline, judgment, travelling difficulty, motor function and cognitive status.

Methods

Cox proportional hazards model. Survival analysis commonly employs the Cox proportional hazards model, renowned for its popularity in statistical prediction (Kurt Omurlu, Ture and Tokatli 2009). As a semiparametric model, it requires fewer assumptions compared to parametric models, making it widely employed (Abadi, et al. 2014). The core assumption of this model lies in the proportionality of the hazard function.

Results

To tackle the aim of our study, we implemented the Cox proportional hazards (CoxPH) model and assessed three scenarios for predicting conversion to MCI using: *1) imaging data, 2) demographic and clinical data,* and *3) imaging, demographic and clinical data*.

The dataset, comprising 1366 individuals, was randomly split into a training set and a test set with an 80:20 ratio. This resulted in 1092 samples (80%) for training and 274 samples (20%) for testing. Additionally, we evaluated each scenario using 10-fold cross-validation.

TABLE 1: Comparison of the accuracy of Cox regression models on a test dataset across different data types

	C-index	Brier score	Mean C-index (10-fold CV)
Imaging data	0.756	0.167	0.710 ± 0.045
Demographic and clinical data	0.861	0.148	0.866 ± 0.033
Imaging + demographic and clinical data	0.861	0.158	0.854 ± 0.036

The CoxPH model trained on only imaging data achieved a C-index of 75.6% and a Brier score of 0.167 on the testing set. The mean C-index from 10-fold cross-validation was 71.0% with a standard deviation of 4.5%. The model performance was significantly improved by incorporating both image and clinical data, with mean C-index = 85.4% (Table 1). Demographic and clinical data alone provided comparable performance to the combined model, achieving a mean C-index of 86.6% with a standard deviation of 3.3% and a Brier score of 0.148 on the testing subset.

Conclusion

Detecting MCI at an early stage, even in people with normal cognition, is crucial. Time-to-event predictive modelling, helps clinicians identify individuals who are at risk of developing Alzheimer's disease in its early stages. We have successfully implemented a predictive statistical Cox regression model. This method has shown strong predictive potential (mean C-index, 86.6%) and has the potential to screen progression to MCI. Imaging data have some predictive power (mean C-index, 71.0%) however, they do not add any further predictive power to demographic and clinical data. Previous research using CNNs reported strong but not necessarily superior performance to the Cox regression model used in this study. However, CNNs typically require larger datasets and more computational resources. Furthermore, the interpretability of CNNs remains challenging, whereas the Cox model offers more straightforward insights into the influence of specific risk factors. In future, interactions of the imaging data with demographic and clinical data can be investigated, as heterogeneity of risk factors

between e.g. males and females has been previously reported. Furthermore, feature selection methods and other machine learning approaches (CNNs, DeepSurv, RF, etc.) can be investigated and compared to enhance predictive performance.

Acknowledgments
We would further like to acknowledge the NACC team for providing detailed documentation and datasets.

References

Abadi, A., et al. "Cox models survival analysis based on breast cancer treatments." *Iranian Journal of Cancer Prevention*, 2014: 124-129.

Arvanitakis, Zoe, Raj C. Shah, and David A. Bennett. "Diagnosis and Management of Dementia: Review." *JAMA*, 2019: 1589-1599.

Billichová, Martina, Lauren Joyce Coan, Silvester Czanner, Monika Kováčová, Fariba Sharifian, and Gabriela Czanner. "Comparing the performance of statistical, machine learning, and deep learning algorithms to predict time-to-event: A simulation study for conversion to mild cognitive impairment." *PLOS ONE*, 2024: 1-20.

Frizzell, Tory O., et al. "Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: A systematic review." *Ageing Research Reviews*, 2022.

Chen, L., J. Jiao, and Y. Zhang. "Therapeutic approaches for improving cognitive function in the aging brain." *Frontiers in Neuroscience*, 2022.

Kurt Omurlu, Imran, Mevlut Ture, and Füsun Tokatli. "The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer." *Expert Systems with Applications*, 2009.

Li, Zhe, et al. "Aging and age-related diseases: from mechanisms to therapeutic strategies." *Biogerontology*, 2021: 165-187.

Lin, Ming, Pinghua Gong, Tao Yang, Jieping Ye, Roger L. Albin, and Hiroko H. Dodge. "Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment." *Alzheimer Disease and Associated Disorders*, 2018.

Liu, W., et al. "MRI-based Alzheimer's disease-resemblance atrophy index in the detection of preclinical and prodromal Alzheimer's disease." *Aging*, 2021.

Ocasio, E., and T. Q. Duong. "Deep learning prediction of mild cognitive impairment conversion to Alzheimer's disease at 3 years after diagnosis using longitudinal and whole-brain 3D MRI." *PeerJ. Computer science*, 2021.

Yang, Su, Jose Miguel Sanchez Bornot, Kongfatt Wong-Lin, and Girijesh Prasad. "M/EEG-Based Bio-Markers to Predict the MCI and Alzheimer's Disease: A Review from the ML Perspective." *IEEE Transactions on Biomedical Engineering* (IEEE), 2019.

Zhao, Lei, et al. "Risk estimation before progression to mild cognitive impairment and Alzheimer's disease: An AD resemblance atrophy index." *Aging*, 2019.

Enhanced segmentation via a shared encoder with interpretable classifier for breast tumor analysis

Author
Youngmin Kim – Kyungpook National University, IPA Lab, Korea
Sungjoon Park – Kyungpook National University, IPA Lab, Korea
Hyejeong Kim – Kyungpook National University Medical Center, Korea
Wonhwa Kim – Kyungpook National University Medical Center, Korea
Jaeil Kim – Kyungpook National University, IPA Lab, Korea

Citation
Kim, Y., Park, S., Kim, H., Kim, W., Kim, J. Enhanced segmentation via a shared encoder with interpretable classifier for breast tumor analysis.

Abstract
Accurate segmentation of tumors in breast MRI is always a critical issue in medical imaging. Breast MRI images are used for surgical planning and post-operative care; therefore, they need to be analyzed accurately. The application of deep learning techniques for the automated segmentation of breast tumors in MRI has been introduced, leading to significant advancements. Nevertheless, breast tumors vary in size and shape, accurate segmentation remains a challenge.

In addition, diagnosing the entire MRI image slices can be inefficient, as only a small percentage of slices in a patient's entire MRI image slices contain tumors. Therefore, our study aims to classify the presence of tumor in MRI slices and accurately segment tumors of different sizes and types.

Introduction

Our proposed model is based on a multi-task learning framework that consists of sharing an encoder with a classification network and segmentation network. Multi-task learning has been previously applied to the analysis of medical data, leveraging the simultaneous training of multiple related tasks to improve model performance [1, 2]. The difference between our proposed model and the existing methods is that our classifier has the capability to not only classify tumors but also, as it learns knowledge, predict theapproximate location of the tumors. This is the basis for the classifier to consider it interpretable. The learning process of the classification network influences the segmentation network as both networks share an encoder, thereby enhancing segmentation performance. Additionally, it is possible to consolidate the slices predicted as containing tumors by the classifier to ascertain the precise location of each patient's tumor. Our proposed model structure is shown at Figure 1.

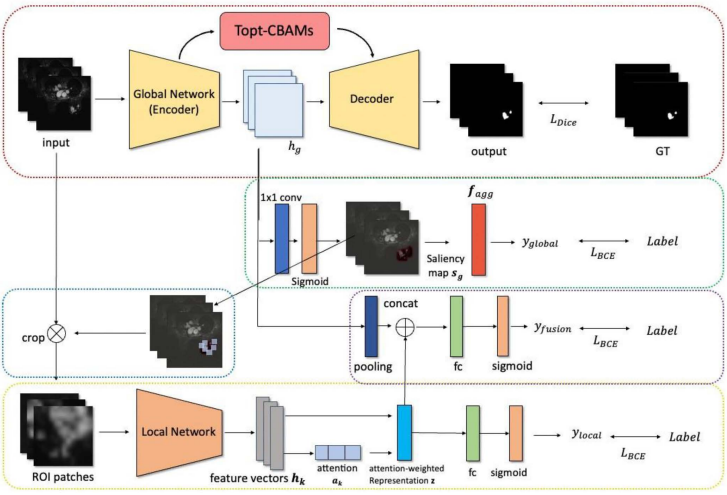


FIGURE 1
An overview of multi-task learning process. Classification and segmentation share an encoder, as calledGlobal Network.

Methods

We utilize the Globally-Aware Multiple Instance Classifier(GMIC) [3] as a classification network. This model is a multiple instance classifier that uses a two-step process. First, a low-capacity Global Network identifies key regions in an image. Then, a high-capacity Local Network analyzes these regions in detail.

Finally, a fusion module combines global and local information for the final prediction. A decoder network is added to the global network of this classification model to perform segmentation simultaneously. The global network and decoder network utilize the structure of U-Net. This approach allows the classification network and segmentation network to share an encoder (the Global Network) so that they can improve each performance by influencing the learning process through mutual exchange of information. We also utilized an attention module the skip-connection of each encoder and decoder stage. This module is a variant the Convolutional Block Attention Module(CBAM) called Topt-CBAM, aimed at enhancing segmentation performance.

Dataset and Experiments

The study used a private DCE-MRI dataset containing images from 1,010 breast cancer patients, which were acquired at the KNUMC institution. Tumor masks are annotated by radiologists. The data utilized in this study comprises the subtraction image, obtained by subtracting the post-contrast image from the pre-contrast image. By subtracting the pre- and post-contrast images, only the changes due to the contrast enhancement are highlighted, making it easier to distinctly visualize abnormalities such as tumors orinflammatory diseases. Mohamed Eid et al. [4] also shows that subtraction images are helpful. The dataset is split into training, validation, and test sets in the ratio of 0.7, 0.1, and 0.2. Considering that the number of slices containing tumors

is significantly lower than those without tumors, we opted for the under-sampling method to balance the training and validation datasets. In our study, due to the subtraction process, pixel values in these images may be negative. We address this by replacing any negative pixel values with zero prior to normalization. Subsequently, images undergo Min-Max normalization to scale the pixel values to the range [0,1].

To compare our proposed model, we conducted experiments with existing segmentation models. Additionally, we compare a version of our proposed model without an attention module, a version using the traditional CBAM module instead of Topt-CBAM. Segmentation evaluation was conducted by calculating the DSC, Sensitivity, and Average Hausdorff distance(Avg HD). Evaluation metrics was calculated individually for each patient, and the final score was determined by averaging the scores across all patients. Additionally, the classification task was evaluated using the AUC metric. We also performed subgroup analyses based on tumor type and size.

Results

Given that our model, the classification results exhibit a significant level of performance, attaining an AUC of 97.1%. The AUC was computed on a per-slice basis. This high classification performance results are a reliable indicator that our model can accurately classify slices containing tumors. Segmentation results for test data patients are shown in Table 1, demonstrating the superior performance of our model. Our proposed model showed the lowest performance without the attention module. Performance improved with the use of CBAM, and the highest performance was achieved with the Topt-CBAM module.Our model's performance is further improved when classification is performed on slices with tumors classified first and then segmented only on those slices. Fig. 2 illustrates a qualitative example.

TABLE 1: Comparison of experimental results for breast tumor segmentation. We estimate the 95% confidence interval(CI) using 10,000 samples

	DSC ↑(%)	Sensitivity ↑ (%)	Avg HD ↓ (mm)
UNet [5]	70.50(CI:67.76,73.13)	70.55(CI:67.38,73.65)	13.00(CI:10.42,15.83)
UNet++ [6]	70.09(CI:66.88,73.12)	72.09(CI:68.55,75.48)	15.25(CI:11.81,19.05)
SegNet [7]	66.89(CI:63.70,70.02)	66.42(CI:62.65,70.03)	19.69(CI:15.96,23.68)
SwinUNETR [8]	72.57(CI:69.62,75.30)	77.42(CI:74.43,80.24)	10.88(CI:7.33,15.44)
Ours (without attention module)	72.53(CI:69.66,75.26)	76.64(CI:73.73,79.37)	10.36(CI:7.86,13.10)
Ours (with CBAM [2])	73.03(CI:70.14,75.87)	77.01(CI:73.86,80.07)	10.01(CI:7.08,13.43)
Ours (with Topt-CBAM)	73.37(CI:70.70,75.88)	80.46(CI:78.06,82.74)	8.49(CI:6.68,10.48)
Ours (after classification, with Topt-CBAM)	**77.29(CI:75.00,79.44)**	**82.20(CI:79.86,84.37)**	**6.93(CI:5.21,8.83)**

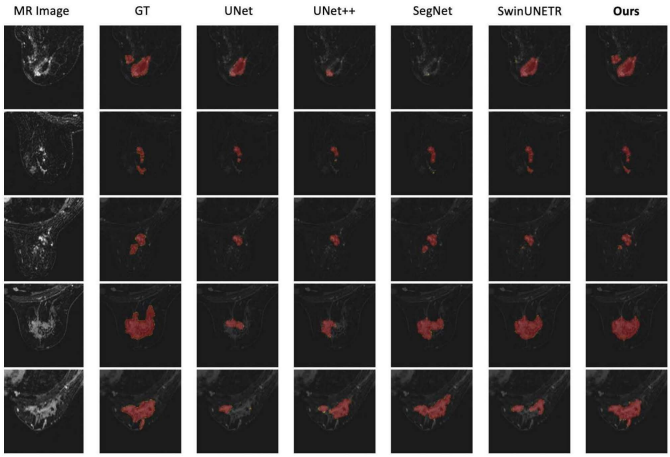


FIGURE 2
Qualitative examples of model. It shows the capability to detect various types and sizes of tumors that other models could not identify.

TABLE 2: Tumor type comparison. Our proposed model performs well, demonstrating a sensitivity of 82.96% for Non-mass tumors, which is particularly challenging due to tumor blurring

	model	DSC ↑(%)	Sensitivity ↑ (%)	Avg HD↓(mm)
Mass	SwinUNETR [8]	75.23 (CI: 71.04,79.04)	75.36 (CI: 70.51,79.83)	14.24 (CI: 8.43,22.88)
	Ours	**80.52 (CI: 77.46,83.35)**	**80.16 (CI:76.08,83.85)**	**9.62 (CI:6.55,13.09)**
Mass + Non-mass	SwinUNETR	79.16 (CI:75.16,82.79)	83.85 (CI:79.41,87.66)	4.86 (CI:1.90,8.82)
	Ours	**81.86 (CI:78.11,84.91)**	**85.78 (CI: 81.17,89.37)**	**3.90 (CI:1.16,8.49)**
Non-mass	SwinUNETR	68.94 (CI:61.95,75.45)	77.72 (CI: 70.50,84.47)	**3.15 (CI: 1.74,4.94)**
	Ours	**70.27 (CI: 63.60,76.08)**	**82.96 (CI: 76.99,88.47)**	3.97 (CI: 1.73,7.08)

TABLE 3: Tumor size comparison. Our model outperforms the SwinUNETR model across all categories, demonstrating superior capability in accurately capturing small tumors

	model	DSC↑ (%)	Sensitivity↑ (%)	Avg HD↓(mm)
Small (size≤ 18mm)	SwinUNE-TR [8]	70.87 (CI: 64.89,76.29)	73.61 CI: 66.73,79.87)	19.77 (CI: 10.32,33.98)
	Ours	**77.40 (CI: 72.67,81.65)**	**77.94 (CI:71.78,83.58)**	**13.59 (CI:8.20,19.58)**
Medium (18mm<-size≤32mm)	SwinUNETR	75.29 (CI:70.65,79.54)	76.46 (CI:71.06,81.41)	6.89 (CI:4.39,9.88)
	Ours	**78.76 (CI:74.78,82.46)**	**81.19 (CI: 76.71,85.21)**	**5.17 (CI:3.39,7.16)**
Large (size>32mm)	SwinUNETR	77.94 (CI:73.20,81.84)	83.73 (CI: 78.99,87.86)	2.92 (CI: 1.46,4.84)
	Ours	**79.68 (CI: 75.60,82.90)**	**87.23 (CI: 84.05,90.13)**	**2.91 (CI: 1.38,5.09)**

Subgroup analysis involved the comparison of two distinct subgroups. We compared performance by tumor type and tumor size using our proposed model and the second best performing SwinUNETR model.

In the tumor type category as shown in the Table 2, our model outperformed SwinUNETR in all categories except Average HD in the Non-mass category. At the tumor size category as shown in the Table 3, our model outperformed SwinUNETR in all categories. Our model demonstrated the ability to segment small tumors, with a Sensitivity of 77.94% in the Small tumor category. This subgroup analysis demonstrates that our proposed model effectively segments even the tumors that are challenging to locate.

Conclusion

In this paper, we propose a multi-task learning approach for tumor classification and segmentation of breast MRI images. This method can help radiologists analyze tumors efficiently because it can achieve accurate segmentation by classifying only slices that contain tumors. We found that an interpretable classifier and segmentation network can better capture tumor features by sharing an encoder. Furthermore, we leveraged a modified attention module called Topt-CBAM to improve segmentation performance. Our model has been trained on a variety of tumors using data from over 1,000 patients, thereby equipping it with the ability to segment between various tumor types.

Acknowledgments

This work was supported by the Technology Innovation Program(or Industrial Strategic Technology Development Program)(20011875, Developed AI diagnostic technology for imaging diagnostic medical devices) funded By the Ministry of Trade Industry Energy(MOTIE, Korea). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2019R1G1A1098655 and 2022R1A2C2009415).

References

[1] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Shan E Ahmed Raza, Fayyaz Minhas, David Snead, and Nasir Rajpoot. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. 2022.

[2] Zishang Kong, Min He, Qianjiang Luo, Xiansong Huang, Pengxu Wei, Yalu Cheng, Luyang Chen, Yongsheng Liang, Yanchang Lu, Xi Li, and Jie Chen. Multi-task classification and segmentation for explicable capsule endoscopy diagnostics. Frontiers in Molecular Biosciences, 8, 2021.

[3] Nan Wu. et al. Yiqiu Shen. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical Image Analysis, 68:101908, 2021.

[4] Mohamed Eid and Ahmed Abougabal. Subtraction images: A really helpful tool in non-vascular mri. The Egyptian Journal of Radiology and Nuclear Medicine, 45(3):909–919, 2014.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597, 2015.

[6] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. CoRR, abs/1807.10165, 2018.

[7] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. CoRR, abs/1511.00561, 2015.

[8] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, 2022.

[9] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. CoRR, abs/1807.06521, 2018.

Generating brain MRI with subject-specific and generalised learning using 3D GAN-based models

Author
Hari Kala Kandel, Carl Barton – Birkbeck, University of London, United Kingdon

Citation
Kandel, H.K., Barton, C. Generating brain MRI with subject-specific and generalised learning using 3D GAN-based models.

Abstract
MRIs are a common imaging modality used in the diagnosis and management of conditions related to the brain. One common type of MRI is structural MRIs, these record the structure of the brain and have high spatial resolution. Another common type is functional MRIs (fMRI) which are used to capture neural activities of the brain by examining the variation of oxygen levels in the blood flow. fMRI is non-invasive and has a high temporal resolution but low spatial resolution. A recent paper on enhancing the spatial resolution of fMRIs to near structural MRI quality was recently published by Ota et al[3]. This paper also introduced the idea of subject-specific learning to this problem. In this approach, the model is trained on data from one patient only to eliminate any anatomical variation that might be caused by training on other subject's data. This is opposed to the more well-known approach of using multiple patients to train the model, referred to as generalised training in

this article. It was suggested by Ota et al[3] that subject-specific training is important for fMRIs as changes in functional activation can be very small and artefacts caused by training on multiple patients may mean information is lost or incorrectly analysed, but this hypothesis wasn't tested. A further limitation of the work was that the model presented was in 2D. Our research is focused on developing 3D architectures for this problem and analysing the impact of using subject-specific vs generalised training. To this end, we have designed 3D architectures for this problem based on three GAN-based [1] models- SRGAN, DCGAN, and CycleGAN- and tested them using subject-specific and generalised training. The resulting models from DCGAN and CycleGAN-based architectures can successfully generate high-quality structured MRI images from resting-state fMRI (rs-fMRI). The novelty of our work is to extend 2D architectures to 3D and test subject-specific and generalised learning with 3D medical images.

The research is conducted on the Human Connectome Project dataset[2]. Structural MRI is considered as high resolution MRI and fed to the model as ground truth whereas rs-fMRI is used as input to the generator. rs-fMRI of each subject is acquired in four different sessions, each session has two runs. We have used a fairly standard (Min-Max) loss function and gradient penalty with a penalty coefficient of 5. In subject-specific learning, we split the data so that data from different sessions is used to train, validate, and test the models whereas, in generalised learning, a model is trained with data from multiple subjects' images and other unseen subjects' images are used to validate and test the model. Through this, we wish to test the hypothesis that subject-specific training is more appropriate for sensitive medical data.

Experimental Comparison of Subject-Specific and Generalised Training

In this section, we compare the performance of subject-specific training against generalised training to determine the suitability of subject-specific and generalised training for sensitive medical applications. To test this, we have trained models using the three architectures using both subject-specific and generalised approaches. For subject-specific training, we trained the

model using four runs from the first two imaging sessions', we validated and tested the model with two runs from the third and fourth session's images respectively which are completely unseen data for the model. For generalised training, we trained the model using 70 subjects with all eight runs' data from all four sessions of each subject. The model is evaluated using all eight runs of four sessions' data from 15 different (unseen) subjects each for validation and testing. We use statistical metrics and box plot methods to compare our models.

Statistical Analysis

The quantitative evaluation metrics for subject-specific and generalised training are shown in Table 1 and Table 2 respectively. From this we can see that images generated via subject-specific training have generally higher SSIM and PSNR than the images generated from generalised training. Images generated from subject-specific training also have lower MSE and MAE than the images generated from generalised training. This suggests that subject-specific model may be much easier to train and can more easily produce higher quality images than in the case of generalised training.

TABLE 1: Average of SSIM, PSNR, MSE and MAE of SRGAN, DCGAN and CycleGAN- based models on subject-specific learning

Subject-specific Learning				
Model	**SSIM**	**PSNR**	**MSE**	**MAE**
SRGAN	0.4516	21.4841	0.1687	0.3316
DCGAN	**0.9867**	**41.2408**	**0.0001**	**0.0052**
CycleGAN	0.9056	31.3349	0.00073	0.0166

TABLE 2: Average of SSIM, PSNR, MSE and MAE of DCGAN and CycleGAN- based model on generalised learning

Generalised Learning				
Model	**SSIM**	**PSNR**	**MSE**	**MAE**
SRGAN	0.4558	unavailable	unavailable	unavailable
DCGAN	0.7202	24.7775	0.0037	0.0334
CycleGAN	**0.7451**	**25.9598**	**0.0035**	**0.0327**

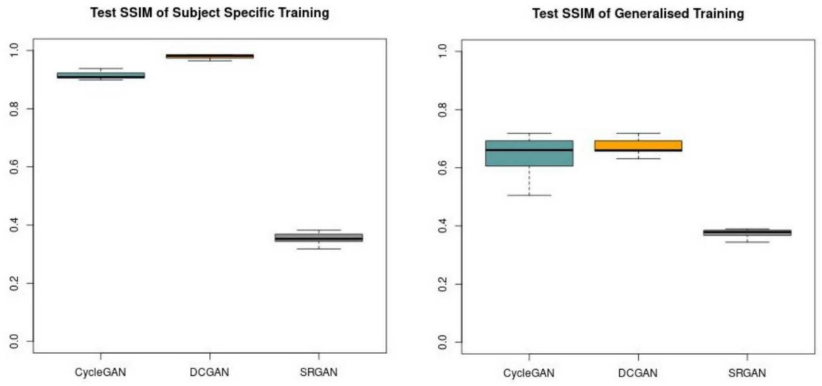


FIGURE 1
Box plot of Test SSIM of DCGAN, CycleGAN and SRGAN- based models on both subject-specific and generalised learning.

Looking at the statistics, presented in Tables 1 & 2 subject-specific models are easier to train and more accurate than generalised learning. This can be seen by the significantly higher SSIM and PSNR values and the significantly lower MAE and MSE reported for subject-specific models. This is further confirmed by the plots in Fig. 1 which demonstrates that in addition to achieving the best average values for the evaluation metrics, the best performing subject-specific models also showed lower variation of SSIM when compared with the best performing generalised models.

Conclusion

These results support the idea that subject-specific training is more appropriate for sensitive medical applications. If small variations are important (such as in MRI/fMRI analysis) then generalised models may not be appropriate as they are much higher variation in SSIM, suggesting that they may introduce small differences not present in the original data. Generalised learning may be more appropriate for applications trying to identify broad trends in the data from many patients, as small variations are less likely to impact the final results.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

[2] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681–4690, 2017.

[3] J. Ota, K. Umehara, J. Kershaw, R. Kishimoto, Y. Hirano, Y. Tachibana, H. Ohba, and T. Obata. Super-resolution generative adversarial networks with static t2* wi-based subject-specific learning to improve spatial difference sensitivity in fmri activation. Scientific Reports, 12(1):10319, 2022.

Implicit neural networks for breast ultrasound image segmentation

Author
Michal Byra – Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

Citation
Byra, M. Implicit neural networks for breast ultrasound image segmentation.

Abstract
Breast cancer is the most common cancer in women, and ultrasound (US) imaging is important for breast mass assessment. Accurate automatic breast mass segmentation facilitates mass characterization. Traditional deep learning methods, such as convolutional networks and transformers, have achieved high performance in breast mass segmentation. Recently, implicit neural representations, which use continuous, nonlinear, coordinate-based approximations through multi-layer perceptrons, have shown promise in various fields, including medical image segmentation. In this work, we present an implicit network for breast mass segmentation in US. We train a coordinate-based implicit network to jointly output the US image pixel values and the segmentation pixel scores. The network is conditioned using latent codes, effectively associating the regression and segmentation tasks with the mass type (benign/malignant) and BI-RADS category. Additionally, a trainable image-specific code is used. During inference, given a US image, we fix the weights of the network and use the backpropagation algorithm to determine the latent codes, facilitating the image regression task. This process, due to

the learned associations, also provides the segmentation mask. Our results confirm the feasibility of using implicit networks for breast mass segmentation and other tasks leveraging learned associations between latent codes and image/mask appearances.

Introduction

Breast cancer is the most common cancer in women, and ultrasound (US) imaging is widely used for breast mass assessment. Accurate automatic breast mass segmentation facilitates mass characterization, as malignant lesions often have more variable shapes compared to benign lesions. Over recent years, various deep learning methods, primarily based on feed-forward convolutional networks or transformers, have been proposed for this task, achieving excellent performance with high agreement with manual annotations by medical experts.

Recently, implicit neural representations (INRs) have gained attention in computer vision and computational physics. INRs provide a continuous, nonlinear, coordinate-based approximation of target quantities using a multi-layer perceptron. In computer graphics, INRs are used to efficiently handle high-resolution data and are well-suited for tasks like differentiable rendering and scene reconstruction due to their ability to interpolate and generalize from sparse data. Singh et al. demonstrated that a single polynomial implicit network could represent large datasets like ImageNet, enabling tasks such as guided image generation or style-mixing. [1].

Implicit networks have emerged as a novel approach for medical image segmentation, offering several advantages over convolutional networks or transformers. Stolt-Anso et al. used implicit networks for cardiac segmentation in MRI. Authors trained a multi-task coordinate-based implicit network to jointly output MRI image pixel intensity values and segmentation mask scores [2]. This network was conditioned with an image-dependent latent code, efficiently relating the tasks of joint image regression and segmentation mask computation.

Methods

In this work, we extend the approach by Stolt-Anso et al. and develop an implicit network for breast mass segmentation in US. To our knowledge, this is the first application of implicit networks for this purpose. Our framework is presented in Fig. 1. Compared to the original approach, our method includes several innovations. In addition to a trainable latent code sampled from a normal distribution, we condition the network using latent codes related to BI-RADS category and breast mass type (malignant/benign), which allows the network to associate US image regression and mass segmentation tasks

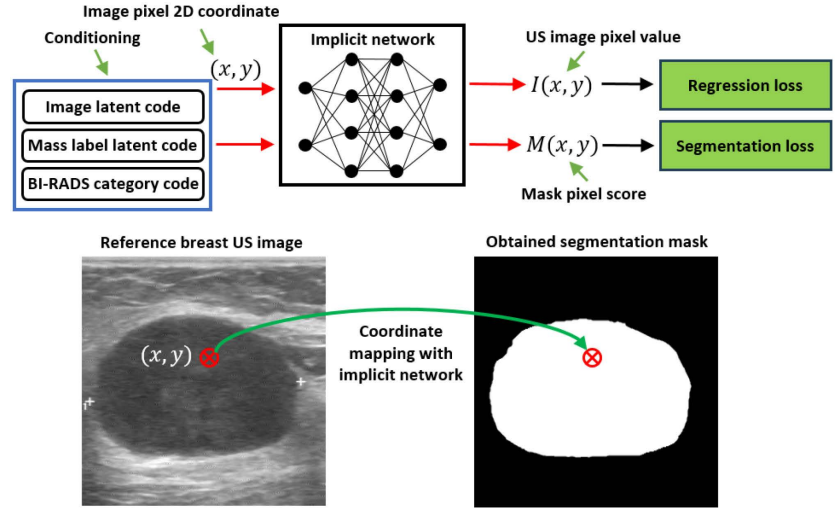


FIGURE 1

We train a coordinate-based implicit network to jointly output the US image pixel values and the segmentation pixel scores. The network is conditioned using latent codes, effectively associating the regression and segmentation tasks with the mass type (benign/malignant) and BI-RADS category. Additionally, a trainable image-specific code is used. During inference, given a US image, we fix the weights of the network and use the backpropagation algorithm to determine the latent codes, facilitating the image regression task. This process, due to the learned associations, also provides the segmentation mask.

with clinical descriptors of the pathology. This conditioning mechanism opens new possibilities. For example, the network may learn to associate the complex shape of a breast mass with malignancy. During inference, we freeze the network weights and use the backpropagation algorithm to regress the test US image and determine the latent code along with the segmentation mask. Furthermore, the learned associations between US images, segmentation masks, and clinical descriptors can be leveraged for several tasks. For instance, we can determine the latent codes corresponding to the BI-RADS category and classification label, effectively transforming the implicit network into a classification model. Additionally, given a segmentation mask and a latent code, our method can address tasks related to image inpainting or interpolation.

Experiments and Discussion

We trained our network using the BUS-BRA dataset, which includes around 1800 breast US images with BI-RADS categories and malignant/benign labels [3]. Our approach was also evaluated using the UDIAT dataset [4]. The segmentation branch was trained using a Dice score-based loss function, while mean squared error was used for the image regression task. Results confirmed the feasibility of using implicit networks for breast mass segmentation and other tasks leveraging learned associations between latent codes and image/mask appearances. However, the proposed method achieved lover performance compared to deep learning methods utilizing transfer learning with models pre-trained on large scale datasets.

Although still in their infancy compared to convolutional models, implicit networks offer a promising approach to breast mass image analysis. Our study is an important preliminary step exploring the usefulness of implicit networks in US image analysis.

Acknowledgement

This study was supported by the National Science Center of Poland (2019/35/B/ST7/03792).

References

[1] Singh, Rajhans, Ankita Shukla, and Pavan Turaga. "Polynomial implicit neural representations for large diverse datasets." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Stolt-Anso, Nil, et al. "Nisf: Neural implicit segmentation functions." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.

[3] Gómez-Flores, Wilfrido, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. "BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems." Medical Physics 51.4 (2024): 3110-3123.

[4] Yap, Moi Hoon, et al. "Automated breast ultrasound lesions detection using convolutional neural networks." IEEE journal of biomedical and health informatics 22.4 (2017): 1218-1226.

Lateral ventricle shape modeling using peripheral area projection for longitudinal analysis

Author
Wonjung Park, Suhyun Ahn, Jinah Park – Korea Advanced Institute of Science and Technology, Republic of Korea

Citation
Park, W., Ahn, S., Park, J. Lateral ventricle shape modeling using peripheral area projection for longitudinal analysis.

Abstract
The deformation of the lateral ventricle (LV) shape is widely studied to identify specific morphometric changes associated with diseases. Since LV enlargement is considered a relative change due to brain atrophy, local longitudinal LV deformation can indicate deformation in adjacent brain areas. However, conventional methods for LV shape analysis focus on modeling the solely segmented LV mask. In this work, we propose a novel deep learning-based approach using peripheral area projection, which is the first attempt to analyze LV while considering surrounding areas. Our approach deforms the follow-up LV mesh to match the baseline shape, while optimizing the corresponding points between baseline and follow-up LVs are located on the surfaces with the same adjacent brain area. Applying this approach, we quantitatively evaluate the deformation of the left LV in normal (n=10) and demented subjects (n=10). Noticeable differences are observed on local LV surfaces adjacent to thalamus, caudate, hippocampus, amygdala and right LV.

Introduction

Enlargement of the lateral ventricles (LV) is observed alongside the loss of brain cells during normal aging and is more pronounced in diseases such as dementia. Thus, LV shape deformation is widely studied to find specific changes associated with brain atrophy. The primary method for shape analysis compares baseline and follow-up LV obtained from longitudinal brain images. Longitudinal changes are derived by deforming the original LV to match the target shape and comparing the original and deformed one. For example, deformable mesh-based methods (Styner et al. (2006)) deform the source LV mesh to the target shape, while diffeomorphic mapping-based methods (Qiu and Miller (2008); Djamanakova et al. (2013)) warp the source brain image to align with the target image.

Despite the successful alignment of the LV shapes, previous approaches cannot ensure that the corresponding points on the baseline and follow-up LV surfaces accurately represent local deformations. To be more specific, these methods do not guarantee that points along brain parts such as the amygdala and hippocampus in the baseline LV are moved to the same part in the follow-up LV.

In our work, taking into account that LV shape deformation is accompanied by surrounding brain atrophy, we utilize peripheral area information to find corresponding points more accurately. Our deep learning-based method induces the vertices of the source mesh to move to the corresponding points located in the same surrounding brain area.

To demonstrate the quantitative local deformation of the LV while considering peripheral areas, we applied our approach to longitudinal brain MRIs of normal (n=10) and demented (n=10) subjects. By projecting peripheral brain structures onto the LV, we analyzed local deformations relative to adjacent brain regions. Noticeable differences in local surface deformations adjacent to the thalamus, caudate, hippocampus, amygdala, and right LV were observed.

Method

As illustrated in Fig.1, our method for longitudinal analysis of LVs consists of two stages. In the first stage, the LV and its peripheral areas are segmented from registered longitudinal brain MRIs. Using the LV mask, we constitute point cloud-based baseline and mesh-based follow-up LVs. Then, using surrounding area masks, we classify the points of the baseline point cloud and the vertices of the follow-up mesh into the nearest peripheral areas.

In the second stage, the follow-up mesh is iteratively deformed to match the baseline shape using a vertex deformation module based on Pointnet (Qi et al. (2017)) architecture with the input of the vertex position. The objective function to minimize is the sum of the distance loss $\boldsymbol{L_{dist}}$ between the deformed mesh and the baseline point cloud, and the regularization loss $\boldsymbol{L_{reg}}$ for the deformed mesh. The distance loss is:

$$L_{dist} = \lambda_{cf} L_{cf} + \lambda_{pm} L_{pm} + \sum_{i=1}^{m} \lambda_{pm_i} L_{pm_i}$$

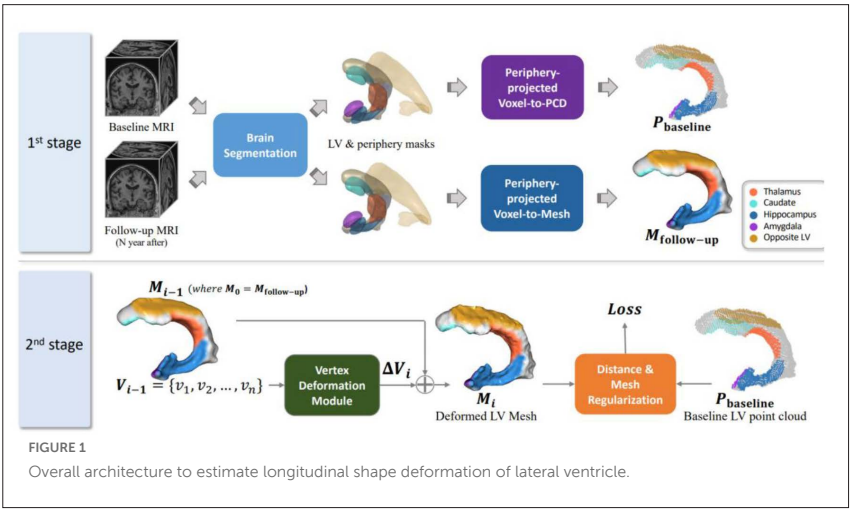


FIGURE 1
Overall architecture to estimate longitudinal shape deformation of lateral ventricle.

L_{cf} is chamfer distance between the selected points on the deformed mesh M_i and the point cloud $P_{baseline}$. L_{pm} is bidirectional distance between $P_{baseline}$ and the faces of M_i. L_{pm_i} represents the bidirectional distance between the points and the part of the mesh faces classified into the i-th area of m surrounding area classes. This loss encourages the vertices of the mesh to deform towards the positions of the same surrounding area. The regularization loss is:

$$L_{reg} = \lambda_{vert}L_{vert} + \lambda_{edge}L_{edge} + \lambda_{normal}L_{normal} + \lambda_{lap}L_{lap}$$

L_{vert} represents the root mean square distance of the vertices moved, which induces the smallest vertex movements while approaching the target shape. L_{edge} is edge length regularization to prevent skewed mesh. Normal consistency L_{normal}, and Laplacian smoothness L_{lap} are used to derive a smooth deformed mesh.

Results

We analyzed LV shape deformation using our approach in the OASIS longitudinal dataset (Marcus et al. (2010)) of normal (n=10) and demented (n=10) subjects whose baseline ages ranged from 78 to 88 years. To derive the deformation, the distance between the baseline and follow-up shapes is calculated by comparing the original vertex V_0 with the deformed vertex V_k after k optimization iterations. As depicted in Fig.2, we analyzed changes in regions adjacent to thalamus, caudate, hippocampus, amygdala and the opposite LV, as well as in the entire LV. In addition, we conducted a comparative analysis between normal male (n=19, age range=80.1$\pm$5.5) and female (n=19, age range=80.6$\pm$5.0) subjects sourced from the OASIS dataset. Across all examined regions, no statistically significant differences in deformation were observed between sexes (p-value>0.26).

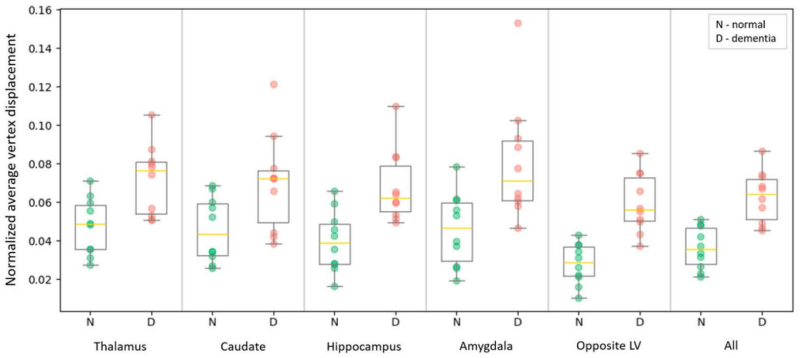


FIGURE 2
Comparison of longitudinal averaged vertex displacement in peripheral brain areas between two groups: normal (n=10) and demented (n=10) subjects. The displacement is normalized by the product of time interval between two longitudinal MRIs (in year) and one-third power of the individual estimated total intracranial volume.

Conclusion

We suggest a deep learning-based LV shape modeling for longitudinal analysis using peripheral area projection that enables the interpretation of local deformations along surrounding areas. We applied our method to the longitudinal dataset, demonstrating noticeable differences in deformation patterns between normal and demented subjects.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.00223446, Development of object-oriented synthetic data generation and evaluation methods) and Korea Center for Gendered Innovations for Science and Technology Research (GISTeR), through the Center for Women in Science, Engineering and Technology (WISET) funded by the Ministry of Science and ICT (WISET202403GI01)

References

Djamanakova, A., Faria, A.V., Hsu, J., Ceritoglu, C., Oishi, K., Miller, M.I., Hillis, A.E., Mori, S. (2013). Diffeomorphic brain mapping based on T1-weighted images: Improvement of registration accuracy by multichannel mapping. Journal of Magnetic Resonance Imaging, 37(1), 76-84.

[Dataset] Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. Journal of cognitive neuroscience, 22(12), 2677-2684.

Qi, C.R., Su, H., Mo, K., Guibas, L.J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 652-660).

Qiu, A. and Miller, M. I. (2008). Multi-structure network shape analysis via normal surface momentum maps. NeuroImage 42, 1430–1438

Styner, M., Oguz, I., Xu, S., Brechb{\"u}hler, C., Pantazis, D., Levitt, J.J., Shenton, M.E., Gerig, G. (2006). Framework for the statistical shape analysis of brain structures using SPHARM-PDM. The insight journal, (1071), 242.

Multiple sclerosis diagnosis with deep learning and explainable AI

Author

Nighat Bibi – School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, Ireland

Jane Courtney – School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, Ireland

Kathleen M. Curran – University College Dublin, School of Medicine, UCD Belfield, Dublin 4, Ireland

Citation

Bibi, N., Courtney, J., Curran, M.K. Multiple sclerosis diagnosis with deep learning and explainable AI.

Abstract

Diagnosing multiple sclerosis (MS) presents significant challenges due to its complex clinical presentation and the subjective interpretation of imaging findings. Machine learning (ML) and deep learning (DL) models, despite their potential, often exacerbate these challenges with their opaque decision-making processes, hindering clinical integration. This study addresses these limitations by employing eXplainable Artificial Intelligence (XAI) techniques, specifically integrating Grad-CAM within a Convolutional Neural Network (CNN) framework, EfficientNetB1, for the diagnosis of MS. The primary objective is to enhance the transparency and reliability of MS diagnosis by providing clear visual insights into the model's decision-making process, while also identifying and mitigating potential biases and irrelevant features. Using a dataset comprising FLAIR axial and sagittal MRI images of MS patients and healthy individuals, the CNN model is trained and integrated with Grad-CAM. Post-integration observations revealed

potential biases and irrelevant features, particularly in the erroneous highlighting of certain regions by the model. Subsequent adjustments and re-training using 10-fold cross-validation led to an improved model with accuracy rates of 99.82% for axial, 99.76% for sagittal images, and 99.36% overall. Furthermore, testing on a separate dataset confirmed the model's ability to generalize and perform well across various clinical contexts. In conclusion, this study underscores the critical role of transparent and interpretable models in medical diagnostics, demonstrating that the integration of XAI techniques can significantly enhance the reliability and clinical applicability of models.

Dataset

The dataset used in this study is obtained from a research article published in 2021, collected from Ozal University Medical Faculty [4]. It consists of FLAIR MRI images of the brains of 72 patients diagnosed with MS and 59 healthy individuals. The dataset is organized into four groups: MS-Axial (n = 650), MSSagittal (n = 761), Healthy-Axial (n = 1002), and Healthy-Sagittal (n = 1014). To evaluate the generalizability of the proposed model, an additional publicly available dataset by Muslim et al. [5] is used, which includes FLAIR MRI images from a different population.

Methods & Results

For the classification of MRI images to diagnose MS, a series of steps were taken. Initially, all images were resized to a fixed dimension of 240x240 pixels to ensure uniformity in data processing. The architecture chosen for the classification task is EfficientNetB1 [7], selected for its proven efficacy in image classification and computational efficiency. Using transfer learning, the model leveraged pre-trained weights from the ImageNet dataset to extract meaningful features from the MRI images. The model architecture consisted of a base EfficientNetB1 model followed by custom layers tailored for classification purposes. To assess the robustness of the model performance, a 10-fold cross-validation strategy is implemented. This involved training the model on nine folds of the data while validating the remaining fold for

each iteration. Following cross-validation, the final model is trained on the entire dataset to maximize its learning potential for subsequent evaluation and testing on unseen data. During the initial 10-fold cross-validation process, the model achieved an accuracy of 98.13% for the overall dataset, combining both axial and sagittal images. Gradient-weighted Class Activation Mapping (Grad-CAM) [6] is integrated to provide visual explanations for the model's predictions. Grad-CAM highlighted regions of the input image that contributed the most to the model's decision-making process. Initially, certain regions on the left side of the images were incorrectly highlighted by the model. To address this issue, images were cropped to focus solely on relevant brain regions. Subsequently, the model is re-trained on the cropped images, leading to improved accuracy. Grad-CAM is then reapplied, effectively highlighting the relevant regions in the MRI images (shown in Figure 1).

To evaluate the generalizability of the model, a separate set of images collected from a different source [5] is used. The model tested on some random images and made accurate predictions on these images and Grad-

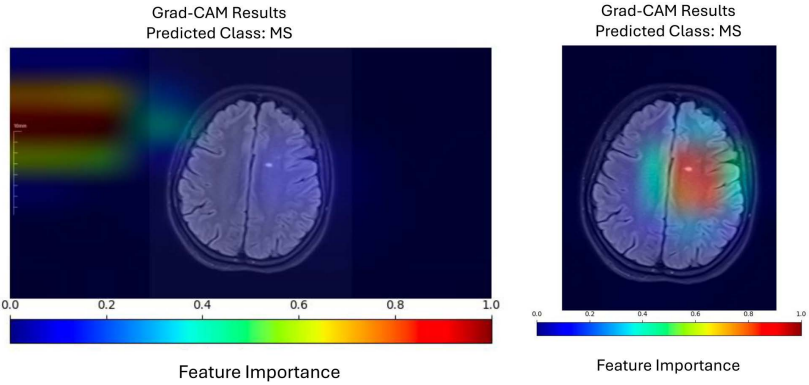


FIGURE 1
Grad-CAM visualisation for MS class before cropping and after cropping the images.

CAM effectively highlighted the regions that influence classification decisions (shown in Figure 2). In general, the integration of Grad-CAM improved the diagnostic capabilities of the model by accurately highlighting regions indicative of Multiple Sclerosis. The visual insights provided by Grad-CAM facilitated better understanding and trust in the model's decision-making process. Upon completion of the 10-fold cross-validation, the proposed CNN architecture achieved a final accuracy of 99.82% for axial images, 99.76% for sagittal images, and 99.36% overall. Furthermore, a comparative analysis demonstrated in Table 1 with existing studies showcased the competitive performance of the proposed model.

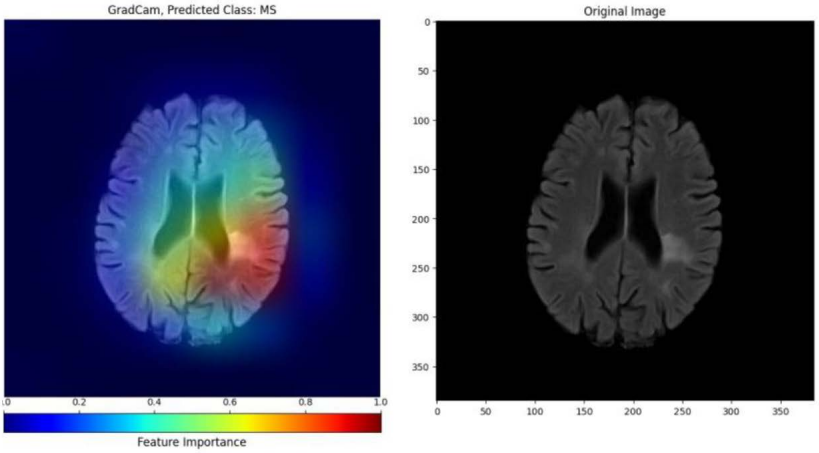


FIGURE 2
Grad-CAM visualisation for MS class on an image from Muslim et al. [5].

TABLE 1: Comparison of model performance with existing studies

Study	Dataset	Accuracy (%)	Methodology	XAI Method
Eitel et al. [1]	Overall dataset	87.04	CNN	LRP
Lopatina et al. [3]	Overall dataset	91 to 95	CNN	Attribution algorithms
	Axial plane images	98.37	(ExMPLPQ,	
Macin et al. [4]	Sagittal plane images	97.75	INCA,	No
	Overall dataset	98.22	KNN)	
	Axial plane images	99.76	(exemplar MobileNetV2,	
Ekmekyapar et al. [2]	Sagittal plane images	99.48	IMrMr,	No
	Overall dataset	98.02	KNN)	
	Axial plane images	**99.82**		
Our Work	**Sagittal plane images**	**99.76**	**EfficientNetB1**	**Grad-CAM**
	Overall dataset	**99.36**		

Conclusion

In conclusion, this work presents an approach to the diagnosis of multiple sclerosis using eXplainable Artificial Intelligence techniques to improve the explainability of the model. Achieving higher classification accuracy than existing work on the same dataset with Grad-CAM integration provides valuable insights into the decision-making process, highlighting potential biases and irrelevant features learned by the model. By addressing these issues, this research underscores the importance of transparent and interpretable models in medical diagnostics. Furthermore, it demonstrates robust performance on a separate unseen dataset and affirms its generalizability.

Looking ahead, future research will explore the incorporation of multiple modalities and advanced XAI techniques to further improve diagnostic accuracy and broaden the scope of our findings. Clinical user evaluation will

also be performed in collaboration with clinicians to develop methods and metrics toward clinical interpretability.

Acknowledgements
This research is funded by the Science Foundation Ireland Center for Research Training in Machine Learning (18/CRT/6183).

References

[1] Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A.U., Ruprecht, K., Giess, R.M., Kuchling, J., Asseyer, S., Weygandt, M., Haynes, J.D., et al.: Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation. NeuroImage: Clinical 24, 102003 (2019)

[2] Ekmekyapar, T., Taşcı, B.: Exemplar mobilenetv2-based artificial intelligence for robust and accurate diagnosis of multiple sclerosis. Diagnostics 13(19), 3030 (2023)

[3] Lopatina, A., Ropele, S., Sibgatulin, R., Reichenbach, J.R., Güllmar, D.: Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. Frontiers in neuroscience 14, 609468 (2020)

[4] Macin, G., Tasci, B., Tasci, I., Faust, O., Barua, P.D., Dogan, S., Tuncer, T., Tan, R.S., Acharya, U.R.: An accurate multiple sclerosis detection model based on exemplar multiple parameters local phase quantization: Exmplpq. Applied Sciences 12(10), 4920 (2022)

[5] Muslim, A.M., Mashohor, S., Al Gawwam, G., Mahmud, R., binti Hanafi, M., Alnuaimi, O., Josephine, R., Almutairi, A.D.: Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. Data in Brief 42, 108139 (2022)

[6] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

[7] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

Post-processing of perivascular spaces segmentation using k-means

Author

Roberto Duarte Coello – Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

Maria Valdés Hernández – Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

Jose Bernal – Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom; German Centre for Neurodegenerative Diseases, Magdeburg, Germany

Joanna Wardlaw – Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

Citation

Coello, R.D., Hernández, M.V., Bernal, J., Wardlaw, J. Post-processing of perivascular spaces segmentation using k-means.

Abstract

The widespread visibility of perivascular spaces (PVS) in the brain, as seen with magnetic resonance imaging (MRI), is thought to be an early biomarker of brain health dysfunction. PVS can, in principle, be counted using computational methods. However, when numerous PVS are situated near each other, their counts are often underestimated, potentially hindering subsequent analysis. Here, we propose a pruning step using the k-means clustering algorithm to learn the PVS and background intensities of each object and automatically select a threshold for refining the segmentation. In

a preliminary analysis, we qualitatively and quantitatively show that such an approximation leads to improved performance when compared to the state-of-the-art.

Introduction

Perivascular Spaces (PVS) are millimetre-sized tubular structures that surround perforating blood vessels of the brain. PVS can grow in size to the point where they become visible on MRI---a structural alteration that is considered an early biomarker of brain health dysfunction, as it has been found to be associated with a plethora of conditions, including cardiovascular risk factors as well as other features of small vessel disease [1].

PVS can be segmented nowadays using classical as well as machine learning-based techniques [2], and this has proven to be effective. However, all these methods require careful tuning of segmentation parameters (especially thresholds), which can lead to over- or under-segmentation of PVS, hindering any subsequent morphological analysis and limiting comparability across studies. Over-segmentation of PVS, for instance, can give the erroneous impression that a subject with many closely spaced PVS has a lower PVS count than someone else with more dispersed PVS (see Figure 1 – A). To address this problem, we propose a post-processing step relying on intensity information of the region of interest and its surroundings to decide whether voxels are part of the PVS or the background.

Methods

The proposal is as follows. After running the main processing algorithm (e.g., deep learning or Hessian-based filtering strategies), we obtain a heatmap, in which higher values indicate a higher likelihood of a voxel containing a segment of a PVS. We then threshold this heatmap and, for each detected object, we perform the following steps:

- We dilate the mask of each individual object to include its surroundings.
- We then perform k-means [3] clustering with two clusters, one representing the background and another the PVS intensity.

- We compute a threshold by selecting an intensity value that best separates PVS from the background (e.g. (background+(PVS-background)/2).

- We use said threshold to update the segmentation.

We carry out this process iteratively until the number of objects does not change.We analised a subset of 100 scans from the TREAT dataset [4] and compared the automatic count to the PVS visual scores described in [5]. This score is obtained by counting the number of PVS in a representative slice and is rated as follows: 0 (none), 1 (1-10), 2 (11-20), 3 (21-40) and 4 (>40).

Results

We use the Frangi filter---a popular Hessian-based PVS enhancement method---to showcase our proposal. Two examples of PVS clusters can be seen in the first column of Figure 1 and the PVS segmentation of the

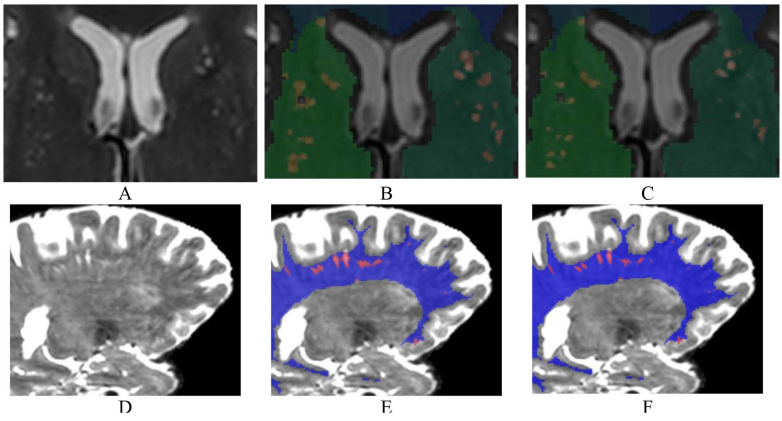


FIGURE 1
The first row is an example of the pruning algorithm in the Basal Ganglia (BG) region (green). Original T2-weighted image (A). PVS segmentation using Frangi + Thresholding (B). Post-processed PVS segmentation using the proposed pruning algorithm (C). The second row is an example of the pruning algorithm in the Centrum Semiovale (CSO) region (blue). Original T2-weighted image (D). PVS segmentation using Frangi + Thresholding (E). Post-processed PVS segmentation using the proposed pruning algorithm (F).

images using the Frangi filter + Thresholding method is shown in the second column. Due to the close proximity of the PVS, the filter erroneously identifies the voxels in the middle as part of a larger object. The result of the pruning algorithm can be seen in the last column of Figure 1. The clusters have been successfully separated, and the PVS shape is better preserved.

The total count of the PVS in the given region is expected to be highly correlated with the PVS visual scores. As seen in Figure 2, a higher number

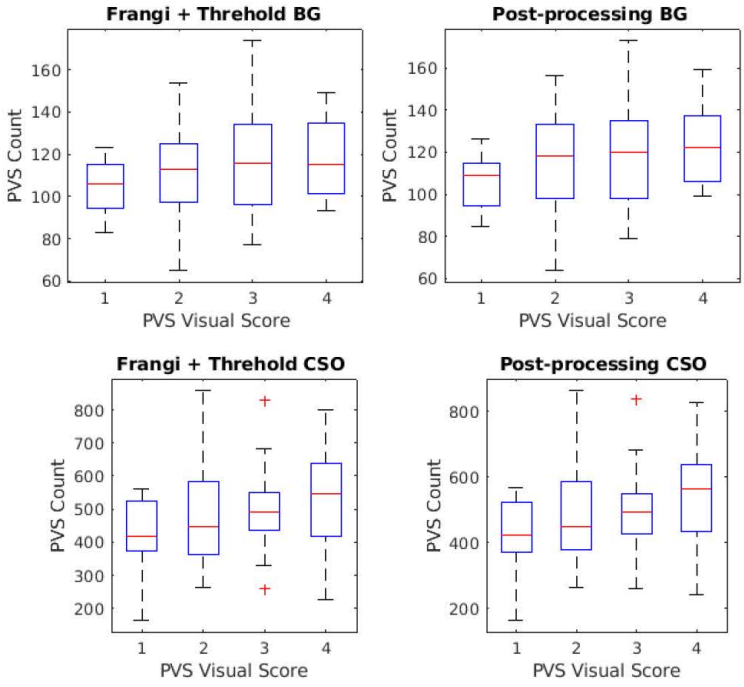


FIGURE 2
Comparison between PVS visual scores and the total count of PVS using+ both methods in the BG region (left) and the CSO (right).

of PVS corresponds to higher PVS visual score in the CSO. The count of PVS without the post-processing is similar for the 3 and 4 visual scores due to the clusters in the BG. The post-processing step led to an increase in the Spearman's rank correlation coefficient for both the BG and CSO regions. For the BG, the Spearman's rank correlation coefficient increased from 0.1605 to 0.1955. For the CSO, the Spearman's rank correlation coefficient increased from 0.2360 to 0.2781. These results indicate that the post-processing step effectively improved the correlation between the PVS visual scores and the PVS total count, potentially reflecting a more accurate count of individual PVS with the proposed approach.

References

[1] Wardlaw, Joanna M., et al. "Perivascular spaces in the brain: anatomy, physiology and pathology." Nature Reviews Neurology 16.3 (2020): 137-153.

[2] Waymont, Jennifer MJ, et al. "A Systematic Review and Meta-Analysis of Automated Methods for Quantifying Enlarged Perivascular Spaces in the Brain." medRxiv (2024): 2024-03.

[3] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the royal statistical society. series c (applied statistics) 28.1 (1979): 100-108.

[4] Kopczak, Anna et al. "The EffecTs of Amlodipine and other Blood PREssure Lowering Agents on Microvascular FuncTion in Small Vessel Diseases (TREAT-SVDs) trial: Study protocol for a randomised crossover trial." European stroke journal vol. 8,1 (2023): 387-397.

[5] Potter, Gillian M et al. "Cerebral perivascular spaces visible on magnetic resonance imaging: development of a qualitative rating scale and its observer reliability." Cerebrovascular diseases (Basel, Switzerland) vol. 39,3-4 (2015): 224-31.

Predictive Bayesian Active Learning in Stargardt disease diagnosis

Author

Biraja Ghoshal – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Shihan Zhao – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom

William Woof – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Bernardo Mendes – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Saoud Al-Khuzaei – Oxford Eye Hospital, Oxford, United Kingdom

Thales Antonio Cabral De Guimaraes – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Malena Daich Varela – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Yichen Liu – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom

Sagnik Sen – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Siying Lin – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Yu Fujinami-Yokokawa – Laboratory of Visual Physiology, Division of Vision Research, National Institute of Sensory Organs, National Hospital Organization Tokyo Medical Center, Japan
Andrew R. Webster – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom
Omar A. Mahroo – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom
Kaoru Fujinami – Laboratory of Visual Physiology, Division of Vision Research, National Institute of Sensory Organs, National Hospital Organization Tokyo Medical Center, Japan
Savita Madhusudhan – St Paul's Eye Unit, Liverpool University Hospitals NHS Foundation Trust, Liverpool, United Kingdom
Konstantinos Balaskas – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom
Susan M Downes – Oxford Eye Hospital, Oxford, United Kingdom
Michel Michaelides – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom
Nikolas Pontikos – University College London Institute of Ophthalmology, 11 - 43 Bath Street, London EC1V 9EL, United Kingdom; Moorfields Eye Hospital, NHS Foundation Trust, 162 City Road, London EC1V 2PD, United Kingdom

Citation
Ghoshal, B., Zhao, S., Woof, W., Mendes, B., Al-Khuzaei, S., De Guimaraes, T.A.C., Varela, M.D., Liu, Y., Sen, S., Lin, S., Fujinami-Yokokawa, Y., Webster, A.R., Mahroo, O.A., Fujinami, K., Madhusudhan, S., Balaskas, K., Downes, S.M., Michaelides, M., Pontikos, N. Predictive Bayesian Active Learning in Stargardt disease diagnosis.

Abstract
Deep Learning has achieved state-of-the-art performance for medical image analysis but requires a large number of labelled images to obtain

good performance. Labelling a sufficiently large number of these images is often challenging in resource-constrained environments. In this paper, three different uncertainty-based active learning (AL) strategies (Random, Entropy and Bayesian Active Learning by Disagreement) combined with the Monte Carlo dropout approach were compared to baseline using inherited retinal diseases (IRD) genetic diagnosis to investigate the utility of uncertainty in gene classification of retinal images. Genetic variants in four genes (***ABCA4***, ***PROM1***, ***BEST1*** and ***PRPH2***) have been reported to be associated with a phenotype similar to the appearance seen in Stargardt disease (STGD). A refined hybrid active learning approach, using autoencoder-KMeans clustering and combining uncertainty in Bayesian deep learning, was applied to the classification of associated genes from retinal images of IRD patients affected by Stargardt or its phenocopies. These approaches reduce the required number of genetically diagnosed Stargardt patients for training the gene classifier. The hybrid approach significantly enhanced the model accuracy in the last iteration ($p = 0.014$), and the maximal accuracy was $85 \pm 1.8\%$. The classification uncertainties were also mapped to the image pixels as interpretability maps. The hybrid AL approach can enhance model prediction accuracy with limited labelling in Stargardt genetic diagnosis, potentially optimising the prescription of genetic testing to patients in resource-constrained environments, while improving model interpretability with classification uncertainty mapping.

Introduction

In recent years, uncertainty-aware active learning (AL) strategies have emerged that integrate both the whole-image and feature-wise uncertainty approaches for enhancing medical image analysis Shi et al. (2019); Smailagic et al. (2020); Ghoshal et al. (2021); Wu et al. (2021); Huang et al. (2017); Deng et al. (2009); Beluch et al. (2018); Paul et al. (2022); Zheng et al. (2020). These frameworks have demonstrated impressive performance and annotation benefits. However, extending these frameworks to different domains poses a challenge due to cross-domain adaptation complexities Guan and Liu (2021); the model performance fluctuates due to the diversity of the selected samples even within the same sample but in different modalities.

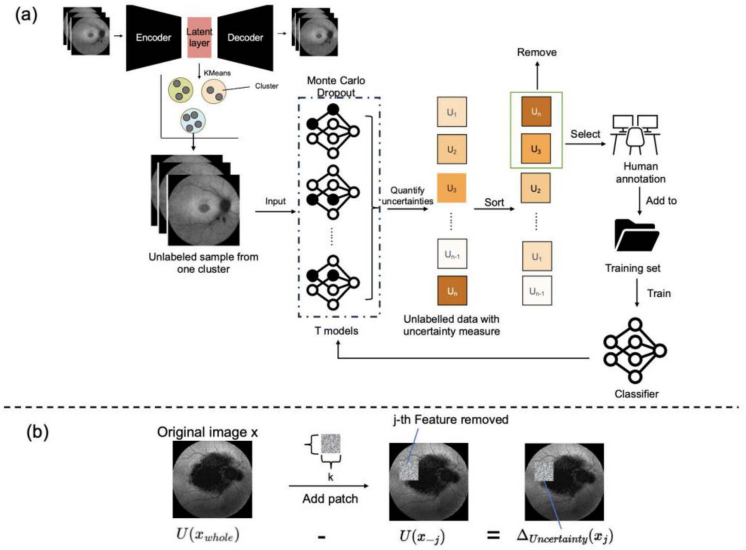


FIGURE 1

End-to-End Active Learning Framework: **(a)** The unlabelled images from the pool are passed through the trained VGG16 Autoencoder and they are clustered using the latent presentation in the latent layer using K-Means clustering. From each cluster, the uncertainties of every sample are captured by Monte Carlo dropout (dashed-line box; black nodes: dropped node set by specified dropout rate). The batch of the most uncertain samples (represented by colour darkness) will be taken for annotation and added to the training dataset, which is then used for model training. The model enters the next iteration to select more samples for labelling. **(b)** An explanation of the uncertainty alteration mapping. The uncertainty alteration ΔU after the j-th feature is shown after patching. An array of patched images is generated, and their corresponding epistemic uncertainty is quantified Houlsby et al. (2014). The resulting uncertainty alterations are visualised as a saliency map, calculated by the formula Chai (2018).

Inherited retinal diseases (IRDs) are monogenic disorders of the retina and a leading cause of blindness in children and working-age adults Galvin et al. (2020). Genetic variants in four genes (ABCA4, PROM1, BEST1 and PRPH2) have been reported to be associated with a phenotype similar to the appearance seen in Stargardt disease (STGD) which is the most common form of IRD patients affected by Stargardt or its phenocopies. Recent work

has shown that deep-learning-based approaches can be effective at identifying the causative gene from widely available retinal imaging Pontikos et al. (2022), however, such approaches are data intensive. Due to the sparsity of IRDs and the challenges in accessing genetic testing, especially in resource-constrained environments and low-middle income countries, large datasets of patients with both retinal imaging and a corresponding gene diagnosis are scarce, especially given the costs associated with genetic diagnosis of patients. Hence data efficient training regimes are important, particularly for rarer genes with fewer identified cases. This suggests that there's a promising avenue to explore the application of hybrid active learning approaches utilising uncertainty and diversity in this field to optimise the training regime.

We designed a novel Bayesian neural network-based Active Learning sample selection strategy for the prediction of IRD genetic diagnosis using high dimensional retinal fundus autofluorescence (FAF) images, as shown in Figure 1. Using the differentiation of different Stargardts-associated genes as an example, we show that integrating active learning can significantly reduce the number of labelled samples while achieving higher prediction accuracy and model interpretability.

Methods

Dataset

The data was gathered from Moorfields Eye Hospital, Oxford University Hospital, Liverpool University Hospital, and Tokyo Medical Centre. Specifically, patients with IRD and documented genetic diagnoses spanning from 2006 to 2018 were chosen against the data acquisition protocol Nguyen et al. (2023). The resultant dataset of Stargardt patients comprises 6,410 images of blue light autofluorescence across four distinct genes: ABCA4 (3501 images), PROM1 (180 images), BEST1 (855 images), and PRPH2 (815 images). This dataset was then partitioned into 5,351 images for the training set and 1,059 images for the validation set. The images were pre-processed to a size of 224x224x3.

Implementation

The images from the pool were grouped using the VGG16 Autoencoder and K-Means clustering method, 56 where the number of clusters was set to four indicated by the silhouette score Pedregosa et al. (2011); 57 Simonyan and Zisserman (2014). The model was initialised with 10 labelled samples randomly selected 58 from each gene. Subsequently, 20 images were selected per batch and cluster based on the acquisition functions. Therefore, a total of 200 samples were selected per iteration. A dropout rate of 0.4 was applied to the next hidden dense layer of size 256, and this layer was connected to the following dense layer of size 128 with a dropout rate of 0.25. The downstream output layer was of size 10. Class weights were computed to counteract the potential bias towards the dominant class. The model was compiled with the Adam optimiser, using a learning rate of 0.01 and reducing by 0.5 every 100 epochs until it reaches the minimal learning rate of 1e-5. The categorical cross-entropy was chosen as the loss function. In every iteration, the model was trained for 100 epochs, with batch size 32. A model trained by the entirely whole labelled dataset was used to set the upper boundary of the AL process. For testing, Stochastic passes in MC dropout were set to 100. The experiment used a GPU and the TensorFlow Keras package in Python language Chollet et al. (2015); Abadi et al. (2015). The active learning process was carried out within the modAL framework with customised query strategies Danka and Horvath (2018).

Results

We have observed that the model acquired more information from the images belonging to the ABCA4 class, making patterns from other classes more informative for the model, and the model had a tendency to 72 prioritise under-represented classes. This preference for minor classes occurred in the middle of the entire process, with the class distribution reverting to a pattern resembling random sampling in the final iterations. Whereas Pixel-wise uncertainty (typical images and their corresponding mappings) is depicted in Figure 2. Regions that contribute to the classification would exhibit higher uncertainty after those regions were removed, as visualised by the red colour. In contrast, regions that confused

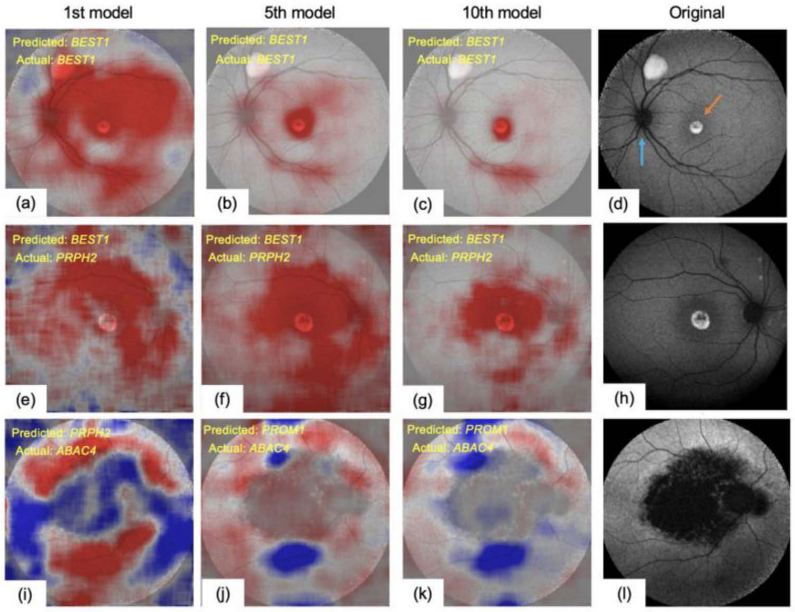


FIGURE 2
Three examples of applying epistemic uncertainty alteration to interpret model performance for gene classification. Image components are highlighted (orange arrow: macula, blue arrow: optic disc). The red regions contribute to the classification, while the blue regions confuse the model classification. For BEST1 in the first row, the smaller central area of the macula with the vitelliform lesion characteristic of this form of the disease is highlighted in red. For PRPH2 in the second row, a broader region of the macular is highlighted showing that the disease is less localised as would be expected. For ABCA4 in the third row, more peripheral regions are highlighted in red which corroborates with the disease being more peripheral due to the presence of flecks in the outer periphery.

the classification were depicted in blue. The intensity of coloured regions represented the degrees of contribution or confusion. In the mapping, the 1st model 'considered' that the optic disc, macula, and blood vessels contributed to the classification, as highlighted by red regions. There were also some pale blue regions around the peripheral area, suggesting that

the peripheral regions around the fundus did not contribute significantly to the classification. However, by the 5th model, the red regions had narrowed to the macula only, with only faint red areas remaining around the optic disc and vessels. Finally, in the last model, the red regions further concentrated on the macula and exhibited strong intensity. Similar patterns of bright fluorescence in the central macula were shown in Figure 2(e)-(g), and all three models misclassified the image as BEST1 based on the hyperfluorescent macula and its surrounding regions. The model at the first iteration misclassified 2(l) as PRPH2, and the corresponding mapping did not indicate any apparent patterns, suggesting that this misclassification was more like a random guess. The subsequent two models also performed misclassification, but Figure 2(j)(k) showed that the components in the peripheral regions made the model more certain about its classification.

Conclusion

Although the uncertainty-based Active Learning approach (Random, Entropy and Bayesian Active Learning by Disagreement) did not significantly improve the model performance, however, the hybrid approach significantly enhanced the model accuracy in the late iteration ($p = 0.014$), and the maximal accuracy was $85 \pm 1.8\%$. In summary, these results demonstrate that uncertainty can serve a dual purpose, not only as a sampling strategy in the Active Learning process but also as a means to map important features for classification onto the image, enhancing interpretability, such as, for example, capturing flecks in the peripheral region Cremers et al. (2020). If more subtle or smaller resolution features need to be captured with the uncertainty map, smaller patches could be employed Zintgraf et al. (2017).

Acknowledgments

The views expressed are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 9368–9377

Chai, L. R. (2018). Uncertainty estimation in Bayesian neural networks and links to interpretability. Master's Thesis, Massachusetts Institute of Technology

Chollet, F. et al. (2015). Keras. Available at: https://github.com/fchollet/keras.

Cremers, F. P., Lee, W., Collin, R. W., and Allikmets, R. (2020). Clinical spectrum, genetic complexity and therapeutic approaches for retinal disease caused by abca4 mutations. Progress in retinal and eye research 79, 100861

Danka, T. and Horvath, P. (2018). modAL: A modular active learning framework for Python. arXiv preprint arXiv:1805.00979

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (Ieee), 248–255

Galvin, O., Chi, G., Brady, L., Hippert, C., Del Valle Rubido, M., Daly, A., et al. (2020). The impact of inherited retinal diseases in the Republic of Ireland (ROI) and the United Kingdom (UK) from a Cost-of-Illness perspective. Clin. Ophthalmol. 14, 707–719

Ghoshal, B., Swift, S., and Tucker, A. (2021). Bayesian deep active learning for medical image analysis. In Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings (Springer), 36–42

Guan, H. and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering 69, 1173–1185

Houlsby, N., Hernández-Lobato, J. M., and Ghahramani, Z. (2014). Cold-start active learning with robust ordinal matrix factorization. In International conference on machine learning (PMLR), 766–774

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4700–4708

Nguyen, Q., Woof, W., Kabiri, N., Sen, S., Varela, M. D., Guimaraes, T. A. C. D., et al. (2023). Can artificial intelligence accelerate the diagnosis of inherited retinal diseases? protocol for a data-only retrospective cohort study (eye2gene). BMJ Open 13. doi:10.1136/bmjopen-2022-071043

Paul, S. K., Pan, I., and Sobol, W. M. (2022). Efficient labeling of retinal fundus photographs using deep active learning. Journal of Medical Imaging 9, 064001–064001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830

Pontikos, N., Woof, W., Krawitz, P., Arno, G., Hess, K., Varela, M. D., et al. (2022). Eye2gene: prediction of causal inherited retinal disease gene from multimodal imaging using ai. Investigative Ophthalmology & Visual Science 63, 1161–1161

Shi, X., Dou, Q., Xue, C., Qin, J., Chen, H., and Heng, P.-A. (2019). An active learning approach for reducing annotation cost in skin lesion analysis. In Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10 (Springer), 628–636160

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Smailagic, A., Costa, P., Gaudio, A., Khandelwal, K., Mirshekari, M., Fagert, J., et al. (2020). O-medal: Online active deep learning for medical image analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, e1353

Wu, X., Chen, C., Zhong, M., Wang, J., and Shi, J. (2021). COVID-AL: The diagnosis of COVID-19 with deep active learning. Medical Image Analysis 68, 101913

Zheng, H., Zhang, Y., Yang, L., Wang, C., and Chen, D. Z. (2020). An annotation sparsification strategy for 3d medical image segmentation via representative selection and self-training. In Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, 6925–6932

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595

Specimen-to-tumor bed deformable registration to inform re-resection in otolaryngologic procedures

Author

Morgan Ringel – Vanderbilt University, Department of Biomedical Engineering, Nashville, TN

Ayberk Acar – Vanderbilt University, Department of Computer Science, Nashville, TN USA

Qingyun Yang – Vanderbilt University, Department of Computer Science, Nashville, TN USA

Marina Aweeda – Vanderbilt University Medical Center, Department of Otolaryngology-Head and Neck Surgery, Nashville, TN USA

Carly Fassler – Vanderbilt University Medical Center, Department of Otolaryngology-Head and Neck Surgery, Nashville, TN USA

Jon Heiselman – Vanderbilt University, Department of Biomedical Engineering, Nashville, TN USA ; Memorial Sloan-Kettering Cancer Center, Department of Surgery, New York, NY, USA

Jie Ying Wu – Vanderbilt University, Department of Biomedical Engineering, Nashville, TN USA; Vanderbilt University, Department of Computer Science, Nashville, TN USA

Michael Topf – Vanderbilt University, Department of Biomedical Engineering, Nashville, TN USA; Vanderbilt University Medical Center, Department of Otolaryngology-Head and Neck Surgery, Nashville, TN USA

Michael Miga – Vanderbilt University, Department of Biomedical Engineering, Nashville, TN USA; Vanderbilt University, Department of Computer Science, Nashville, TN USA; Vanderbilt University Medical Center, Department of Otolaryngology-Head and Neck Surgery, Nashville, TN USA

Citation
Ringel, M., Acar, A., Yang, Q., Aweeda, M., Fassler, C., Heiselman, J., Wu, J.Y., Topf, M., Miga, M. Specimen-to-tumor bed deformable registration to inform re-resection in otolaryngologic procedures.

Introduction

Surgical resection is a common treatment for head and neck cancers, and re-resection procedures are indicated when close or positive margins are detected on an excised specimen. Frozen section histology is often used for intraoperative margin assessment, and there is a clinical need to determine correspondence from histopathological imaging back to the surgical tumor bed [1]. Previous work has focused on using 3D scanned specimen models to improve communication between pathologists and surgeons to assist in identifying remaining high-risk areas for further resection [2]. However, deformation of the resected cancer specimen remains a concern with this method. Correcting for these deformations is crucial if 3D scans of surgical specimens are to be used in clinical care to assist the surgical team in more precisely identifying the corresponding positive margin on the 3D specimen model and the precise area for re-resection in the tumor bed.

This work proposes a framework for registering a 3D excised specimen model back to the tissue bed for re-resection guidance. This methodology is demonstrated in a proof-of-concept mock study of a left partial rhinectomy procedure performed on a cadaver specimen. While true intraoperative feasibility and utility has yet to be evaluated, this framework may be useful for locating regions of interest for additional resection in a tumor bed in instances where positive or close margins have been detected.

Experimental Setup

This study investigated cadavers with approval from the Vanderbilt University Medical Center's Center for Experiential Learning and Assessment. Dissection was performed on one fresh-frozen cadaver. Prior to incision, the surgical plan for a left partial rhinectomy was marked on the skin surface with an ink pen (Figure 1A). Sutures were placed to mark three corresponding points

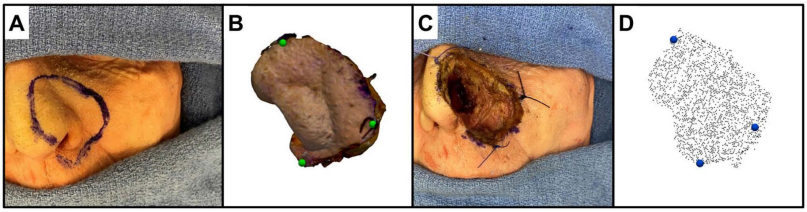


FIGURE 1
Data collection and pre-processing of a left partial rhinectomy specimen. (A) Marked resection plan. (B) 3D specimen model with suture points marked in green. (C) Tissue bed. (D) Tissue bed point cloud data with suture points marked in blue.

between the tumor bed and specimen during the procedure. These three points were used for initial rigid alignment between the tumor bed and specimen prior to deformable correction. After excision, a 3D model of the specimen geometry was generated using a commercially available structured light 3D scanner (EinScan SP, Shining 3D, Hangzhou, China) and companion software (EXScan, Shining 3D), (Figure 1B). Geometric tissue bed data after excision was captured using an overhead mounted stereo camera (ZED 2i, Stereolabs Inc., San Francisco, CA, USA), (Figure 1C).

Data pre-processing steps included cropping the stereo camera point cloud data and manually labeling the suture points on the specimen model and in the tissue bed. The point cloud data was smoothed and down-sampled to create a geometric representation of the tissue bed to use for registration (Figure 1D).

Specimen-to-Tumor Bed Cavity Registration

The registration task is to deform the 3D specimen model to better align with the resection cavity geometric data. After resection, tissue specimens are known to deform and compromise correspondence to the resection bed [3], [4]. In the instance of a positive margin, the loss of correspondence between the resected specimen and the resection bed makes it challenging to spatially locate the appropriate area and extent for re-resection. In fact, in oral cavity cancer surgery, re-resections only contain additional cancer

29% of the time [5]. The method proposed reestablishes correspondence for guiding re-resection.

A previously reported sparse data image-to-physical registration method, the linearized iterative boundary reconstruction (LIBR) method, is proposed to reestablish correspondence [6]. An adaptation of the LIBR method that uses regularized Kelvinlet functions to reduce computation time was also employed [7], [8]. Briefly, the LIBR method deploys a set of control points on the specimen boundary to create a deformation basis. Then, the linear combination of these deformations that reduces the model-data error between the deformed specimen model and the tissue bed geometric data is computed using Levenberg-Marquardt optimization. Rigid registration parameters, including a scaling factor, and a strain energy regularization parameter are also included in the optimization. For more implementation details, see [6]. The scanned posterior specimen surface (acquired by Einscan SP), the corresponding scanned tissue bed geometric data (acquired by ZED 2i), and a corresponding set of suture fiducial locations (3 sutures on specimen and tissue bed) were the only data used. After initializing with a point-based rigid registration computed from the suture fiducial points, 45 control points distributed on the specimen boundary with a regularized Kelvinlet radial scale parameter of 0.01 m and a strain energy regularization weight of 10^{-11} Pa^{-2} were utilized in the optimization that non-rigidly aligned specimen to tissue bed. Model material properties were set at a Young's Modulus of 2100 Pa and Poisson's ratio of 0.45 to approximate biological tissue. To measure registration performance, fiducial registration error for the suture points and the average normal projected surface error between the specimen model surface and cavity data were computed.

Results and Discussion

Registration results are shown in Figure 2 before (A) and after (B) deformable correction. Fiducial registration errors and surface errors before and after deformable correction are reported in Table 1. As expected, both fiducial and surface errors improve with deformable correction compared to rigid registration alone.

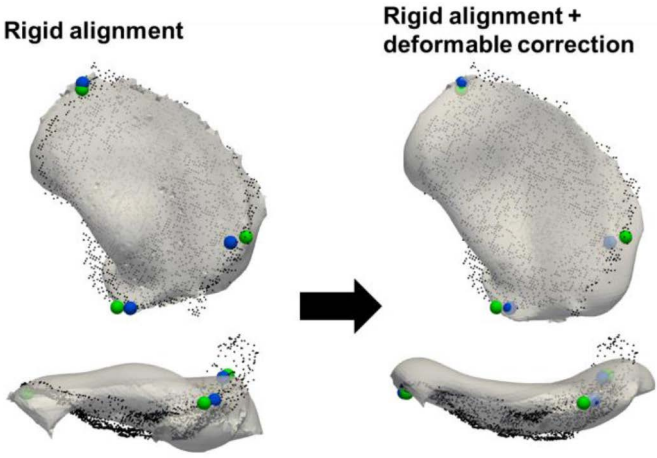


FIGURE 2
Specimen model (gray) registered to tissue bed point cloud data (black) before and after deformable correction. Suture points from the specimen and tissue bed are shown in green and blue, respectively.

TABLE 1: Registration results. Avg ± Std (Maximum) mm

Suture fiducial error (mm)		Projected surface error (mm)	
Rigid	Deformable	Rigid	Deformable
2.0 ± 0.9 (2.7)	1.8 ± 0.5 (2.3)	3.3 ± 2.4 (10.5)	2.3 ± 1.8 (7.5)

The methodology proposed here could also be expanded for other soft-tissue surgeries. For example, additional specimens are often excised after a first-pass breast lumpectomy surgery if positive margins are found during rapid intraoperative margin assessment. This same methodology could be utilized to help surgeons relocate the area for re-resection in the tumor bed with more precision than relying on anatomical annotations alone (i.e., "Excise an additional superior-posterior specimen"). Novel methods to display registration results during surgery, such as those utilizing augmented reality head-mounted displays to overlay specimen models on the surgical

scene, could benefit from more accurate specimen-to-tumor bed alignment [9]. While the results are encouraging, additional studies are needed with more resected specimens under conditions of varied presentation and will be the focus of future work.

Acknowledgements
This work is supported by NIH award R01EB027498, Vanderbilt University Seeding Success Grant, the Vanderbilt Institute for Surgery and Engineering (VISE) Physician-in-Residence program, a Vanderbilt Clinical Oncology Research Career Development Program (K12 NCI 2K12CA090625-22A1) and AHNS/AAO-HNSF Young Investigator Combined Award.

References

[1] F. J. Voskuil *et al.*, "Intraoperative imaging in pathology-assisted surgery," *Nature Biomedical Engineering 2021 6:5*, vol. 6, no. 5, pp. 503–514, Nov. 2021, doi: 10.1038/s41551-021-00808-8.

[2] K. F. Sharif *et al.*, "The computer-aided design margin: Ex vivo 3D specimen mapping to improve communication between surgeons and pathologists," *Head Neck*, vol. 45, no. 1, pp. 22–31, Jan. 2023, doi: 10.1002/HED.27201.

[3] L. A. Umstattd, J. C. Mills, W. A. Critchlow, G. J. Renner, and R. P. Zitsch, "Shrinkage in oral squamous cell carcinoma: An analysis of tumor and margin measurements in vivo, post-resection, and post-formalin fixation," *Am J Otolaryngol*, vol. 38, no. 6, pp. 660–662, Nov. 2017, doi: 10.1016/J.AMJOTO.2017.08.011.

[4] R. C. Mistry, S. S. Qureshi, and C. Kumaran, "Post-resection mucosal margin shrinkage in oral cancer: quantification and significance," *J Surg Oncol*, vol. 91, no. 2, pp. 131–133, Aug. 2005, doi: 10.1002/JSO.20285.

[5] K. Prasad *et al.*, "How Often is Cancer Present in Oral Cavity Re-resections After Initial Positive Margins?," *Laryngoscope*, vol. 134, no. 2, p. 717, Feb. 2024, doi: 10.1002/LARY.30959.

[6] J. S. Heiselman, W. R. Jarnagin, and M. I. Miga, "Intraoperative Correction of Liver Deformation Using Sparse Surface and Vascular Features via Linearized Iterative Boundary Reconstruction," *IEEE Trans Med Imaging*, vol. 39, no. 6, pp. 2223–2234, Jun. 2020, doi: 10.1109/TMI.2020.2967322.

[7] F. De Goes and D. L. James, "Regularized Kelvinlets: Sculpting Brushes based on Fundamental Solutions of Elasticity," *ACM Trans. Graph*, vol. 36, 2017, doi: 10.1145/3072959.3073595.

[8] M. Ringel, J. Heiselman, W. Richey, I. Meszoely, and M. Miga, "Regularized Kelvinlet Functions to Model Linear Elasticity for Image-to-Physical Registration of the Breast," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14228 LNCS, pp. 344–353, 2023, doi: 10.1007/978-3-031-43996-4_33.

[9] G. Tong *et al.*, "Development of an augmented reality guidance system for head and neck cancer resection," *Healthc Technol Lett*, vol. 11, no. 2–3, pp. 93–100, Apr. 2024, doi: 10.1049/HTL2.12062.

Support classification system for glaucoma detection

Author

Dmytro Furman – FIIT, Slovak University of Technology, Bratislava, Slovak Republic
Bryan Williams – SCC, Lancaster University, Lancaster, UK
Silvester Czanner – FIIT, Slovak University of Technology, Bratislava, Slovak Republic; FSBE, University of Chester, Parkgate Road, Chester, UK
Gabriela Czanner – FIIT, Slovak University of Technology, Bratislava, Slovak Republic; FET, Liverpool John Moores University, Liverpool, U

Citation

Furman, D., Williams, B., Czanner, S., Czanner, G. Support classification system for glaucoma detection.

Abstract

Glaucoma is the leading cause of irreversible blindness worldwide affecting 3% of the global population. Early detection of glaucoma cases is therefore the way to prevent vision loss. Fundus Photography, which captures images of the retina, is a common imaging methodology used in ophthalmology. Automated analysis of fundus images using machine learning techniques has shown promise for detecting glaucoma. In this study, we propose a novel approach that integrates Variational Mode Decomposition (VMD) with Convolutional Neural Networks (CNNs) for automated glaucoma detection in fundus images. Preprocessing fundus images is crucial for glaucoma detection as it enhances relevant features while reducing noise, ensuring accurate identification of subtle structural changes associated with the disease. By integrating signal processing techniques like VMD with deep learning models, we want to enhance the preprocessing of medical images

and improve diagnostic accuracy. Our findings contribute to the ongoing efforts to develop robust and efficient tools for early glaucoma detection, ultimately leading to improved patient outcomes and reduced healthcare costs.

Introduction

Glaucoma is a chronic progressive optic neuropathy and the leading cause of irreversible vision impairment globally, with cases continuously rising worldwide [1]. Changes in the structure of the optic nerve head (ONH) and retinal nerve fiber layer (RNFL) are associated with visual defects [3] and manifest by a slow progressive narrowing of the neuroretinal rim. Early detection is crucial to allow timely intervention and prevent further visual field loss. This requires examination of the optic nerve head via fundus imaging by a clinician. At the center of diagnosis is the assessment of the optic cup and disc boundaries [2]. Fundus imaging is noninvasive and low-cost [4]; however examination relies on subjective, time-consuming, and costly expert assessments.

AI for glaucoma detection can be split into two approaches: one-step and two-step approach. In a one-step, the AI detects glaucoma in a single step. The best way to do it is via a deep learning black-box approach, also called end-to-end approach [3]. Two-step approach, as the name of this approach suggests, is the process of AI involves two distinct steps, the first of which consists of AI that segments images automatically detect and trace the optic cup and disc contours. An image segmented in this way facilitates the extraction of valuable clinically interpretable features (e.g. CDR and NRR area)[5] and abstract features (e.g. texture and color features). Then a second step works with this segmented data and features to decide whether a patients suffering from glaucoma or not. This work presents improved fundus image preprocessing for feature extraction and AI optimization for glaucoma classification with discussion of the clinical implications and potential future directions.

Methodology

We propose a classification system that detects glaucomatous eye by evaluation of color fundus photograph of retina. In doing so, we are innovatively combining VMD and Convolutional Neural Networks (CNNs). The system comprises several key components:

Data Preprocessing

We propose the following preprocessing pipeline of color fundus images of the retina, firstly it was basic steps like *conversion of images to grayscale* to focus on relevant features and *resizing images to a standardized resolution* to ensure consistency across the dataset. After that - *the application of histogram equalization* to enhance contrast and improve image quality.

The next step is *filtering images by Gaussian filters to reduce noise and enhance edges* by smoothing pixel intensities and emphasizing transitions between regions. Also, *Variational Mode Decomposition (VMD)* as a preprocessing step to enhance the content of the image further. VMD is a data-driven signal processing technique that decomposes a signal into multiple modes, each representing a specific frequency band. By decomposing the fundus images into their constituent modes, we are able to extract relevant features that may not have been captured adequately by CNN alone. VMD Preprocessing Pipeline: Signal Decomposition, Feature Extraction and Integration with CNN (extracted features are integrated into the CNN architecture as additional input channels).

Data Augmentation

The classification system incorporates data augmentation techniques to enhance the diversity and robustness of the training dataset we use the following methods: ***Shear Range***: Shearing is a geometric transformation that displaces one part of an image relative to another along a given axis, resulting in a deformation similar to stretching or skewing. In this system, after a batch of experimentation, a shear range of 0.2 is employed as optimal, indicating that the images can be sheared by a random angle between -0.2

and 0.2 radians. ***Zoom Range***: In this system, a zoom range of 0.2 is utilized, indicating that the images can be zoomed in or out by a factor randomly chosen from the range [1-0.2, 1+0.2]. For example, if the original image size is 100x100 pixels, the zoomed images can have sizes ranging from 80x80 to 120x120 pixels.

Model Architecture

The DL classification model comprises convolutional layers responsible for feature extraction from fundus images, followed by max-pooling layers for spatial down-sampling and dimensionality reduction. The extracted features are then fed into densely connected layers for classification into glaucoma or non-glaucoma categories. Dropout regularization is incorporated to prevent overfitting and enhance model generalization by randomly dropping neurons during training.

Training Strategy

The augmentation described above increases the diversity of the training data and helps the model learn robust features. Additionally, the model is trained using an Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss function. The training process is monitored using metrics as classification accuracy and loss, with early stopping implemented to prevent overfitting. Early stopping involves halting training when the performance on a separate validation dataset fails to improve or starts to degrade, thus ensuring the model generalizes well to unseen data.

Evaluation

The performance of the classification system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. A validation dataset is used to assess the model's performance on unseen data and avoid overfitting. The training and test data were partitioned with a distribution of 70% for training and 30% for testing, ensuring a robust evaluation of the model's performance on unseen data.

Data

The data utilized in this study encompassed a diverse amalgamation of datasets to overcome the scarcity of glaucomatous images in individual repositories. We have 1,545 images of glaucoma eyes (including the suspicious) and 3,609 images of normal eyes (from RIM-ONE, ORIGA, DRISHTI and LAG). Combining these datasets helped us to deal with the significant imbalance in each particular dataset.

Results and Conclusion

Our classification model achieves 99% accuracy on training and 89% on validation (see Fig. 1), as well as 99% and 90% F1-Score on the training and validation, respectively. Also, the training loss of 0.0182 suggests that the model has effectively minimized its objective function.

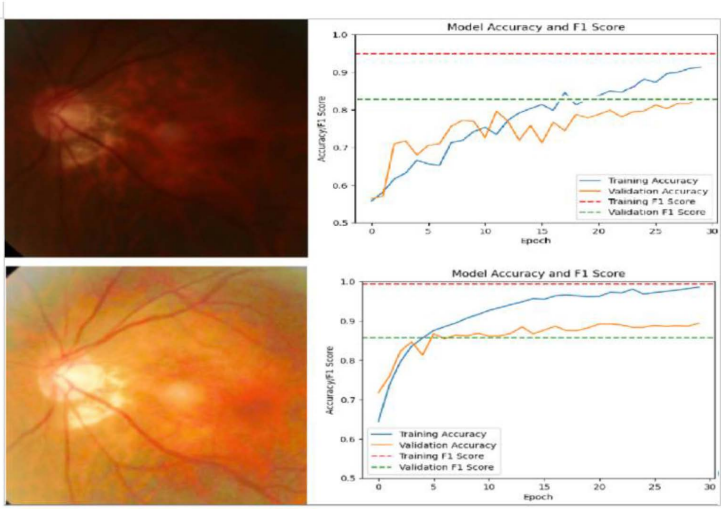


FIGURE 1

1) Example image (basic preprocessing); 2) NN performance with basic preprocessing; 3) Same example image with additional preprocessing; 4) NN performance with additional preprocessing.

In contrast, VGGNet-based models typically achieve 93% validation accuracy, ResNet models 95%, and hybrid models up to 96%. Multi-modal approaches, integrating data from different modalities, report the highest validation accuracy at 97%. Our model demonstrates exceptional performance on training, and its validation accuracy is competitive with potential for further improvement in generalization.

We propose VMD to capture fine-grained details and texture patterns in fundus images and observed improved generalization performance, particularly on validation with unseen fundus images. VMD is helpful in this specific task because the resulting decomposition isolates different frequency components, enhancing specific features that are indicative of glaucoma while suppressing irrelevant details. Overall, the integration of shear and zoom range augmentation parameters enhances the versatility and generalization capabilities of the glaucoma classification system, enabling it to effectively handle variations in image orientation, perspective, and scale commonly encountered in clinical practice.

Our proposed glaucoma classification system demonstrates promising results in accurately detecting glaucomatous features from fundus images. By integrating deep learning models with comprehensive preprocessing techniques and data augmentation strategies, the system offers a robust framework for automated glaucoma diagnosis. Future work may focus on expanding the dataset, refining the architecture, and integrating additional clinical features to further enhance diagnostic accuracy and clinical utility.

References

[1] Faizan Abdullah, Rakhshanda Imtiaz, Hussain Ahmad Madni “A Review on Glaucoma Disease Detection Using Computerized Techniques ” {https://ieeexplore.ieee.org/document/9360745}, In: 2016

[2] Ian J. C. MacCormick, Bryan M. Williams, Yalin Zheng, Kun Li, Baidaa Al-Bander, Silvester Czanner, Rob Cheeseman, Colin E. Willoughby, Emery N. Brown, George L. Spaeth, Gabriela Czanner "Accurate, fast, data efficient and interpretable glaucoma diagnosis with automated spatial analysis of the whole cup to disc profile" {https://doi.org/10.1371/journal.pone.0209409}, In: 2019

[3] Lauren J. Coan, Bryan M. Williams, Silvester Czanner, Gabriela Czanner, Venkatesh Krishna Adithya, Swati Upadhyaya, Ala Alkafri, Colin E. Willoughby Professor and other "Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review" {https://linkinghub.elsevier.com/retrieve/pii/S0039625722001163}, In: 2023

[4] William Paul, Yinzhi Cao „Defending Medical Image Diagnostics against Privacy Attacks using Generative Methods: Application to Retinal Diagnostics" {url: https://arxiv.org/abs/2103.03078}, In: 2021.

[5] Zheheng Jiang, Hossein Rahmani, Plamen Angelov, Ritesh Vyas, Huiyu Zhou, Sue Black, Bryan Williams „Deep orientated distance-transform network for geometric-aware centerline detection" {url: https://doi.org/10.1016/j.patcog.2023.110028}, In: 2023.

Synthetic cerebral blood vessel generator for training anatomically plausible deep learning models

Author

Georgia Kenyon – Australian Institute for Machine Learning, School of Computer and Mathematical Sciences, The University of Adelaide, Australia; South Australian Health and Medical Research Institute, Adelaide, Australia; Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham, UK

Stephan Lau, – Australian Institute for Machine Learning, School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide, SA, Australia

Antonios Perperidis – Australian Institute for Machine Learning, School of Computer and Mathematical Sciences, The University of Adelaide, Australia; Adelaide's Women's and Children's Hospital Network, Australia

Michael Chappell – Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham, UK

Mark Jenkinson – Australian Institute for Machine Learning, School of Computer and Mathematical Sciences, The University of Adelaide, Australia; South Australian Health and Medical Research Institute, Adelaide, Australia

Citation

Kenyon, G., Lau, S., Perperidis, A., Chappell, M., Jenkinson, M. Synthetic cerebral blood vessel generator for training anatomically plausible deep learning models.

Introduction

Blood vessel networks, with their complex geometrical and topological characteristics, play a significant role in diagnosing and understanding

various cerebrovascular diseases. Deep learning (DL) segmentation methods can aid in analysing these structures; however, models often produce anatomically implausible segmentations, overlooked by simple segmentation metrics. Extensive literature on cerebral vessel geometry rules, like branching patterns and vessel length-radius ratios, enable the creation of synthetic vessel label generators that can create data that adhere to or deviate from these rules. This data can be used to train DL networks, that score vessel label's anatomical plausibility and implausibility. Trained networks can then be used to evaluate segmentation networks' label outputs based on their anatomical plausibility, to go beyond commonly used, but mathematically simple, segmentation evaluation metrics. This work presents a novel synthetic cerebral vessel data generator, facilitating the generation of both anatomically plausible and implausible vasculature for the purpose of training DL models to assess the plausibility, or quality, of vessel segmentations in medical imaging.

Method

Our approach to simulating cerebral vessels involves a particle-based model using Python code, where the path of each particle represents a vessel segment. This model is driven by a set of functions, each tailored to replicate specific aspects of cerebral vessel anatomy guided by predefined anatomical rules derived from literature to simulate the branching characteristics of the vessels (Helthuis, 2019). The vessels are initially represented as an evolving 3D skeleton image, which is then transformed into a 3D partial volume image at the desired resolution.

Particle Initialization and Growth Dynamics

The simulation starts by initializing particles, each tracing out the initial vessel segments, with specific attributes. These properties are; the particle's initial velocity and location (location of the Internal Carotid Artery), the length and radius of the vessel (Rai, 2013), and the distance to a branching event, that determines the next branching point. This initialization step acts as a realistic starting point for vessel growth. The iterative particle update function then

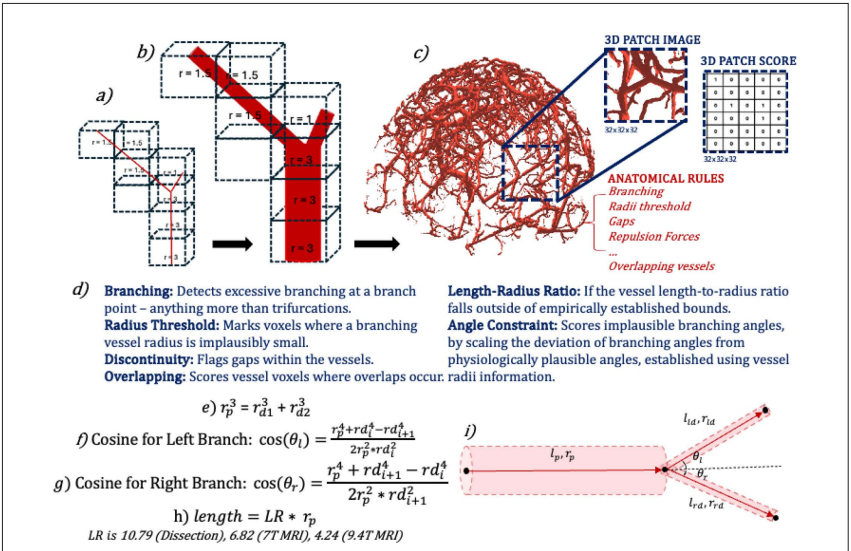


FIGURE 1

Synthetic Vessel Generation (**a**) 3D skeleton image, where each voxel also stores the vessel radius for that location. (**b**) Partial Volume image: Generated from the fractional volume of each voxel occupied by vessels, based on cylindrical segments. (**c**) Full synthetic vessel image d) Implausibility types and scoring method. (**e**) Murray's Law for Bifurcations, (**f**) & (**g**) Bifurcation Angles, (**h**) Vessel length-radius ratio, (**i**) Constrained bifurcation configuration.

acts as the engine of the simulation, iteratively updating the properties of all current particles, which trace out the continuous vessel structures. The 'step-size' of the iterations are reduced as the radii of the vessels reduce, to optimize the speed and resolution of image generation. A vessel length update function adjusts new vessel branch lengths using empirical data to maintain a realistic length-radius ratio (Figure 1h) (Helthuis, 2019).

The vessel growth direction is influenced by repulsive forces from nearby vessels and brain structures. Each iteration uses the current velocity vectors and repulsive forces to move the particles to simulate vessel geometry and the shape of vessels between branch points. The amount of repulsion integrates both vessel overlap and brain edge repulsion forces to ensure that

the simulated vessels do not intersect unnaturally (with set repulsion radii and strengths) and remains within the boundaries of the brain.

Branching Mechanism
Once a branch point is reached, the number of new branches is initialized using set probability functions derived from literature to determine if the branch has a bifurcation, trifurcation or no branching – 70%, 30% and 20% respectively (Al Fauzi, 2021). When a new branch is initiated, the function dynamically adjusts the radii of the daughter branches dependent on the parent and daughter branch characteristics, as per Figure 1e. After generating new vessel radii, the branching angles between daughter vessels are adjusted and constrained, by calculating the vector components of the new branches based on the parent vessel's direction (Figure 1f and g). This obeys Murray's law, as it ensures the conservation of blood flow and energy efficiency in the simulated vessels (CD, 1926).

Partial Volume Rendering
The simulator renders a 3D skeleton image of the vascular network. Each voxel in the vessel image also stores the vessel radius at that point, which is used to linearly interpolate partial volume contributions that were pre-calculated numerically for a range of cylinders of different sizes and locations within the voxel. This ensures that the resultant partial volume image accurately reflects the volumetric contributions of the vessels based on their radius.

Vessel Implausibility
The vessel generator is designed to be adaptable to produce anatomically plausible vessels and implausible vessels, that do not obey, to a controllable degree, the known literature rules- outlined in Figure 1d. The generator then produces a score image, of the same label image dimension to represent the location of the implausibility and the scale of inaccuracy.

Anatomical Quality Assessment Network (AQA)
After generating a dataset of plausible and implausible vascular images and labels, the data is used to train multiple fully convolutional AQA networks.

The networks are trained with patches extracted from synthetic images, and labelled according to their anatomical plausibility (e.g. gaps). The AQA networks can employ varying loss functions to learn to differentiate between plausible and implausible vessel features by using binary decision making (plausible vs implausible) or regression losses to score the level of implausibility.

Results and Discussion

The adaptable vessel generator, with its multiple configurable parameters, can create a wide array of cerebral vessel trees that are anatomically plausible and implausible based on literature guidelines. This data can be used to train a network to detect and quantify anatomical implausibility in vessel images. Results are shown in Table 1, however further work is required with different implausibility types. Trained networks can be used to evaluate the plausibility of vascular segmentations from DL segmentation models, which is the focus of future work.

The integration of a synthetic generation tool with a plausibility assessment model based on anatomical literature offers significant opportunities to enhance the evaluation of segmentation outputs, and is adaptable to alternative anatomy.

TABLE 1: Accuracy of DL models for specific implausibility

Implausibility	Loss Function	Accuracy (%)	Precision (%)	Recall (%)
Gaps	Weighted BCE	92	92	92
Branching Numbers	Weighted BCE	78	77	78

Bibliography

Al Fauzi, A. a. A. Y. K. a. G. R. a. S. N. S., 2021. Neuroangiography patterns and anomalies of middle cerebral artery: A systematic review. *Surgical Neurology International,* Volume 12.

CD, M., 1926. The Physiological Principle of Minimum Work: I. The Vascular System and the Cost of Blood Volume.. *Proceedings of the National Academy of Sciences of the United States of America,* Issue 0027-8424, pp. 207-14.

Helthuis, J. H. a. V. D. T. P. a. H. B. a. B. R. L. a. H. A. A. a. H. J. a. V. d. T. A. a. B. M. a. Z. J. J. a. V. d. Z. A., 2019. Branching pattern of the cerebral arterial tree. *The Anatomical Record,* Volume 302, pp. 1434 -- 1446.

Rai, A. T. a. H. J. P. a. C. B. a. H. G., 2013. Cerebrovascular geometry in the anterior circulation: an analysis of diameter, length and the vessel taper. *Journal of neurointerventional surgery,* Volume 5, pp. 371--375.

Set 2: Low-Quality Medical Images, Pathology, Microscopic, Dental and Bone Imaging

A histology-informed network for white blood cell recognition at subpixel level

Author
Qian Wang – School of Computer Science and Technology, Donghua University, Shanghai, China
Zhao Chen – School of Computer Science and Technology, Donghua University, Shanghai, China; Department of Computer Science, Tissue Image Analytics Centre, University of Warwick, Coventry, UK

Citation
Wang, Q., Chen, Z. A histology-informed network for white blood cell recognition at subpixel level.

Abstract
Recognition of white blood cells (WBCs) in digital slides by deep learning can facilitate blood cancer diagnosis and treatment. Each type of WBC exhibits distinct phenotypes, such as differences in shapes, cellular structures and spectral signatures. Being mixtures of multiple substances, WBCs reflect light consisting of entangled spectral signatures (1–3). If disentangling the spectrum of each pixel of the WBC slides, one could know the substances

in WBCs and the types of WBCs (4–6). Motivated by this hypothesis, we use Microscopic Hyperspectral Imagery (MHSI) comprising tens or hundreds of band images which allow for subpixel spectral disentanglement (unmixing) and identification (7–9). As there is an inevitable tradeoff between the spatial resolution and the spectral resolution (10), we fuse low-spatial-resolution (LR) HSIs and some co-registered high-spatial-resolution (HR) multispectral images (MSIs) with the same view as LR-HSIs to obtain HR-HSIs, providing enhanced details of cellular structures for accurate WBC recognition (11,12). Therefore, we propose a novel model that simultaneously performs hyperspectral image enhancement, subpixel feature representation and white blood cell recognition. The model is named Histology-Informed Network (HINet), as there a variety of histology knowledge considered, including the hyperspectral-biomedical components (HBCs) and their proportions of each pixel of HSIs obtained by spectral disentanglement (13,14), the multispectral counterparts (MBCs and their proportions) from the co-registered MSIs, the spectral correlation in HSIs and the structures of WBCs. The contributions are as follows.1) This work is perhaps the first study of WBCs by deep learning from the hyperspectral subpixel level. 2) Given the histology knowledge, HINet leverages the subpixel features to accurately delineate WBCs and refine pixel-wise classification and semantic segmentation. 3) Instead of fusing synthetized images fully supervised by reference images, it performs subpixel-aligned fusion and enhances HSIs without any reference.

Method

The proposed HINet is illustrated by Figure 1. It consists of three parts, multimodal image fusion (MIF), cross-modality spectral disentanglement (CSD) and subpixel-guided cell recognition (SCR). After being pretrained independently, each module is fine-tuned in a self-supervised fashion: as CSD attains subpixel information to regularize MIF and SCR, it disentangles the HR-HSIs enhanced by MIF and estimates proportions of HBCs in each HSI pixel of the cell regions segmented by SCR. Let X, Y and Z denote LR-HSIs, HR-MSIs and HR-HSIs, respectively. To fuse without reference images, pseudo reference Z′ are generated by an off-line unsupervised

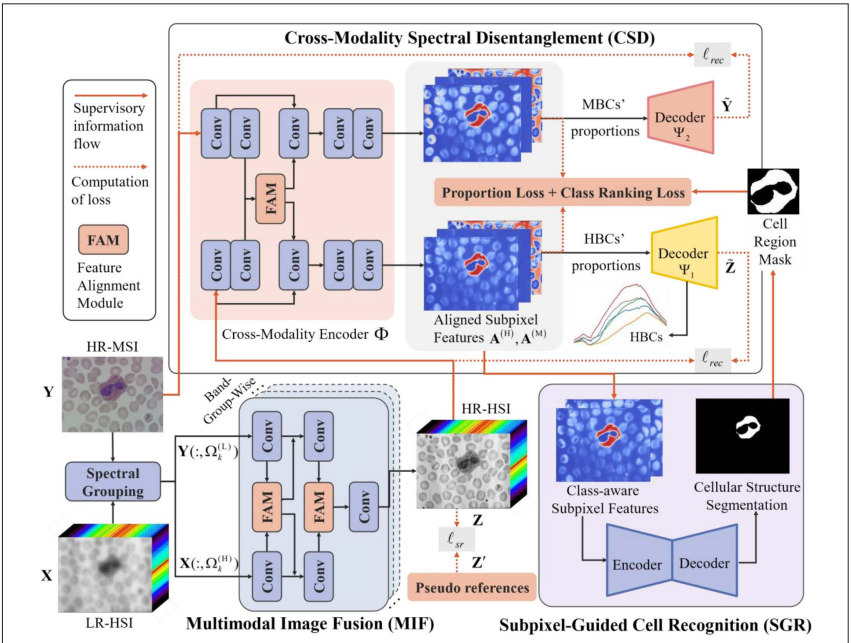


FIGURE 1
The proposed HINet framework.

fusion method (15) to train MIF and reconstruct Z with the super-resolution error ℓ_{sr}. For CSD, the linear mixing model (LMM) (16) is used and optimized via minimization of the sum-to-one constraint ℓ_{asc}, and the image reconstruction loss ℓ_{rec}. Total-Variation (TV) smoothing taking effect by the loss ℓ_{tv} is applied to HBCs/ MBCs to depict the similarity of substances of the neighboring pixels. SCR is optimized by the dice loss ℓ_{dice} for WBC segmentation and the class ranking loss ℓ_{rank} to learn discriminative subpixel features for WBC classification. The total loss is $\ell_{total} = \ell_{sr} + \ell_{rec} + \lambda_{asc}\ell_{asc} + \lambda_{tv}\ell_{tv} + \lambda_{rank}\ell_{rank} + \lambda_{dice}\ell_{dice}$, where the trade-off coefficients λ_{asc} = 0.003, λ_{tv} = 0.001, λ_{rank} = 0.02 and λ_{dice} = 0.5. The final output of HINet is the segmentation and classification mask of each enhanced WBC image, i.e., each set of HR-HSIs.

Results

We performed model evaluation on a public Hyperspectral White Blood Cell Dataset[1]. Each set of LR-HSIs is captured under 100x magnification with 51 spectral bands ranging from 450-780nm at an interval of 6.6nm, while HR-MSIs are the RBG images registered with the 26th band of LR-HSIs.

Three types of WBCs, eosinophil, neutrophil, and lymphocyte, are selected for their unique substance composition. Both HSIs and MSIs are resized to 256×320, while 586 pairs of images are partitioned into training, validation and testing sets by the ratio 6:2:2. The model is trained by AdamW optimizer with the learning rate of 1e-4. Table 1 indicates that HINet outperforms other pixel-level models on every metric for segmentation. Without fine-tuning stage, each task in HINet operates independently rather than in a self-supervised manner, which significantly reduces performance. Moreover, the removal of ranking loss prevents HINet from leveraging cellular information to learn discriminative subpixel features, leading to a decrease in performance. Figure 2 shows that the learned HBCs are similar to but smoother than the initial ones. Specifically, the internal structures of each WBC phenotype learned from HSIs are consistent with those of MSIs, as highlighted in red in Figure 2. (c). It suggests that the histology knowledge of WBCs is well preserved and enhanced, which contributes to accurate WBC recognition.

TABLE 1: Comparison with SOTA segmentation methods and ablated models

Method	OA (%) ↑	DI (%) ↑	JC (%) ↑	HD95↓	ASSD↓
HyperNet (17)	95.12±0.50	92.57±0.61	87.27±0.75	3.55±0.89	1.25±0.34
nnUnet (18)	90.32±0.75	90.43±0.80	83.41±1.00	5.54±2.03	2.90±0.96
ACCUnet (19)	91.53±0.74	91.48±0.53	84.92±0.62	3.28±0.60	1.03±0.27
SwinUNETR (20)	94.44±0.73	93.21±0.78	88.46±1.02	2.49±1.10	0.86±0.50
HINet w/o fine-tuning	94.44±0.24	90.56±0.73	83.26±1.01	4.00±1.13	1.39±0.29
HINet w/o ranking loss	93.99±0.36	93.67±0.59	88.41±0.78	2.58±0.62	0.83±0.25
HINet	**95.20±0.27**	**94.61±0.40**	**89.94±0.62**	**1.89±0.51**	**0.63±0.05**

[1] Hyperspectral White Blood Cell Dataset Homepage, https://bio-hsi.ecnu.edu.cn/

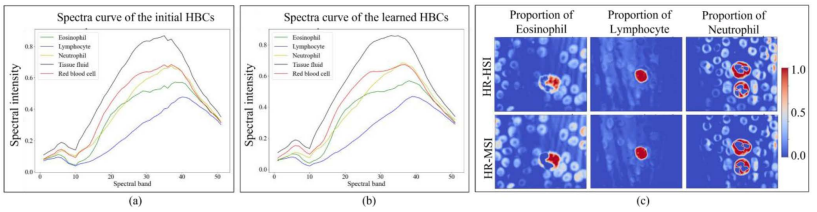


FIGURE 2

Feature visualization of the CSD results. (**a**)-(**b**) plot the initial HBCs and the learned HBCs of the decoder ψ_1 in Figure 1. (**c**) shows multimodal subpixel features of three images, each belonging to the class of eosinophil, lymphocyte and neutrophil.

References

(1) Müller D, Schuhmacher D, Schörner S, Großerueschkamp F, Tischoff I, Tannapfel A, Reinacher-Schick A, Gerwert K, Mosig A. Dimensionality reduction for deep learning in infrared microscopy: a comparative computational survey. *Analyst* (2023) 148:5022–5032. doi: 10.1039/D3AN00166K

(2) Li W, Wang L, Luo C, Zhu Z, Ji J, Pang L, Huang Q. Characteristic of Five Subpopulation Leukocytes in Single-Cell Levels Based on Partial Principal Component Analysis Coupled with Raman Spectroscopy. *Appl Spectrosc* (2020) 74:1463–1472. doi: 10.1177/0003702820938069

(3) Duan Y, Wang J, Hu M, Zhou M, Li Q, Sun L, Qiu S, Wang Y. Leukocyte classification based on spatial and spectral features of microscopic hyperspectral images. *Opt Laser Technol* (2019) 112:530–538. doi: https://doi.org/10.1016/j.optlastec.2018.11.057

(4) Banik PP, Saha R, Kim K-D. An Automatic Nucleus Segmentation and CNN Model based Classification Method of White Blood Cell. *Expert Syst Appl* (2020) 149:113211. doi: https://doi.org/10.1016/j.eswa.2020.113211

(5) Martin FL, Kelly JG, Llabjani V, Martin-Hirsch PL, Patel II, Trevisan J, Fullwood NJ, Walsh MJ. Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nat Protoc* (2010) 5:1748–1760.

(6) Paraskevaidi M, Matthew BJ, Holly BJ, Hugh BJ, Thulya CP V, Loren C, StJohn C, Peter G, Callum G, Sergei KG. Clinical applications of infrared and Raman spectroscopy in the fields of cancer and infectious diseases. *Appl Spectrosc Rev* (2021) 56:804–868.

(7) Ortega S, Halicek M, Fabelo H, Callico GM, Fei B. Hyperspectral and multispectral imaging in digital and computational pathology: a systematic review. *Biomed Opt Express* (2020) 11:3195–3233.

(8) Signoroni A, Savardi M, Baronio A, Benini S. Deep learning meets hyperspectral image analysis: A multidisciplinary review. *J Imaging* (2019) 5:52.

(9) Kumar N, Uppala P, Duddu K, Sreedhar H, Varma V, Guzman G, Walsh M, Sethi A. Hyperspectral tissue image segmentation using semi-supervised NMF and hierarchical clustering. *IEEE Trans Med Imaging* (2018) 38:1304–1313.

(10) Veganzones MA, Simões M, Licciardi G, Yokoya N, Bioucas-Dias JM, Chanussot J. Hyperspectral Super-Resolution of Locally Low Rank Images From Complementary Multisource Data. *IEEE Transactions on Image Processing* (2016) 25:274–288. doi: 10.1109/TIP.2015.2496263

(11) Dian R, Li S, Kang X. Regularizing Hyperspectral and Multispectral Image Fusion by CNN Denoiser. *IEEE Trans Neural Netw Learn Syst* (2021) 32:1124–1135. doi: 10.1109/TNNLS.2020.2980398

(12) Ma Q, Jiang J, Liu X, Ma J. Multi-Task Interaction Learning for Spatiospectral Image Super-Resolution. *IEEE Transactions on Image Processing* (2022) 31:2950–2961. doi: 10.1109/TIP.2022.3161834

(13) Lu G, Qin X, Wang D, Chen ZG, Fei B. Estimation of tissue optical parameters with hyperspectral imaging and spectral unmixing. *Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging*. International Society for Optics and Photonics (2015). p. 94170Q

(14) ul Rehman A, Qureshi SA. A review of the medical hyperspectral imaging systems and unmixing algorithms' in biological tissues. *Photodiagnosis Photodyn Ther* (2021) 33:102165.

(15) Chen Z, Pu H, Wang B, Jiang G-M. Fusion of Hyperspectral and Multispectral Images: A Novel Framework Based on Generalization of Pan-Sharpening Methods. *IEEE Geoscience and Remote Sensing Letters* (2014) 11:1418–1422. doi: 10.1109/LGRS.2013.2294476

(16) Borsoi RA, Imbiriba T, Bermudez JCM, Richard C, Chanussot J, Drumetz L, Tourneret J-Y, Zare A, Jutten C. Spectral Variability in Hyperspectral Data Unmixing: A comprehensive review. *IEEE Geosci Remote Sens Mag* (2021) 9:223–270. doi: 10.1109/MGRS.2021.3071158

(17) Wang Q, Sun L, Wang Y, Zhou M, Hu M, Chen J, Wen Y, Li Q. Identification of Melanoma From Hyperspectral Pathology Image Using 3D Convolutional Networks. *IEEE Trans Med Imaging* (2021) 40:218–227. doi: 10.1109/TMI.2020.3024923

(18) Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* (2021) 18:203–211. doi: 10.1038/s41592-020-01008-z

(19) Ibtehaz Nabil and Kihara D. ACC-UNet: A Completely Convolutional UNet Model for the 2020s. In: Greenspan Hayit and Madabhushi A and MP and SS and DJ and S-MT and TR, editor. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland (2023). p. 692–702

(20) He Y, Nath V, Yang D, Tang Y, Myronenko A, Xu D. SwinUNETR-V2: Stronger Swin Transformers with Stagewise Convolutions for 3D Medical Image Segmentation. In: Greenspan H, Madabhushi A, Mousavi P, Salcudean S, Duncan J, Syeda-Mahmood T, Taylor R, editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Cham: Springer Nature Switzerland (2023). p. 416–426

A novel method of determining Bone Mineral Density from pre-surgical CT scans to aid in surgical planning

Author
Niall C. Maguire, Alan D. Brett – Stryker, Imorphics Clinical Trials Group, Manchester, UK

Citation
Maguire, N.C., Brett, A.D. A novel method of determining Bone Mineral Density from pre-surgical CT scans to aid in surgical planning.

Introduction

Patients with a reduced Bone Mineral Density (BMD) may have worse outcomes after joint replacement surgery. Concern surrounds the use of cementless implants, which rely on press-fit fixation in the surrounding bone. Where BMD is reduced, fixation is greatly enhanced by the addition of cement. Cementless implants are often only indicated for younger patients with higher BMD. Assessment of BMD is not commonly performed prior to surgery. Quantitative CT (QCT) scans may be used to measure BMD by calibrating Hounsfield Units (HU) to BMD values, but robotic surgery planning CT scans do not include BMD calibration phantoms. Here we describe automated BMD calibration based on air, fat and the aluminium motion-detection rod included in Mako robotic knee arthroplasty CT images.

Method

Image voxels containing air were identified using histogram analysis. The rod surface was identified using Active Appearance Model (AAM) search and eroded by a 4mm radius to exclude partial voxels and beam hardening. Median voxel intensity was used as the rod HU value. Normals to the AAM segmented femur bone surface were sampled until a discontinuity was detected representing a soft-tissue boundary to air, bone or cartilage. A two-gaussian mixture was used to model the histograms of fat and muscle tissue along this normal path. Mean of the lower gaussian peak was reported as fat HU value. 133 CT knee arthroplasty planning images contained a commercial BMD phantom (Mindways, TX) used to determine the average BMD equivalent values of air, fat and rod as calibration standards. We used 10 iterations of 10-fold cross-validation to compare calibration from the phantom or air-fat-rod standards in trabecular bone at femur and tibia.

Results

An example of rod and fat identification is shown in Figure 1. Trabecular bone BMD from phantom calibration were highly correlated with BMD values calculated using the air-fat-rod standards for both femur and tibia ($r \approx 0.98$).

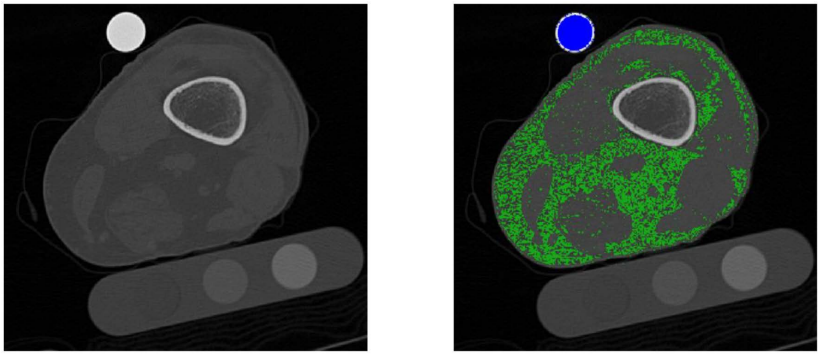


FIGURE 1
Left: axial slice through the femur with rod above and BMD phantom below. Right: after detection, rod is labelled blue and fat is labelled green.

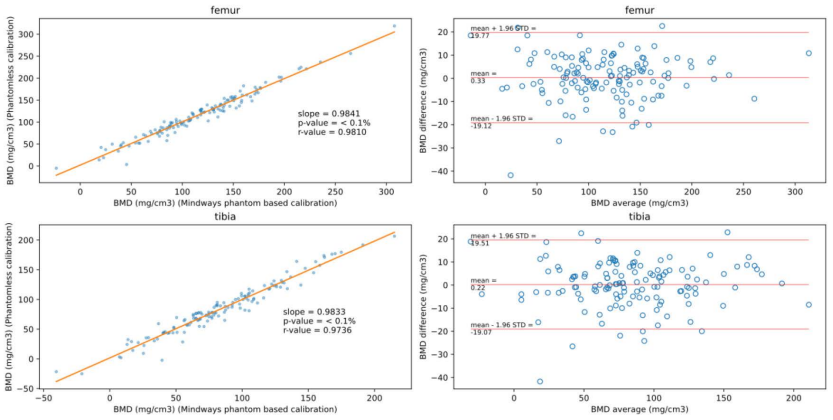


FIGURE 2
Comparison of trabecular bone BMD from phantom calibration and air/fat/rod calibration. Left: correlation of femur (top) and tibia (bottom). Right: Bland-Altman plots of agreement for femur and tibia.

Bland-Altman analysis showed negligible bias of ~0.3mg/cm3 (Figure 2). The SD of differences is 9.9mg/cm3 which suggests an upper bound of precision of 9mg/cm3 (precision error for Mindways phantom is 3.6mg/cm3), which produces a CoV=7.2% at mean BMD=125mg/cm3.

Conclusion

Accurate volumetric BMD values may be derived at the knee from CT images acquired for robotic arthroplasty planning. This information may be useful in informing surgical decision about the use of cementless implants, particularly in postmenopausal women or patients that have received androgen blocking therapy as part of treatment for cancer, who may be at risk of osteoporosis.

Applying likelihood-based out-of-distribution detection to malaria microscopy using Deep Diffusion Models

Author

Joseph Goodier – Dept. of Computer Science, University of Bath, Bath, United Kingdom
Richard Bowman – School of Physics & Astronomy, University of Glasgow, Glasgow, United Kingdom
Pietro Cicuta – Dept. of Physics, University of Cambridge, Cambridge, United Kingdom
Joe Knapper – School of Physics & Astronomy, University of Glasgow, Glasgow, United Kingdom
Samuel McDermott – Dept. of Physics, University of Cambridge, Cambridge, United Kingdom
Joram Mduda – Ifakara Health Institute, Ifakara, Tanzania
Catherine Mkindi – Ifakara Health Institute, Ifakara, Tanzania
Joel Collins – Dept. of Computer Science, University of Bath, Bath, United Kingdom
Julian Stirling –
William Wadsworth – Dept. of Computer Science, University of Bath, Bath, United Kingdom
Boyko Vodenicharski – Dept. of Computer Science, University of Bath, Bath, United Kingdom
Jessica Nicholson – Dept. of Computer Science, University of Bath, Bath, United Kingdom
Neill Campbell – Dept. of Computer Science, University of Bath, Bath, United Kingdom

Citation
Goodier, J., Bowman, R.,Cicuta, P., Knapper, J., McDermott, S., Mduda, J., Mkindi, C., Collins, J., Stirling, J., Wadsworth, W., Vodenicharski, B., Nicholson, J., Campbell, N. Applying likelihood-based out-of-distribution detection to malaria microscopy using Deep Diffusion Models.

Abstract
Malaria is a serious febrile illness affecting nearly a quarter of a billion people per year, and responsible for half a million deaths. The gold standard for malaria diagnosis is microscopic examination of blood films. Poor slide quality and suboptimal imaging can lead to misdiagnosis and are common problems when collecting training data for diagnostics. We propose leveraging Out-of-Distribution detection to improve diagnostics by employing probabilistic generative models to detect deviations from healthy samples. We use a class-conditioned diffusion model to detect potentially suboptimal and pathological images in a Giemsa-stained, thin-film microscopy dataset. This is achieved by using a Deep Denoising Diffusion Model to build a Diffusion Classifier model. Our results demonstrate the effectiveness of this approach, offering a promising avenue for enhancing malaria detection and triaging care in resource-limited settings.

Introduction
Efficient diagnosis of malarial disease from the various *Plasmodium* parasite species is critical in triaging care. Accurately detecting malaria in clinics in resource-limited settings can be challenging, as the disease burden is borne predominantly by sub-Saharan African countries. Diagnostics involve microscopy, specifically thick and thin film blood smear examinations. However, the quality of these microscopy slides are often suboptimal for diagnostics due to a number of possible preparation factors (over-staining, prepared with contaminated water, unclean slides) or suboptimal imaging conditions (microscope not in focus). This context presents an opportunity for Out-of-Distribution (OOD) detection, where a probabilistic generative model can be trained to learn a distribution over the healthy data, which can

then be used to detect samples that are not from the healthy distribution. Furthermore, if the generative model is conditional, and has some knowledge of the types of suboptimal microscopic images, a generative classifier can be used to identify why an image is suboptimal. This ensures that malaria can be confidently identified. This process could streamline detection by: (1) examining anything that appears OOD to the healthy class, (2) checking if the sample is OOD to the healthy class because of suboptimal imaging (3) and identifying if the sample is OOD for the healthy and suboptimal image classes, which implies there is likely pathology present.

In this paper, we apply a conditional Denoising Diffusion Probabilistic Model (DDPM) to detect suboptimal and pathological thin-film microscopy images using likelihood-based OOD detection, to facilitate malaria diagnostics in resource-limited settings.

Methodology

We use thin-film Giemsa-stained microscopy images from the OFM dataset. This dataset is curated from samples collected from clinics across Tanzania using low-cost, Open Flexure 3D-printed microscopes Collins et al. (2020). The dataset consists of both optimal and suboptimal microscopy images. Images are deemed optimal if they can be used for malaria microscopy to diagnose malaria and the amount of parasitemia. Images are deemed suboptimal if they are of insufficient quality to be used for diagnostics. We assigned suboptimal images to Contaminated, Overstained, and Out-of-Focus classes. We also curated a withheld set of 29 cropped images, from a malaria-positive patient, containing cells with inclusions that appeared to be well-focused and reasonably stained. We refer to them as pathological images.

We use the pixel-space DDPM, **simple diffusion** Hoogeboom et al. (2023). The **simple diffusion** model uses the **v**-parameter as an objective, where $\mathbf{v}_{\alpha_t} \epsilon - \sigma_t \mathbf{x}$. Here, ϵ is the sample of isotropic Gaussian noise and $\alpha_t, \sigma_t \in (0,1)$ are parameters that control how much noise is added to the sample as a

function of the time variable t, where $0 \leq t \leq T$ Salimans and Ho (2022). The **x** term is the original sample. The **simple diffusion** Evidence Lower-Bound (ELBO) is,

$$L_\theta = \frac{1}{T}\sum_{t=0}^{T}\left[w(t)|v\hat{}_{\theta,t} - v|^2\right],$$

where $\epsilon\hat{}_{\theta,t}$ is an estimate of the noise distribution from the network, the ELBO weighting is $w(t)=\frac{-\mathrm{d}}{\mathrm{d}t}\log \mathrm{SNR}(t)$ and . Our Diffusion Classifier is based on the method developed by Li et al. (2023). We train a class-conditioned DDPM using classifier-free guidance Ho and Salimans (2022). We calculate the posterior probability over the different classes of our class-conditioned DDPM,

$$p_\theta(c_i|x) = \frac{\exp\{-E_{t,v}\left[|v\hat{}_{\theta,t}(c_i) - v|^2\right]\}}{\sum_j exp\{-E_{t,v}\left[|v\hat{}_{\theta,t}(c_j) - v|^2\right]\}},$$

where **c** are the image classes. In place of the ELBO, we use the DDPM-based likelihood ratio CCLR, proposed by Goodier and Campbell (2023),

$$\mathrm{CCLR}_k = L_\theta - L_\theta^k,$$

where $0 < k < T$. For all experiments, we set $k/T = 0.1$. This stabilises the stochasticity of the ELBO for OOD detection and class prediction.

We run three experiments. **First**, we test if a sample is OOD with respect to the healthy class, by calculating the ROCAUC between the healthy and the various suboptimal test classes. **Second**, we test the Diffusion Classifier and calculate the classification performance on the test classes. We use a Resnet50 He et al. (2016) classifier as a baseline. **Third**, we test the withheld pathological image class by calculating the ROCAUC between the pathological image class and all test classes.

For the **first** experiment, Fig. 1 Left displays the histogram of the CCLR score for each test class when conditioned on the healthy image label. Strong class

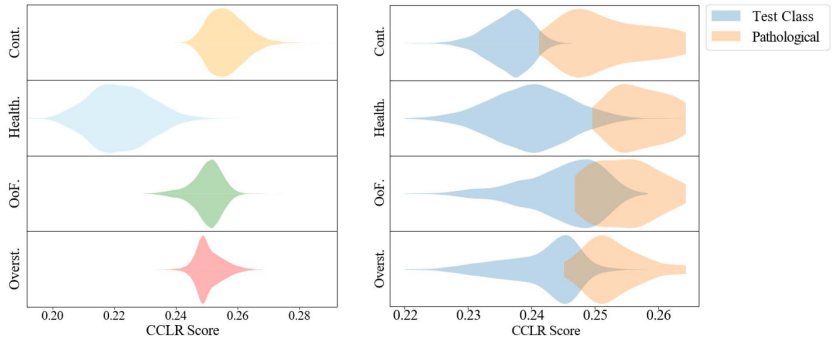


FIGURE 1

Left: Violin plots for the thin-film dataset displaying histograms of CCLR score for each test class conditioned on the optimal label. The ROCAUC scores between the optimal test set and the suboptimal image classes are: Cont. = 0.998, OoF = 0.993, Overst. = 0.996. **Right:** Violin plots for the thin-film dataset displaying histograms of the CCLR score between each test class and the withheld pathological images. The ROCAUC scores are: Cont. = 0.998, Health. = 0.990, OoF = 0.899, Overst. = 0.968.

separation is demonstrated visually and quantitatively with the ROCAUC scores between optimal and suboptimal image class pairs. For the **second** experiment, Tab. 1 shows that there is varying classification performance among classes for both the Diffusion Classifier and Resnet50: all suboptimal classes performed strongly, Healthy class accuracy was the weakest and Overstained class the most often misclassified too. Our Diffusion Classifier performs only marginally worse than Resnet50 at classifying optimal and suboptimal image classes. For the **third** experiment, the violin plots in Fig. 1 Right demonstrate that pathological images show clear class separation from each test class, despite the limited number of samples. Indicating that our model can effectively detect all optimal, suboptimal and pathological image types using OOD detection. These results demonstrate the potential for likelihood-based OOD to automate suboptimal image detection and flag pathology, in the domain of malaria microscopy.

TABLE 1: (a) A confusion matrix displaying the results of the Diffusion Classifier across all classes (Contaminated, Healthy, Out-of-Focus, and Overstained) on the test set of the thin-film dataset. The performance across all classes was: Accuracy = 0.965, Precision = 0.966 and Recall = 0.965. (b) A confusion matrix displaying the results of Resnet50 across all classes (Contaminated, Healthy, Out-of-Focus, and Overstained) on the test set of the thin-film dataset. The performance across all classes was: Accuracy = 0.977, Precision = 0.978 and Recall = 0.977

	Cont.	Health.	OoF	Overst.
Cont.	0.979	0.002	0.003	0.015
Health.	0.017	0.951	0.001	0.031
OoF.	0.002	0.003	0.959	0.036
Overst.	0.017	0.001	0.008	0.974

	Cont.	Health.	OoF	Overst.
Cont.	0.983	0.012	0.000	0.000
Health.	0.002	0.944	0.007	0.047
OoF.	0.000	0.000	0.995	0.005
Overst.	0.001	0.010	0.000	0.990

Acknowledgements

This work was supported UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent AI (grant number EP/S023437/1), the UKRI CAMERA project (EP/T022523/1), EPSRC Global Challenges Research Fund (EP/R013969/1) and EPSRC (EP/R011443/1).

References

Collins, J. T., Knapper, J., Stirling, J., Mduda, J., Mkindi, C., Mayagaya, V., et al. (2020). Robotic microscopy for everyone: the openflexure microscope. *Biomedical Optics Express* 11, 2447–2460

Goodier, J. and Campbell, N. D. (2023). Likelihood-based out-of-distribution detection with denoising diffusion probabilistic models. *arXiv preprint arXiv:2310.17432*

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*

Hoogeboom, E., Heek, J., and Salimans, T. (2023). simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. (2023). Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*

Salimans, T. and Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*

Enhancing mitotic figure detection using attention modules in digital pathology

Author

May Hlaing Kyi – University of West London, School of Computing and Engineering, London, United Kingdom

Massoud Zolgharni – University of West London, School of Computing and Engineering, London, United Kingdom

Syed Ali Khurram – University of Sheffield, School of Clinical Dentistry, Sheffield, United Kingdom

Neda Azarmehr – University of West London, School of Computing and Engineering, London, United Kingdom

Citation

Kyi, M.H., Zolgharni, M., Khurram, S.A., Azarmehr, N. Enhancing mitotic figure detection using attention modules in digital pathology.

Abstract

Mitotic figure counting is crucial in cancer grading and prognosis. However, manual counting is tedious and time-consuming. The diverse appearances of mitoses lead to considerable discordance among pathologists. Developing an automated detection model is challenging due to complex growth patterns and similarities with non-mitotic cells. This work utilizes a detection and classification task (mitosis versus mimics) based on the RetinaNet model. To improve performance in capturing distant feature dependencies, we investigated state-of-the-art attention modules like the Convolutional Block Attention Module (CBAM). These modules are integrated into the ResNet

backbone of RetinaNet. By focusing on essential features, this approach aims to improve mitosis detection and classification. All models were evaluated using the public Canine Mammary Carcinoma (ODAEL) dataset. Results reveal that models with a ResNet50 backbone incorporating CBAM and Squeeze-and-Excitation achieved F1 scores of 0.793 and 0.784, respectively, outperforming standard RetinaNet.

Introduction

Mitosis, a cell division process, has a direct connection with tumour prognosis. Abnormal mitosis and its increase indicate genetic damage in tumours and a loss of controlled proliferation (1). Counting mitoses is crucial for assessing the severity and predicting the outcome of various cancers on digital Whole Slide Images (WSIs). However, manual detection of mitotic figures is time-consuming, requires specialised expertise, and is subject to variability between pathologists. Convolutional neural networks (CNNs) have been employed to automate mitosis detection to address these limitations. However, despite recent efforts to improve this automation, the results remain inadequate for clinical applications (2). This study aims to investigate the potential of incorporating various state-of-the-art attention mechanisms into a ResNet backbone to improve the performance of mitotic figure detection, particularly in distinguishing mitotic from mimics in WSIs.

Methods and Dataset

This study employs a detection and classification task based on the RetinaNet (3), inspired by (4) with ResNet50 (5) backbone. In CNNs, the layers near the input typically focus on capturing local features but may struggle with emphasising and accurately locating more complex and abstract features. To address this, we integrated into ResNet50 backbone state-of-the-art attention mechanisms including CBAM (6), Squeeze-and-Excitation (SE) (7), Enhanced Channel Attention (ECA) (8), and Hybrid attention module (HAM) (9). Figure 1 provides an overview of the framework.

We used the ODAEL dataset (10) with 14,151 mitotic figures and 36,135 look-alikes, split into 70% training (15 WSI) and 30% testing (6 WSI). For each

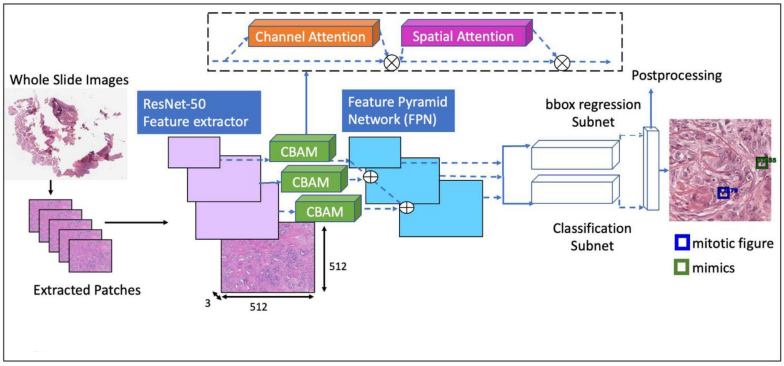


FIGURE 1
The mitotic detection framework.

epoch, 25,500 training and 4,500 validation patches (512×512) were randomly extracted from the WSIs, continuing for 100 epochs using the Adam optimizer with a learning rate of 0.0001. To prevent overfitting, an early stopping method was used.

Experiment Results and Discussion

The models' performance was evaluated on the test dataset using a sliding window with a tile size of 512×512 pixels (stride of 51×51 pixels). Non-maximum suppression technique was used to retrieve the detected mitoses that surpassed the confidence level of 0.4. The detected mitoses are compared against the ground truth using the K-dimensional tree algorithm. Table 1 presents an overview of the comparative analysis and as can be seen the model integrated with the CBAM module achieved a higher F1-score (0.793 compared to 0.775- details Table 1). You can find a visual representation of the automatically detected mitoses in Figure 2.

TABLE 1: Comparison of experimental results

Models (ResNet50 backbone)	Precision	Recall	TP	FP	FN	F1
RetinaNet	0.802	0.751	2238	553	744	0.775
CBAM	0.792	0.794	2368	621	614	**0.793**
SE	0.773	0.796	2374	699	608	**0.784**
ECA	0.776	0.771	2298	664	684	0.773
HAM	0.765	0.782	2333	715	649	0.774

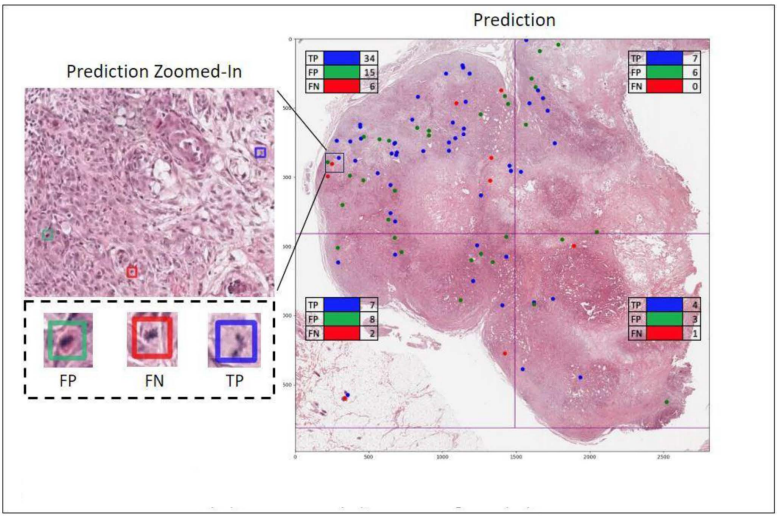


FIGURE 2
Displays the prediction overlay on the WSI level. The numbers at each quarter represent the counts of True Positives (TP), False Positives (FP), and False Negatives (FN).

Conclusion

Initial results indicate that CBAM attention outperforms RetinaNet in detecting mitosis with a F1 score of 0.793. Given the insufficient clinical applicability of this score, future work will focus on improving accuracy.

Acknowledgements

MHK is supported by the Vice Chancellor's Scholarship at the University of West London.

References

(1) Chauhan I, Wonhyeong K, Trisal M. Mitotic Index and its Role in Squamous Cell Carcinoma of Oral Cavity. Recent Advances in Pathology and Laboratory Medicine (ISSN: 2454-8642). (2021) 7(3&4):19-22.

(2) Pan X, Lu Y, Lan R, Liu Z, Qin Z, Wang H, Liu Z. Mitosis detection techniques in H&E stained breast cancer pathological images: A comprehensive review. Computers & Electrical Engineering. (2021) 91:107038.

(3) Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. InProceedings of the IEEE international conference on computer vision (2017). p. 2980-2988.

(4) Aubreville M, Wilm F, Stathonikos N, Breininger K, Donovan TA, Jabari S, Veta M, Ganz J, Ammeling J, van Diest PJ, Klopfleisch R. A comprehensive multi-domain dataset for mitotic figure detection. Scientific Data. (2023)10(1):484.

(5) He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition (2016). p. 770-778.

(6) Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. InProceedings of the European conference on computer vision (ECCV) (2018). p. 3-19.

(7) Hu J, Shen L, Sun G. Squeeze-and-excitation networks. InProceedings of the IEEE conference on computer vision and pattern recognition (2018). p. 7132-7141.

(8) Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020). p. 11534-11542.

(9) Li G, Fang Q, Zha L, Gao X, Zheng N. HAM: Hybrid attention module in deep convolutional neural networks for image classification. Pattern Recognition. (2022) 129:108785.

(10) Aubreville M, Bertram CA, Donovan TA, Marzahl C, Maier A, Klopfleisch R. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. Scientific data. (2020) 7(1):417.

FLy-HEi nuclear distribution clusters associate with clinical features in Follicular Lymphoma

Author

Volodymyr Chapman – Faculty of Biological Sciences, University of Leeds, Leeds, United Kingdom; Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom

Alireza Behzadnia – The Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom

Cathy Burton – The Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom

Dan Painter – Department of Health Sciences, University of York, York, United Kingdom

Alex Smith – Department of Health Sciences, University of York, York, United Kingdom

Reuben Tooze – Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom; The Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom

Andrew Janowzcyk – Wallace H. Coulter Department of Biomedical Engineering, Emory University Atlanta, Georgia, United States of America; Department of Oncology, Geneva University Hospitals, Geneva, Switzerland

David Westhead – Faculty of Biological Sciences, University of Leeds, Leeds, United Kingdom; Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom

Citation

Chapman, V., Behzadnia, A., Burton, C., Painter, D., Smith, A., Tooze, R., Janowzcyk, A., Westhead, D. FLy-HEi nuclear distribution clusters associate with clinical eatures in Follicular Lymphoma.

Introduction

Follicular Lymphoma (FL) is an incurable, common hematological cancer with over 2,000 diagnoses per year in the UK [1]. Grouping of patients

into clinically useful clusters has proven difficult [2]. Though pathologist assessment is routine in diagnosis of FL, the resulting reports provide little value over clinical features such as stage, performance status and presence of B-symptoms in risk prediction and treatment choice.

Given widespread digitisation of pathology slides, there is now scope for analyses otherwise highly impractical for pathologists, such as comparison of all nuclei in a biopsy. Recent immunofluorescence work [3] has focused on individual lymphocyte subpopulations, including regulatory T-cells. Results evidenced better survival in patients with high cellular diversity (hazard ratios high vs rest: 0.22). This resolution of lymphocyte subgrouping is not possible in routine Hematoxylin & Eosin (H&E) biopsy slides while immunofluorescence studies are too costly to become routine. We present a patient subgrouping method that aims to capture cellular profiles in FL within routine H&E – the Follicular Lymphoma H&E index (FLy-HEi). In this, patients were grouped into 3 clusters on cell distribution similarity. Though there were no survival differences between clusters, associations existed between clinical features, including blood albumin level, bone marrow involvement and Ann Arbor stage.

Materials and Methods

N=286 cases of the Hematological Malignancy Research Network FL dataset [4], with corresponding clinical data were used. One formalin fixed paraffin embedded (FFPE) Hematoxylin & Eosin (H&E) biopsy image was available per patient, scanned at 40x magnification (resolution of 0.253μM per pixel).

Non-overlapping tissue patches of size 1000x1000 pixels were extracted from slides. These were filtered for patches predicted to be within naive B-cell regions using a UNet model, trained on 9 FL slides exhaustively-annotated using Quick Annotator [5]. Between 20 and 145 patches were used per slide, depending on tissue availability. The MoNuSaC weight set of HoVerNet [6] was used to segment unlabelled nuclei in patches and nuclear features describing morphology, texture, and intensity were measured. Z-scaled features were sampled using k-means clustering to balance

sampling across the feature space. In total, 43,200 nuclei from a sample of non-Hodgkin's Lymphoma slides (Burkitt, Diffuse Large B-Cell and Follicular Lymphomas) were used to create a non-Hodgkin's Lymphoma 1-dimensional UMAP embedding. Likely cell identities were informed by a pathologist.

Pairwise comparison was conducted on nuclear distributions between all patients using the Anderson-Darling statistic as a distance metric. Agglomerative clustering grouped patients by nuclear distribution group, termed FLy-HEi clusters, with optimal cluster number determined using maximum Calinski-Harabasz metric score. Cox Proportional Hazard model hazard ratios were used to investigate overall survival while Pearson correlation was used to compare clinical features between clusters. Welch's t was used for calculation of p-values.

Results

The 1D UMAP embedding separated visually distinct nuclei, from large, light nuclei with multiple nucleoli (centroblasts) on the negative pole to small, dark nuclei (mature B- & T-lymphocytes and centrocytes) on the positive side (Figure 1A). Patients clustered into 3 nuclear distribution groups. Cluster 1 (n = 190) was enriched for centrocytes. Cluster 2 (n = 45) presented with a centrocyte/immunoblast phenotype while cluster 3 (n = 51) presented with a centroblast/immunoblast phenotype, with near equal proportions of each (Figure 1B & C).

There were no overall survival differences between clusters (Cox proportional hazard ratios = 0.98-1.00) but comparison with clinical features evidenced differences in clinical presentation of clusters (Figure 2). Cluster 1 patients associated with lower ECOG ($p<0.01$) and stage ($p=0.02$). Cluster 2 associated with blood-related features, namely bone marrow involvement, higher blood hemaglobin and albumin (all $p = 0.01$) and normal LDH ($p = 0.03$). Cluster 3 appeared to associate with physical symptoms, including night sweats and weight loss, though these were not significant ($p = 0.05$-0.06), in addition to lower hemaglobin ($p =0.03$) and albumin ($p < 0.01$).

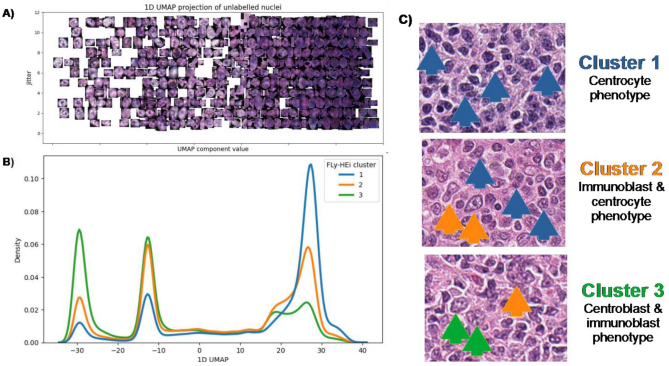


FIGURE 1

Nuclear distribution clusters in Follicular Lymphoma. A) Images of non-Hodgkin's lymphoma nuclei across the 1-dimensional UMAP space. B) Mean density of nuclei for each FLy-HEi cluster C) B-cell regions at 40x magnification, displaying characteristic nuclei of each cluster. Arrow key: Blue = centrocyte, Orange = immunoblast, Green = centroblast.

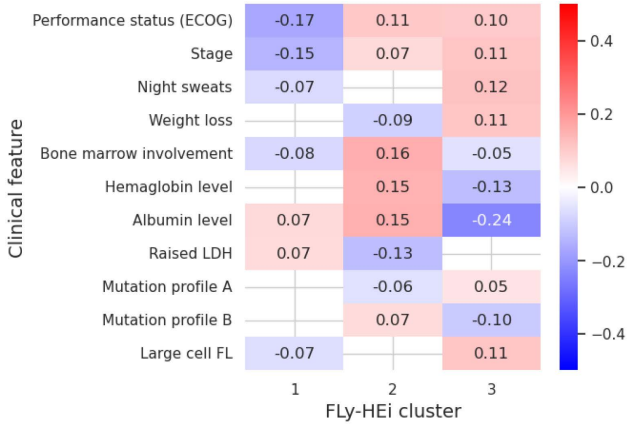


FIGURE 2

Pearson correlation between FLy-HEi clusters and clinical features. Data gathered as reported in Smith et al. [4]. Mutation profiles relate to Crouch et al. BIC mutation groups [2].

Discussion

Previous immunofluorescence studies reported predictive utility of cellular variation in FL [3]. The resolution of H&E is far lower than that achieved in immunofluorescence analyses, where individual lymphocyte subtypes such as $CD4^{+}FoxP3^{+}$ Regulatory T-cells have been distinguished. This study aimed to distinguish distinct cellular phenotypes using standard H&E biopsy slides and explore potential clinical utility of such information.

Patients clustered into distinct cellular groups, with clear differences in nuclear densities at 4 key points (Figure 1B), relating to centroblasts, immunoblasts, centrocytes and mature lymphocytes. Since distributions only differed at these 4 cell types, a more computationally efficient approach may use a narrower scope of cell types to achieve comparable patient clustering. The clinical utility of these groups in risk stratification appears low, with no predictive value in overall survival. Yet, the FLy-HEi clusters associated with patterns of clinical features, with significant differences in some. These distribution clusters may therefore provide an insight into specific pathologies and patterns of pathogenesis in follicular lymphoma. Finally, multiple testing correction was not applied, so there is a risk in the presented work of spurious correlations. The presented work represents a preliminary study into objective, large-scale analyses of H&E cellular phenotypes in FL. Further external validation studies can focus on relationships with the clinical features identified, including potential differences in pathogenesis, and optimal treatment between the FLy-HEi clusters.

Acknowledgements

VC was supported by an EPSRC doctoral training studentship. HMRN dataset retrieval was funded by Blood Cancer UK, formerly Bloodwise (Grant ID 15037). Retrieval of digital pathology data was funded by a Wellcome Trust Institutional Partnership Award.

References

[1] HMRN – Incidence, https://hmrn.org/statistics/incidence, last accessed 2024/05/14

[2] Crouch, S. et al. Molecular subclusters of follicular lymphoma: a report from the United Kingdom's Haematological Malignancy Research Network. *Blood Advances* 6, 5716-5731. (2022)

[3] Tsakiroglou, A.M. et al. Immune infiltrate diversity confers a good prognosis in follicular lymphoma.*Cancer Immunol Immunother,* 70, 3573-3585. (2021)

[4] Smith, A. et al. Cohort Profile: The Haematological Malignancy Research Network (HMRN): a UK population-based patient cohort. *International Journal of Epidemiology,* 47, 700-700g. (2018)

[5] Miao, R., Toth, R., Zhou, Y., Madabhushi, A. & Janowczyk, A. Quick Annotator: an open-source digital pathology based rapid image annotation tool. The Journal of Pathology: Clinical Research 7, 542-547. (2021)

[6] Graham, S. et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis,* 58, 101563. (2019)

Let's strike a balance: Addressing class imbalance issues in haematological images

Author
Thabang F. Isaka – School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, Ireland
Jane Courtney – School of Electrical and Electronic Engineering, Technological University Dublin, Dublin, Ireland
Claire Wynne – School of Biological, Health and Sports Sciences, Technological University Dublin, Dublin, Ireland

Citation
Isaka, T.F., Courtney, J., Wynne, C. Let's strike a balance: Addressing class imbalance issues in haematological images.

Introduction

Applying deep learning to medical imaging, especially in haematology, faces significant challenges due to class imbalance, where infected cells are vastly outnumbered by normal cells. This study addresses this issue using a Customized Repeat Factor Sampling (CRFS) method integrated into the Faster R-CNN architecture, with malaria detection as a use case. By dynamically adjusting sampling weights based on the number of infected instances, CRFS significantly improves model performance. Results show notable increases in precision, recall, and F1 scores for detecting malaria-infected cells, demonstrating the method's effectiveness in enhancing diagnostic accuracy.

Related Work

Traditional class balancing strategies in malaria diagnostics, such as 1:1 ratio adjustment and one-time random undersampling, often fail to dynamically adapt to the fluctuating densities of infections, leading to potential oversights in critical data variations. These static methods highlight the need for more flexible and adaptive solutions [1, 2]. Models such as YOLO have been extensively explored in medical diagnostics and have shown great promise. However, Detectron2, known for its robust architecture and flexibility in handling complex object detection tasks, presents an untapped opportunity in this domain. Its advanced capabilities have shown considerable promise in applications such as environmental monitoring and brain tumour detection from MRI scans but have not been specifically harnessed for malaria diagnostics [3, 4]. This gap in application highlights a significant area for enhancement. By customizing Detectron2's default class-based repeat factor sampling to be instance-aware, this technique can dynamically adjust to varying densities of infection. Class-aware applications have already been applied to general object detection and segmentation tasks, particularly with the LVIS dataset, demonstrating considerable promise [5]. Our objective is to tailor and assess its effectiveness focusing on haematological images, which often exhibit substantial class imbalance challenges.

Methodology

"The 'PlasmoCount' dataset, consisting of 398 blood smear images labelled with 2,377 malaria-infected cells and 33,687 uninfected cells, was utilized, the dataset exhibits a significant class imbalance, with infected cells constituting only 6.59% of the total [2]. To ensure effective model training and assessment, the dataset was divided into training, validation, and test sets in a 70-20-10 ratio. This distribution optimizes learning, allows thorough model refinement during validation, and assesses performance on unseen data during testing.

The Faster R-CNN R50 FPN 3x served as the baseline model and was trained on an NVIDIA Tesla T4 GPU. Systematic adjustments to learning rates and iteration counts were made to optimize performance. The process involved continuous monitoring of total loss, learning rate curves, classification accuracy, and bounding box accuracies, which were crucial for refining the model. Ultimately,

a stable performance was established without any class balancing technique. This configuration was achieved with a learning rate of 0.001, 8500 iterations, and a batch size of 128, utilizing a Stochastic Gradient Descent (SGD) optimizer in a deterministic setting (seed=42). Once the baseline model was established and its results recorded, the second experiment integrated a custom sampling technique into the training process. After careful tuning, the scale_factor was initialized at 16, as this setting demonstrated the best percentage increase. This technique increased the sampling frequency of images with more infected instances to enhance the model's learning from these critical minority class examples. The repeat factors for each image were dynamically adjusted based on the proportion of infected instances relative to the highest count in the dataset. This ensured that images with more infected instances were sampled more frequently, while also including images with no infected instances to accurately identify uninfected cells, as demonstrated in Algorithm 1.

ALGORITHM 1: Custom Repeat Factor Calculation

Input:
image_list: List of images each with a count of infected cells.
Output:
repeat_factors_array: Array containing calculated repeat factors for each image.
Procedure:
Initialize Maximum Infected Count:
max_infected = max(image['infected_count'] for image in image_list)
Compute Repeat Factors for Each Image:
For each image in image_list:
infected_count = image['infected_count']
If infected_count > 0:

$$repeat_{factor} = 1 + \left(scale_{factor} \times \frac{infected_{count}}{max_{infected}} \right)$$

Else:
repeat_factor = 1
Append repeat_factor to repeat_factors
Compile and Return Repeat Factors:
repeat_factors_array = list_to_array(repeat_factors)
Return repeat_factors_array

Results & Discussions

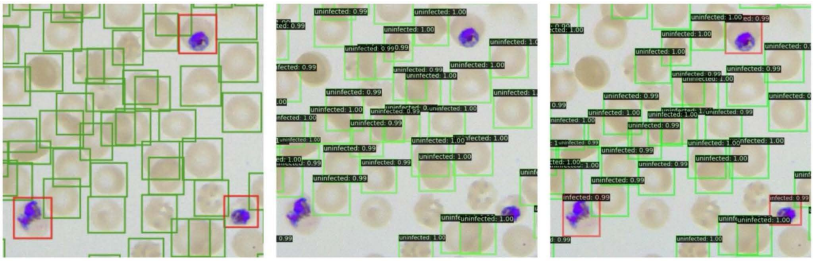
The baseline model reported an average precision (AP) of 58.4% across all classes, which was notably lower for infected cells at 43.3%, as detailed in Table 1. Conversely, the implementation of the CRFS technique significantly enhanced the model's diagnostic capabilities. AP for infected cells improved to 46.6%, with a notable increase in precision, recall, and F1 scores to 0.78, 0.55, and 0.64, respectively, as recorded in Table 2. Figure 1 provides a simple visual contrast between the diagnostic capabilities of the baseline and CRFS-enhanced Faster R-CNN models. The baseline model often failed to identify infected cells, demonstrating a bias towards recognizing uninfected cells. This is evident in Figure 1B, where the baseline model did not correctly identify any infected cells in the sample. In contrast, the CRFS-enhanced model displayed significantly improved accuracy, successfully identifying infected cells with good confidence scores, as shown in Figure 1C. Despite the improvements, challenges remain. The CRFS-enhanced model struggled with densely packed and overlapping cells, leading to some false negatives. While it performed well on non-dense smears, handling complex images with dense cell distributions requires further refinement. Hence, a need to explore advanced techniques such as ensemble learning or integrating attention mechanisms to enhance robustness in complex imaging scenarios [6, 7]. Moreover, the adoption of a dynamic adjustment of the scale_factor could significantly improve adaptability, allowing the model to fine-tune its response to variations in cell density and infection severity in real-time.

TABLE 1: Average Precision (AP) Metrics for Faster R-CNN with and without CRFS

Method	AP (all)	AP (infected)	AP (uninfected)	AP_{50} (all)	AP_{75}(all)
baseline	58.4	43.3	73.5	68.0	65.7
CRFS	**60.6**	**46.6**	**74.7**	**69.3**	**67.5**

TABLE 2: Precision, Recall, and F1 Score for Faster R-CNN with and without CRFS

	Class			
	infected		uninfected	
	baseline	CRFS	baseline	CRFS
Precision	0.71	**0.78**	0.98	0.98
Recall	0.52	**0.55**	0.86	0.86
F1-Score	0.60	**0.64**	0.91	0.91



(A) (B) (C)

FIGURE 1
Visualization: (A) Ground truth labels the red bounding boxes are for infected and green for uninfected cells. (B) Baseline model predictions, failing to classify infected cells. (C). CRFS-enhanced model predictions, successfully classifying infected cells.

Acknowledgments
This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

References

[1] Kudisthalert W, Pasupa K, Tongsima S. Counting and Classification of Malarial Parasite From Giemsa-Stained Thin Film Images. IEEE Access 2020;8:78663–82. Available from: https://doi.org/10.1109/access.2020.2990497.

[2] Davidson MS, Andradi-Brown C, Yahiya S, Chmielewski J, O'Donnell AJ, Gurung P, et al. Automated detection and staging of malaria parasites from cytological smears using convolutional neural networks. Biol Imaging 2021;1. Available from: https://doi.org/10.1017/s2633903x21000015.

[3] Dipu NM, Shohan SA, Salam KMA. Brain Tumor Detection Using Various Deep Learning Algorithms. Proceedings of the 2021 International Conference on Science and Contemporary Technology (ICSCT) 2021;1–6. Available from: https://doi.org/10.1109/icsct53883.2021.9642649.

[4] Abdusalomov AB, Islam BMS, Nasimov R, Mukhiddinov M, Whangbo TK. An Improved Forest Fire Detection Method Based on the Detectron2 Model and a Deep Learning Approach. Sensors 2023;23:1512. Available from: https://doi.org/10.3390/s23031512.

[5] Gupta A, Dollár P, Girshick R. LVIS: A Dataset for Large Vocabulary Instance Segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019;5351–9. Available from: https://doi.org/10.1109/cvpr.2019.00550.

[6] Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University - Computer and Information Sciences 2023;35:757–74. Available from: https://doi.org/10.1016/j.jksuci.2023.01.014.

[7] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing 2021;452:48–62. Available from: https://doi.org/10.1016/j.neucom.2021.03.091.

Streamlining colon biopsy screening with interpretable machine learning

Author
Quoc Dang Vu, Navid Alemi, Johnathan Pocock, David Snead, Nasir Rajpoot, Simon Graham – Histofy Ltd, Coventry, United Kingdom

Citation
Vu, Q.D., Alemi, N., Pocock, J., Snead, D., Rajpoot, N., Graham, S. Streamlining colon biopsy screening with interpretable machine learning.

Introduction
The increasing use of early screening for colon cancer is putting a lot of pressure on pathology resources worldwide. In the UK, 78% of cellular pathology departments already deal with significant staff shortages. Around one-third of endoscopic colon biopsies are reported as normal, requiring minimal intervention, but it can take 2-3 weeks to get the biopsy results. Machine learning methods offer a promising solution to alleviate this burden by effectively filtering out normal slides that do not require further intervention, thereby streamlining the diagnostic process for cancer screening.

Materials and Methods
We conducted colon biopsy screening using both black-box and interpretable machine learning approaches. We predicted the diagnosis for each slide to be normal, non- neoplastic or neoplastic. Normal slides can be automatically filtered out, while abnormal slides can be triaged. In the black-box approach, we combined a series of patch features extracted from

one of the pre-trained Vision Transformers, DINOv2 and Phikon (Oquab et al., 2023; Filiot et al., 2023), by applying a single Multi-Head Attention (MHA) Transformer layer. For the interpretable method, we extracted clinically relevant features from various tissue components detected by a U-Net-based multitask segmentation model (Graham et al., 2023a). The features were divided into patch-level summary statistics and gland-based features, which included: gland morphology, intra-gland lumen morphology, intra-gland nuclear quantification and inter-gland (lamina propria) nuclear quantification features (Graham et al., 2023b). These were then used as input to a two-headed predictive model.

Results

We evaluated different approaches using a dataset of 7,181 endoscopic colon biopsy slides. We used 5-fold cross-validation, ensuring an even distribution across patients and labels. Our baseline results using DINOv2 deep features achieved an AUC-ROC of 0.9438 $\pm$ 0.0137, 0.9921 $\pm$ 0.0033 and 0.9725 $\pm$ 0.0048 for non-neoplastic, neoplastic and abnormal categories, respectively. On the other hand, using Phikon features resulted an AUC-ROC of 0.9514 $\pm$ 0.0178, 0.9920 $\pm$ 0.0021 and 0.9753 $\pm$ 0.0035. Regarding the two-headed interpretable approach, initial findings indicate performance on par with the black-box methods with AUC-ROC greater than 0.975, albeit offering enhanced model transparency. Our two-headed approach identifies the most predictive regions, which can be viewed as a heatmap over the slide. Additionally, by using interpretable features, we can assess the reason why each region has been identified. For this, we perform a local feature ranking.

Conclusions

We introduced two competitive colon biopsy screening methods that show the potential to reduce the burden currently placed on pathologists worldwide. Our proposed interpretable approach can assist pathologists in making diagnostic decisions and enhance their trust in the algorithm, facilitating its eventual integration into clinical practice. Future work will involve further developing the proposed solution on a larger multi-centric dataset and additional assessment of model explanations across various disease subgroups.

References

Graham, Simon, Quoc Dang Vu, Mostafa Jahanifar, Shan E. Ahmed Raza, Fay yaz Minhas, David Snead, and Nasir Rajpoot. "One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification." Medical Image Analysis (2023).

Graham, Simon, Fayyaz Minhas, Mohsin Bilal, Mahmoud Ali, Yee Wah Tsang, Mark Eastwood, Noorul Wahab et al. "Screening of normal endoscopic large bowel biopsies with interpretable graph learning: a retrospective study." Gut 72, no. 9 (2023).

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez et al. "DINOv2: Learning Robust Visual Features without Supervision." Transactions on Machine Learning Research (2023).

Filiot, Alexandre, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean Baptiste Schiratti. "Scaling self- supervised learning for histopathology with masked image modeling." medRxiv (2023).

Self-supervised pre-training improves the prediction of gene mutations and tumor mutational burden in lung adenocarcinoma

Author
Arwa AlRubaian, Nasir M Rajpoot, Shan E Ahmed Raza – Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK

Citation
AlRubaian, A., Rajpoot, N.M., Raza, S.E.A. Self-supervised pre-training improves the prediction of gene mutations and tumor mutational burden in lung adenocarcinoma.

Abstract
Identifying gene mutations and tumor mutational burden (TMB) levels in lung cancer forms the basis for treatment planning and targeted therapies. In this paper, we show that self-supervised pre- training using task-related and pathology-specific augmentation can improve the prediction of both gene mutation and TMB stratification, with around a 4% increase in AUC-ROC.

Introduction
Predicting driver gene mutation directly from Hematoxylin and Eosin (H&E) stained Whole Slide Images (WSIs) have been an active research field, particularly after the proof of concept published by Coudray et. al. [9]. Several studies have followed, proposing machine learning methods to predict mutations in single genes [4, 10] or genes subsets [13, 11, 8, 15, 16].

Despite the variety of network architectures used, ranging from AlexNet [13] to Deep Multi- Magnification Net- work [4], none have surpassed the results of Coudray et. al.'s [9].

Tumor mutational burden (TMB) is the number of mutations per megabase of DNA in a tumor sample. Since the Food and Drug Administration approval of pembrolizumab for tumors with TMB greater than 10, several researchers have been exploring machine learning to stratify patients into TMB-high and TMB- low using routine H&E slides [12, 6, 14, 17]. However, this research is in its early stages, and more advanced approaches are needed to improve such stratification.

This work is among the first to redefine the contrastive learning pre-training basing it on classes of a relevant task rather than augmentations. We showcase how applying such pre-training techniques improves the prediction of gene mutation and TMB stratification in lung adenocarcinoma.

Data and Methods

Gene Mutation Detection

Dataset

A total of 586 H&E stained WSIs of lung adenocarcinoma (LUAD) frozen sections were downloaded from The Cancer Genome Atlas (TCGA) [2]. Corresponding mutation labels were collected from both CbioPortal [1] and GDC [2], and only cases with aligned labels were used, resulting in 340 slides. Of these, 65% are TP53 mutant, 37% are KRAS mutant, 18.5% are STK11 mutant, and 18.5% are EGFR mutant. All slides were processed at 20× magnification.

Self-supervised pre-training

Utilizing DINO [5], we trained a ResNet-50 via self-supervision. We adapted the augmentation strategy for producing positive samples by integrating pathology specific augmentations including: stain augmentation, separation of H&E channels, and random flip, while preserving the novel multi-crop

augmentation proposed in in the original approach. The model was trained using approximately 2000 tiles of size 512x512, randomly selected from each WSI in TCGA-LUAD.

Gene mutation classification model

First, we filtered out the non-tumor tiles using a fine-tuned Resnet-50 trained for tumor detection. Then the resulting tumor area was divided into 512x512 patches and the slide mutation label was broadcasted to all its patches. Leveraging the Resnet-50 model trained via self-supervision to capture pathology specific features, we further fine-tuned the model to predict four gene mutations (EGFR, KRAS, STK11, and TP53). Fine- tuning was done using IDaRS (Iterative Draw and Rank Sampling) pipeline developed by Bilal et. al.[3]. IDaRS ranks the tiles based on their predictive probability, then passes the top ranked tiles of each slide to the next training iteration. We adapted IDaRS to suit the multilabel classification problem. Finally, patch predictions were aggregated via majority voting to obtain the WSI prediction. The proposed model is illustrated in Figure 1 below.

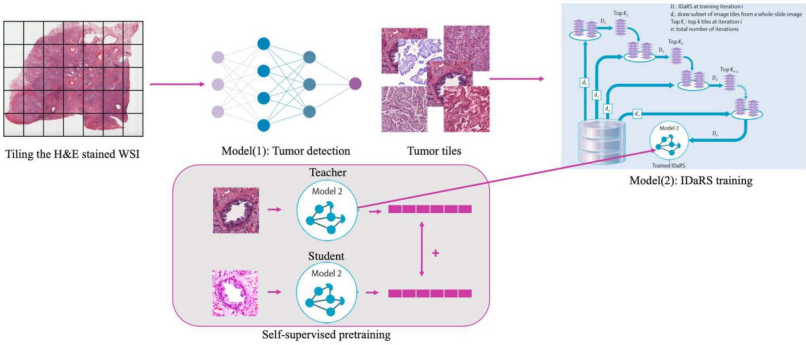


FIGURE 1
The proposed model.

TMB Status Prediction

Dataset

TCGA-LUAD

The dataset consists of 525 H&E diagnostic WSIs with corresponding TMB counts. Patients were stratified into high and low TMB groups using a cut-off of 10 mutations/megabase.

UHCW

The dataset consists of 20 H&E diagnostic WSIs collected from the University Hospital Coventry and Warwickshire, with pathologist-annotated regions indicating different growth patterns (solid, acinar, papillary, micropapillary, and lepidic).

Self-supervised pre-training

To leverage contrastive learning, we used Sim- CLR [7], to train a ResNet-50 on the UHCW dataset. We modified the definition of positive samples to be patches having the same histological growth pattern and the negative samples as patches with different patterns. Moreover, we added pathology specific augmentations including stain augmentation and separation.

TMB classification model

The pretrained ResNet-50 was used in the IDaRS pipeline and applied to TCGA–LUAD to predict TMB high vs low, in the same configuration used in the gene mutation prediction model described above.

Results

We applied 4-fold cross validation to evaluate our gene mutation prediction model on the TCGA-LUAD dataset. The average AUC-ROC results shown in Table 1 assert that self-supervised pre-training improves the model prediction.

We have also replicated Coudray et. al's work [9] on their published splits for binary STK11 mutation prediction achieving an average AUC-ROC of 0.67 ± 0.09, which our model outperformed by approximately 4%.

TABLE 1: Experimental Results

Method	*STK11*	*EGFR*	*KRAS*	*TP53*
ResNet-50	0.62 ± 0.6	0.59 ±0.08	0.62 ± 0.5	0.67 ± 0.08
IDaRS	0.65 ± 0.08	0.61 ± 0.08	0.61 ± 0.03	0.64 ± 0.06
Ours	**0.65 ± 0.03**	**0.66 ± 0.07**	**0.62 ± 0.03**	**0.7 ± 0.06**

For TMB stratification, using Resnet-50 pretrained via SimCLR in the IDaRS pipeline increased the average AUC-ROC to 0.75±0.02, compared to 0.71±0.06 when using Resnet-50 pretrained on ImageNet in the same pipeline.

Conclusion

We demonstrated that task-related augmentations in self-supervised pre-training can enhance gene mutation and TMB prediction. Our experiments suggest a potential link between growth patterns and TMB. However, the classification results lack interpretability, a focus for future improvement.

References

[1] cbioportal for cancer genomics, https://www.cbioportal.org/

[2] GDC, https://portal.gdc.cancer.gov/

[3] Bilal, M., Raza, S.E.A., Azam, A., et al.: Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospec- tive study. The Lancet Digital Health 3 (dec 2021)

[4] Campanella, G., Ho, D., Häggström, I., et al.: H&e-based computational biomarker enables universal egfr screening for lung adenocarcinoma (Jun 2022)

[5] Caron, M., Touvron, H., Touvron, H., et al.: Emerging properties in self-supervised vision transformers (Apr 2021)

[6] Chen, S., Xiang, J., Wang, X., et al.: Pan-cancer computational histopathology reveals tumor mutational burden status through weakly-supervised deep learning (Apr 2022)

[7] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. vol. 1. PMLR (Jul 2020)

[8] Chen, Y., Yang, H., Cheng, Z., et al.: A whole-slide image (wsi)-based immunohistochemical feature prediction system improves the subtyping of lung cancer. Lung Cancer 165 (mar 2022)

[9] Coudray, N., Ocampo, P.S., Sakellaropoulos, T., et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning 24(10) (Sep 2018)

[10] Hong, R., Liu, W., Fenyö, D.: Predicting and visualizing stk11 mutation in lung adenocarcinoma histopathology slides usingg deep learning. bioRxiv (2021)

[11] Huang, K., Mo, Z., Zhu, W., et al.: Prediction of target-drug therapy by identifying gene mutations in lung cancer with histopathological stained image and deep learning techniques 11 (Apr 2021)

[12] Jain, M.S., Massoud, T.F.: Predicting tumour mutational burden from histopatho logical images using multiscale deep learning 2(6) (Jun 2020)

[13] Kather, J.N., Heij, L.R., Grabsch, H.I.o.: Pan-cancer image-based detection of clinically actionable genetic alterations 1(8) (Jul 2020)

[14] Niu, Y., Wang, L., Zhang, X., et al.: Predicting tumor mutational burden from lung adenocarcinoma histopathological images using deep learning 12 (Jun 2022)

[15] Qu, H., Zhou, M., Yan, Z., et al.: Genetic mutation and biological pathway pre- diction based on whole slide images in breast carcinoma using deep learning (Sep 2021)

[16] Teichmann, M., Aichert, A., Bohnenberger, H., et al.: End-to-end learning for image-based detection of molecular alterations in digital pathology (jul 2022)

[17] Xu,H.,Clemenceau,J.R.,Park,S.,etal.:Spatial heterogeneity and organization of tumor mutation burden with immune infiltrates within tumors based on whole slide images correlated with patient survival in bladder cancer. Journal of Pathology Informatics 13 (2022)

Whole slide images classification of salivary gland tumours

Author
John Charlton – University of Sheffield, UK
Ibrahim Alsanie – Laboratory King Saud University, Saudi Arabia
Syed Ali Khurram – University of Sheffield, UK

Citation
Charlton, J., Alsanie, I., Khurram, S.I. Whole slide images classification of salivary gland tumours.

Abstract

This work shows promising results using multiple instance learning on salivary gland tumours in classifying cancers on whole slide images. Utilising CTransPath as a patch-level feature extractor and CLAM as a feature aggregator, an F1 score of over 0.88 and AUROC of 0.92 are obtained for detecting cancer in whole slide images.

Introduction

Salivary gland tumours (SGTs) are a relatively rare group of heterogeneous neoplasms. These tumours represent approximately 3% of all head and neck tumours [5, 6, 9]. Artificial intelligence methods such as deep learning have been applied to many digital histological datasets [4, 7] with very promising results. This includes high accuracy classification [12] and segmentation [10] of numerous types of cancers.

Within the body of literature, there is a gap in knowledge regarding SGTs with applications using artificial intelligence. In particular, there is no work to the authors' knowledge that utilises the entirety of the whole slide image (WSI) in applying artificial intelligence to SGTs. Incorporating knowledge of the entire WSI is important for capturing large-scale histological and morphological information across the whole tissue.

To solve this issue, this work proposes a multiple instance learning (MIL) approach applied to WSIs of SGTs. This work classifies benign/malignant tumours, as well as classification of a particular type of malignant tumour (adenoid cystic carcinoma). The work also compares the accuracy of the model when using two different feature extractors: ResNet-50 and CTransPath. It finds CTransPath to be the more accurate feature extractor, and predicts benign/malignant classification with an F1 score of 0.88 and AUROC of 0.92.

Background and Methodology

Multiple instance learning (MIL) [3, 4] is a variation on supervised learning. For MIL in this work, annotations are made at the WSI level (also known as the bag level in literature).

Salivary gland tumours display a large amount of morphological diversity between tumour types. This can be a challenge for models to accurately classify SGTs. In addition, the relative rarity of SGTs means datasets are difficult to obtain for use in training machine learning models. Machine learning models have been successfully applied to SGTs at the patch level [11, 8], region of interest (ROI) scale [1], and using a graph-based approach [2]. These works are able to classify SGTs with good accuracy, but they can be time-consuming and problematic for cancer subtyping, as high grade tumours are more challenging to annotate accurately.

Within this work two tasks were performed: benign/malignant classification, and adenoid cystic carcinoma/other classification. The first task was tested using two different feature extractors: ResNet-50 and CTransPath. The second task used only CTransPath as the feature extractor.

A dataset of 646 whole slide images of SGTs was used. Each WSI was labelled as either 'benign' (402 cases) or 'malignant' (242 cases). In addition, slides were categorised as adenoid cystic carcinoma (118 cases) or not (528 cases). More images from other clinical groups will be included in future work to help test the model robustness across different clinical workflows.

The workflow for these tasks was similar to other MIL approaches [3]. WSIs were split into smaller patches for feature extraction, then aggregated together utilising a feature aggregation model. For ResNet-50 feature extraction, the square patches were of side length 224 pixels and for CTransPath a patch was 256 pixels. Both ResNet-50 and CTransPath used the default weights of the model. CLAM was used for feature aggregation as was trained on the dataset. Training the CLAM model was performed using k-fold validation for hyperparameter tuning. A ratio of 80%-10%-10% was used for training, validation, and testing respectively. k=10 folds were used, each data point appearing only once in the validation and once in the testing set.

Results

Figure 1 shows the two receiver operating characteristic (ROC) curves of binary classification of cancer. The blue curve is for features generated by ResNet-50. The orange curve is for features generated by CTransPath. It shows an area under the ROC (AUROC) of 0.92 for the method using CTransPath features, and 0.68 when using ResNet-50 features. For the CTransPath method the F1 score is 0.88, the precision is 0.90 and the recall is 0.88. The specificity is 0.92. For the ResNet-50 method the F1 score is 0.72, the precision is 0.72, the recall is 0.77, and the specificity is 0.84.

The figure shows higher accuracy when utilising CTransPath as the feature extractor compared to ResNet-50. This might be due to the datasets they were trained on. CTransPath was trained using histological images, and the features extracted by CTransPath appear to be more useful for this classification task.

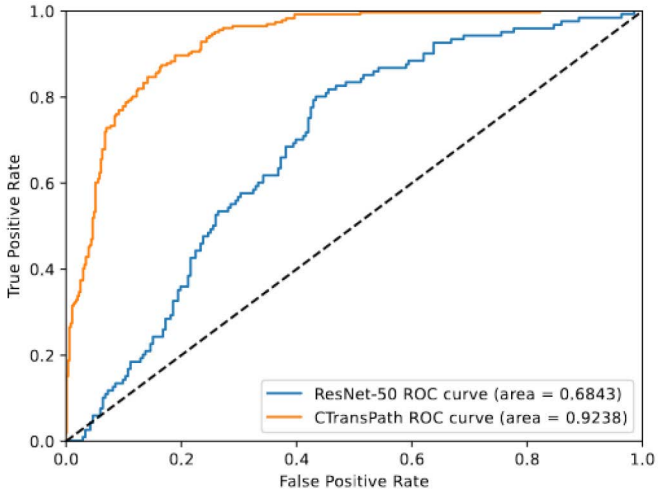


FIGURE 1
ROC curve of Benign/Malignant classification. The blue ROC curve is for ResNet-50 features. The orange curve is for CTransPath features.

The second task, Adenoid cystic carcinoma using CTransPath features with the CLAM feature aggregation model, achieved an AUROC of 0.96 and an F1 score of 0.84, displaying strong initial findings that a high grade SGT can be accurately classified for WSIs. It has a corresponding precision of 0.84, the recall is 0.77, and the specificity is 0.97.

In conclusion, CTransPath features were found to provide greater accuracy in classification of cancer compared to ResNet-50 using a MIL approach. AUROCs of over 90% were obtained for both tasks utilising CTransPath together with CLAM. The applicability of the model to other tasks is still to be explored, as well as more general conclusions about the comparison across more classification tasks. Future work will compare against recent advancements of other architectures, including autoencoders and self-supervised learning to contextualise its performance.

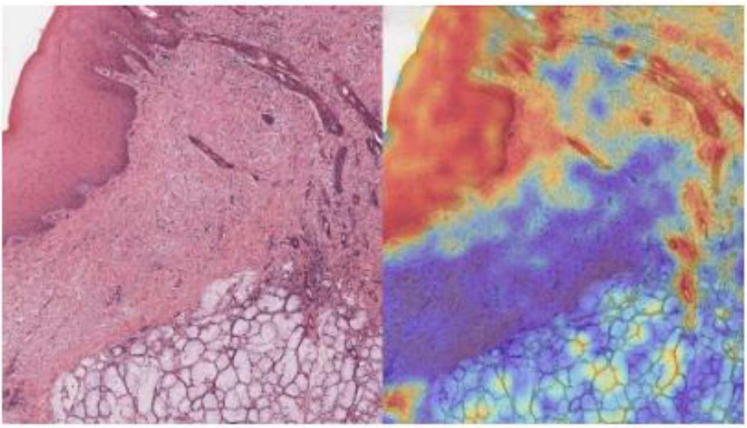


FIGURE 2
Section of a whole slide image (WSI). Heatmap of attention. Areas highlighted in red are more important in deciding the categorisation of the whole image.

The use of the attention mechanism in the CLAM model provides a focus for future study, as it highlights spatial regions within the WSI that are important for classification (see figure 2). It attends differently between different tissue types, demonstrating its ability to account for pathological features. It follows that these regions are important in understanding the behaviour of cancer development within SGTs. This can be explored in future research to examine structural effects on important properties such as cancer behaviour, response to treatment, and patient survival.

References

[1] Isanie, I., Shephard, A., Azarmehr, N., Rajpoot, N., Khurram, S.A.: Using artificial intelligence for analysis of histological and morphological diversity in salivary gland tumor. https://doi.org/10.21203/rs.3.rs-1966782/v1, https://europepmc.org/article/PPR/PPR536069

[2] Alsanie, I.S.: Using artificial intelligence for analysis of histological and morphological diversity in salivary gland tumours, https://etheses.whiterose.ac.uk/32955/

[3 Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications 77, 329–353. https://doi.org/10.1016/j.patcog.2017.10.009, https://www.sciencedirect.com/science/article/pii/S0031320317304065

[4] Gadermayr, M., Tschuchnig, M.: Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential 112, 102337. https://doi.org/10.1016/j.compmedimag.2024.102337, https://www.sciencedirect.com/science/article/pii/S0895611124000144

[5] Gontarz, M., Wyszyńska-Pawelec, G., Zapała, J.: Primary epithelial salivary gland tumours in children and adolescents 47(1), 11–15. https://doi.org/10.1016/j.ijom.2017.06.004, https://www.sciencedirect.com/science/article/pii/S0901502717314923

[6] Ito, F.A., Ito, K., Vargas, P.A., de Almeida, O.P., Lopes, M.A.: Salivary gland tumors in a brazilian population: a retrospective study of 496 cases 34(5), 533–536. https://doi.org/10.1016/j.ijom.2005.02.005, https://www.sciencedirect.com/science/article/pii/S0901502705000718

[7] Mahmood, H., Shaban, M., Rajpoot, N., Khurram, S.A.: Artificial intelligence-based methods in head and neck cancer diagnosis: an overview 124(12), 1934–1940. https://doi.org/10.1038/s41416-021-01386-x, https://www.nature.com/articles/s41416-021-01386-x, publisher: Nature Publishing Group

[8] Prezioso, E., Izzo, S., Giampaolo, F., Piccialli, F., Orabona, G.D., Cuocolo, R., Abbate, V., Ugga, L., Califano, L.: Predictive medicine for salivary gland tumours identification through deep learning 26(10), 4869–4879. https://doi.org/10.1109/JBHI.2021.3120178, https://ieeexplore.ieee.org/abstract/document/9573315, conference Name: IEEE Journal of Biomedical and Health Informatics

[9] Quixabeira Oliveira, G.A., Pérez-DE-Oliveira, M.E., Robinson, L., Khurram, S.A.,Hunter, K., Speight, P.M., Kowalski, L.P., Lopes Pinto, C.A., Sales De Sá, R.,Mendonça, E.F., Sousa-Neto, S.S., de Carlucci Junior, D., Mariano, F.V., Altemani, A.M.d.A.M., Martins, M.D., Zanella, V.G., Perez, D.E.d.C., dos Santos, J.N., Romañach, M.J., Abrahão, A.C., Andrade, B.A.B.d., Pontes, H.A.R., Jorge Junior, J., Santos-Silva, A.R., Lopes, M.A., Van Heerden, W.F.P., Vargas, P.A.: Epithelial salivary gland tumors in pediatric patients: An international collaborative study 168, 111519. https://doi.org/10.1016/j.ijporl.2023.111519, https://www.sciencedirect.com/science/article/pii/S016558762300085X

[10] Raza, S.E.A., Cheung, L., Shaban, M., Graham, S., Epstein, D., Pelengaris, S., Khan, M., Rajpoot, N.M.: Micro-net: A unified model for segmentation of various objects in microscopy images 52, 160–173. https://doi.org/10.1016/j.media.2018.12.003, http://arxiv.org/abs/1804.08145

[11] Schulz, T., Becker, C., Kayser, G.: [comparison of four convolutional neural networks for histopathological diagnosis of salivary gland carcinomas] 71(3), 170–176. https://doi.org/10.1007/s00106-023-01276-z, https://europepmc.org/articles/PMC9950222

[12] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification 81, 102559. https://doi.org/10.1016/j.media.2022.102559, https://www.sciencedirect.com/science/article/pii/S1361841522002043

Set 3: Dermatology, Cardiac Imaging and Other Medical Imaging

Deep texture analysis in whole-body PET using Graph Neural Network analysis of the sub-logit layer

Author
Robert John – Centre for Vision Speech & Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
Ian Ackerley – Centre for Vision Speech & Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
Rhodri Smith – Medical Physics and Clinical Engineering Department, Cardiff and Vale University Health Board, Cardiff, CF14 4XW, UK
Andrew Robinson – National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK
Vineet Prakash – Department of Nuclear Medicine, Royal Surrey County Hospital, Guildford, Surrey,GU2 7XX, UK
Manu Shastry – Department of Nuclear Medicine, Royal Surrey County Hospital, Guildford, Surrey, GU2 7XX, UK
Peter Strouhal – Alliance Medical Ltd, Iceni Centre, Warwick Technology Park, Warwick, CV34 5AH, UK
Kevin Wells – Centre for Vision Speech & Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

Citation
John, R., Ackerley, I., Smith, R., Robinson, A., Prakash, V., Shastry, M., Strouhal, P., Wells, K. Deep texture analysis in whole-body PET using Graph Neural Network analysis of the sub-logit layer.

Abstract
Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs) are increasingly recognised for their potential to enhance medical imaging analysis, specifically in Positron Emission Tomography (PET) scans used for cancer diagnosis. This study explores the integration of GNNs with CNN-derived activations from the sub-logit layer to identify and classify glycolytic volumes in PET images of oesophageal cancer patients. Utilising a dataset of 486 PET/CT scans, we leveraged a pre-trained 5-layer 3D CNN to segment primary tumours, which were then analysed using a GNN for comprehensive metabolic signature mapping. The model achieved a patch-based accuracy and F1 score of 97.34% and 97.33%, respectively, on the test set, demonstrating its capability to effectively distinguish between various critical glycolytic volumes, such as the primary tumour, urinary bladder and liver. These preliminary findings suggest substantial promise for this approach in refining diagnostic accuracy and treatment planning in oncology. Future work will aim to enhance the model's robustness and general applicability in clinical settings.

Introduction
Graph Neural Networks (GNNs) are emerging as a powerful tool in medical applications, due to their ability to model complex relationships and interdependencies between data points that inherently form a graph structure. GNNs show promise in the analysis of Positron Emission Tomography (PET) imaging, a modality with complex metabolic features. These networks are particularly adept at capturing the intricate patterns necessary for detailed analysis in medical diagnostics. Convolutional neural networks (CNNs) can capture characteristic deep metabolic textures from PET imaging (1). Activation values from a CNN can be viewed as a graphical

structure, where each node represents a feature captured from the PET images, and the edges reflect the relationships and dependencies among these features.

By contrast, in this work, we do not use GNN analysis of the actual voxel values in the PET images, but instead consider a graph-based analysis of the activations found in a trained sub-logit layer. In this way we consider the cascaded pattern of features in a simple 5-layer CNN originally trained on primary tumour detection, re-purposed for identifying characteristic metabolic texture features of particular organs within the PET image.

Methodology

Data and CNN Model

Four hundred and eighty-six PET/CT scans of oesophageal cancer patients were leveraged for training and testing. All patients were administered 4MBq/kg of ^{18}F-FDG. Primary tumour segmentation was performed with the ATLAAS algorithm within manually annotated bounding boxes (2). Subsequently, they were visually verified (3). A 5-layer 3D CNN was trained on this data for binary classification (4). Data was split into training, validation and test sets with a ratio of 80:10:10.

GNN Analysis Model

PET data are divided into non-overlapping patches of size 25×25×25 voxels, with the centre voxel of each patch determining its label as one of three key glycolytic volumes: primary tumour, urinary bladder, and liver. Let $\mathbf{X}_i \in R^{25\times25\times25}$ represent an input PET patch from volume i, where $i \in$ {tumour, bladder, liver}.

Each patch $\mathbf{X}_i$ is processed through a 5-layer CNN, with 32 filters per layer, to extract features. The activations from these layers are computed and then normalised to scale the values between 0 and 1. The CNN's activations are passed through an encoder, with a latent space of 32, to reduce their dimensionality:

$$\mathbf{E}_i = \sigma(\mathbf{V}\mathbf{A}_i + \mathbf{d}) \quad (1)$$

where $\mathbf{A}_i$ is the aggregated activations from the CNN, $\mathbf{V}$ and $\mathbf{d}$ are the weights and biases of the encoder, σ is the nonlinear activation function, ReLU, and $\mathbf{E}_i \in R^{32}$ is the encoded feature vector.

Finally, the encoded features $\mathbf{E}_i$ for all volumes are used as input to a Graph Neural Network (GNN) for classification. The GNN consists of two layers:

$$\mathbf{H}^{(k+1)} = \sigma(\mathbf{U}^{(k)}\mathbf{H}^{(k)}\mathbf{A}_{adj} + \mathbf{e}^{(k)}) \quad (2)$$

$$\hat{y} = \text{softmax}(\mathbf{o}^T \mathbf{H}^{(K)}) \quad (3)$$

where $\mathbf{H}^{(k)}$ is the node feature matrix at layer k, $\mathbf{U}^{(k)}$, $\mathbf{e}^{(k)}$ are the layer-specific weights and biases, $\mathbf{A}_{adj}$ is the adjacency matrix representing graph

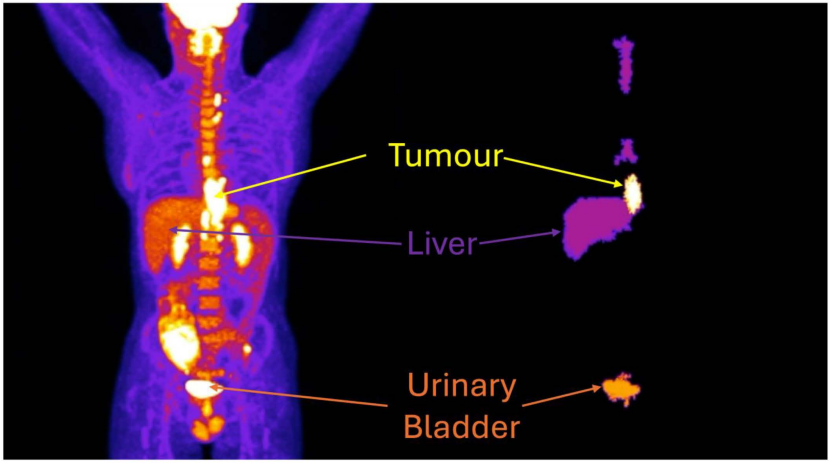


FIGURE 1
Comparative visualisation of a test-case whole-body PET scan and the model output. On the left, a Maximum Intensity Projection (MIP) highlights areas of metabolic activity, including the tumour, urinary bladder, and liver. The tumour, urinary bladder and liver are annotated in yellow, orange and purple, respectively. On the right, the model output illustrates the detection results, with false positives involving liver tissue visible in the spine.

connections, and **o** is the output layer weights. The GNN is trained using the ADAM optimiser to classify the glycolytic volume patches based on the encoded activations.

Results and Discussion

The GNN implemented to classify PET image patches from glycolytic volumes achieved a patch-based labelling accuracy and F1 score of 97.34% and 97.33%, respectively, on the test set. This performance indicator highlights the model's capability to effectively discern between the different classes associated with the primary tumour, the urinary bladder, and the liver based on the deep textural representations derived from the CNN. Figure 1, shows the model's detection of the three glycolytic volumes' deep textural representations. The identification of liver tissue in the spine presents an opportunity to further enhance the model's precision and specificity.

These results suggest that the model could significantly enhance the precision of diagnostic processes in oncology, particularly in the treatment planning of cancers, by providing reliable identification of glycolytic volumes. Such capabilities could lead to more tailored treatment plans and better patient outcomes. To improve accuracy, future work will involve refining the model to effectively minimise false positives. Further efforts will also focus on validating the model across a greater number of glycolytic volumes.

The ability of our model to accurately identify and analyse glycolytic volumes within PET scans has significant implications for clinical practice. By providing detailed insights into the metabolic activity of organs, the model offers a nuanced understanding of organ function that is crucial for effective treatment planning. Such advancements could pave the way for truly personalised medicine, where treatment strategies are optimised for individual patient profiles, significantly enhancing the quality of care and patient outcomes.

Conclusion

This study has demonstrated the effectiveness of using GNNs combined with CNNs for the precise identification and classification of glycolytic

volumes in PET imaging of oesophageal cancer. By leveraging deep texture analysis through CNNs and advanced graph-based modelling with GNNs, we achieved an accuracy and F1 score of 97.34% and 97.33%, respectively, on the test set, which underscores the potential of this methodology to significantly enhance diagnostic processes in oncology. The use of GNNs to interpret complex metabolic activities from PET images enables a deeper understanding of cancer's impact on organ function. This capability is critical for optimising therapeutic strategies, leading to more personalised and effective patient care.

Acknowledgements
The authors greatly acknowledge funding from Alliance Medical Ltd. This work is supported by the UK National Physical Laboratory through the National Measurement System.

References

(1) John RR, Ackerley I, Smith RL, Scuffham J, Robinson A, Prakash V, et al. Automatic labeling of glycolytic volumes in pet using deep texture analysis. *2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD)* (2023), 1–2.

(2) Berthon B, Marshall C, Evans M, Spezi E. Atlaas: An automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. *Phys. Med. Biol.* (2016).

(3) Foley KG, Hills RK, Berthon B, Marshall C, Parkinson C, Lewis WG, et al. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of pet in patients with oesophageal cancer. *European Radiology* **28** (2018) 428–436. doi:10.1007/s00330-017-4973-y.

(4) Ackerley I, Smith R, Scuffham J, Halling-Brown M, Lewis E, Spezi E, et al. Can deep learning detect esophageal lesions in pet-ct scans? (Institute of Electrical and Electronics Engineers Inc.) (2019). doi:10.1109/NSS/MIC42101.2019.9059833.

Detection of extracardiac findings in Cardiac Magnetic Resonance: A comparative study

Author

Edgar Pinto – Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal; ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal; Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal

Patrícia M. Costa – Department of Radiology, Hospital CUF Viseu, Viseu, Portugal

Catarina Silva – Department of Radiology, Hospital Senhora da Oliveira, Guimarães, Portugal

Vitor H. Pereira – Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal; ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal; Cardiology Department, Hospital of Braga, Sete Fontes—São Victor, 4710-243 Braga, Portugal

Jaime C. Fonseca – Algoritmi Center, School of Engineering, University of Minho, Guimarães, Portugal

Sandro Queirós – Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal; ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal

Citation

Pinto, E., Costa, P.M., Silva, C., Pereira, V.H., Fonseca, J.C., Queirós, S. Detection of extracardiac findings in Cardiac Magnetic Resonance: A comparative study.

Introduction

Cardiac Magnetic Resonance (CMR) imaging initiates with a set of sequences, termed anatomical planes, aimed at locating the heart, covering a broad

thoracic area (1). Despite their potential clinical significance, findings outside the scope of the cardiac examination are typically overlooked (2). The only previous attempt to automatically detect incidental extracardiac findings (ECFs) in these images through supervised learning have yielded suboptimal results (3), exacerbated by the impracticality of comprehensive sample collection due to the diverse range of possible anomalies across various organs.

In this study, we investigate the potential of recent anomaly detection (AD) methods to address this challenge. While AD methods have gained prominence due to their superior cluster capabilities for outlier data, their application has primarily been in industrial settings (4–12) or medical contexts like brain magnetic resonance (13–22) or chest X-ray imaging (18,23,24), which typically exhibit lower anatomical variability and complexity compared to CMR anatomical sequences. To assess the effectiveness of recent AD methods in addressing a more complex task – the detection of ECFs in CMR images - we conducted a comparative analysis of state-of-the-art unsupervised, semi-supervised, and open-set supervised AD methodologies, and compared them to a supervised baseline.

Materials and Methods

The dataset used in this study comprises CMR HASTE anatomical coronal sequences acquired during clinical routine at Hospital of Braga (HB; Portugal) between 2018 and 2019. This dataset, retrospectively gathered with the ethical approval from HB's Ethics Committee (ref. 180/2023), contains a total of 690 cases, in a total of 11,361 DICOM images. All images were labelled by one of two radiologists regarding the presence of ECFs (and respective coarse segmentation). Among the 690 cases, 269 have abnormal findings, representing a total of 10.32% of images with anomalies. To ensure robust results, the dataset was randomly split five times into training, validation, and test subsets in a patient-disjoint manner. Validation and test subsets contain equal amounts of normal and abnormal samples (100 and 500 each, respectively). All remaining images were assigned to the training subset.

Given the anisotropic nature of the sequences and the limited number of samples overall, only 2D AD methods were explored. Our benchmark was built upon the work by Lagogiannis *et al.* (18), incorporating state-of-the-art (SOTA) image reconstruction methods (VAE (25), r-VAE (15), f-anoGAN (26), H-TAE-S (16)), feature modelling (DFR (10), FAE (19), RD (5), CFlow-AD (7) and PaDiM (4)), attention-based (expVAE (9) and AMCons (20)) and self-supervised (DAE (17), CutPaste (11) and PII (27)) methods. Additionally, we incorporate recent approaches from the image reconstruction (pDDPM (14)) and feature modelling (ReContrast (8)) sub-categories. Moreover, we include a one-class semi-supervised (DDAD (23)) and two open-set supervised (DRA (6) and BGAD (12)) methods. All methods were contrasted against a supervised baseline (SupAD), employing a ResNet-18 backbone and two fully connected layers (of dimensions 256 and 1, respectively). A hyperparameter tuning was conducted per method to ensure proper convergence.

Results and Discussion

The test results on the coronal CMR dataset are summarized in Table 1, representing the average over the five dataset splits. In general, all AD methods, with the exception of DRA (6), achieved lower image-level classification performance compared to the supervised baseline. Notably, the efficacy of the baseline is strongly influenced by the proportion of abnormal images available for training. When decreasing the number of abnormal samples to 10%, 25% and 50% of those available in the training subset, the supervised baseline's performance decreases significantly, and some AD methods like FAE (19), DAE (17), or RD (5) outperform it. Despite the reliance of supervised methods on abnormal data, these results suggest that it is more effective to identify and learn abnormal patterns with supervised approaches than to learn normal patterns on this complex dataset with actual unsupervised SOTA.

TABLE 1: Test set results for benchmark methods on the coronal CMR dataset

	Method		Pixel-AP	Sample-AUROC	Sample-AP
U/IR	VAE (25)		0.0326±0.0132	0.6722±0.0779	0.6517±0.1146
	r-VAE (15)		0.0182±0.0037	0.5703±0.0529	0.5505±0.0430
	f-AnoGAN (26)		0.0267±0.0074	0.5838±0.0765	0.5496±0.0500
	H-TAE-S (16)		0.0184±0.0051	0.5995±0.0305	0.5823±0.0494
	pDDPM (14)		0.0881±0.0371	0.6121±0.0476	0.6214±0.0466
U/FM	DFR (10)		0.0935±0.0627	0.6500±0.0617	0.6360±0.1173
	FAE (19)		0.2151±0.1063	0.6814±0.0421	0.6970±0.0570
	RD (5)		0.1215±0.0806	0.6770±0.0574	0.6842±0.0894
	ReContrast (8)		0.1962±0.1032	0.6672±0.0760	0.6787±0.0496
	PaDim (4)		0.0855±0.0445	0.6131±0.0451	0.6033±0.0540
	CFlow-AD (7)		0.0458±0.0266	0.6414±0.0625	0.6219±0.1113
U/AB	expVAE (9)		0.0164±0.0041	0.4703±0.0926	0.4833±0.0716
	AMCons (20)		0.0147±0.0049	0.5648±0.0417	0.5511±0.0615
U/S-S	DAE (17)		0.1317±0.0459	0.7015±0.0465	0.7156±0.0407
	CutPaste (11)		0.0442±0.0214	0.6354±0.0792	0.6314±0.0946
	PII (27)		0.2137 ± 0.4466	0.6406±0.1080	0.6571±0.0983
SS	DDAD (23)		0.0381±0.0218	0.6620±0.0720	0.6536±0.0710
WS	BGAD (12)		0.1819±0.0240	0.6605±0.0250	0.6920±0.0360
	DRA (6)	10%	-	0.6141±0.0396	0.6261±0.0649
		25%	-	0.6348±0.0799	0.6559±0.0762
		50%	-	0.7075±0.0552	0.7183±0.0554
		100%	-	0.7387±0.0928	0.7484±0.0978
S	SupAD	10%	-	0.5614±0.0702	0.5888±0.0346
		25%	-	0.5987±0.1398	0.6411±0.1204
		50%	-	0.6410±0.0422	0.6931±0.0468
		100%	-	0.7145 ± 0.0540	0.7458 ± 0.0553

AP: average precision. AUROC: area under the receiver operating characteristic curve. U: unsupervised; SS: semi-supervised; WS: weak/open-set supervised; S: supervised; IR: image reconstruction; FM: feature modelling; AB: attention-based; S-S: self-supervised. For each metric, best and second-best results were bold and underlined, respectively. DRA and SupAD were tested with different proportions of abnormal training images: 10%, 25%, 50% and 100%.

Figure 1 illustrates some results of several high-performing AD methods. FAE (19) exhibits promising results, but the low-resolution feature maps employed generate coarse anomaly maps. In contrast, RD (5) anomaly maps have higher resolution due to the use of shallow feature maps but exhibited lower localization accuracy and overall performance. Compared to RD (5), ReContrast (8) achieved better results in localizing abnormalities, suggesting that fine-tuning the encoder on the target domain leads to more representative and sensitive features. Notably, DAE (17) emerged as an interesting option, particularly excelling when lesions resemble the synthetic noise but facing challenges with larger, homogeneous lesions. Although PII (27) has demonstrated strong pixel-level results, it is susceptible to overfitting during training, resulting in square-shaped anomalies. Finally, the supervision of BGAD (12) seemed to help improve specificity, but the pixel-AP results do not demonstrate a significant advantage of this method compared to unsupervised AD approaches. Overall, these results indicate suboptimal performance of SOTA methods, highlighting the need for further research in this domain.

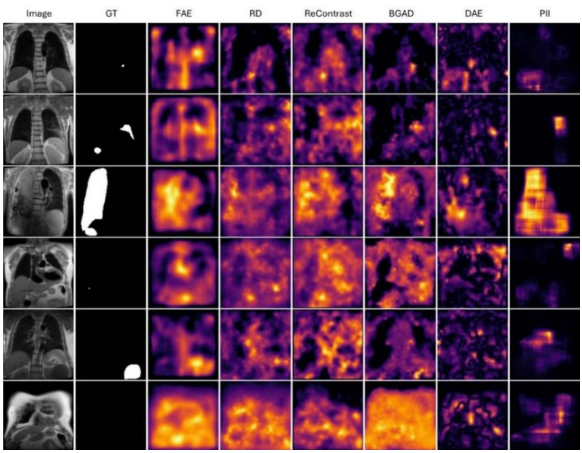


FIGURE 1
Anomaly maps generated by select high-performing methods on 6 test samples, and associated ground truth (GT).

Acknowledgements
This work was supported by Portuguese National funds, through the Foundation for Science and Technology (FCT) - projects UIDB/50026/202, UIDP/50026/2020, LA/P/0050/2020, and PTDC/EMD-EMD/1140/2020, and grant CEECIND/03064/2018 (S.Q.). The authors would also like to acknowledge the donation of a RTX A6000 GPU by NVIDIA Corporation (USA).

References

(1) Kramer CM, Barkhausen J, Bucciarelli-Ducci C, Flamm SD, Kim RJ, Nagel E. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. Journal of Cardiovascular Magnetic Resonance. 2020 Feb 24;22(1).

(2) Ufuk F, Yavaş HG, Sağtaş E, Kılıç İD. The prevalence and clinical significance of incidental non-cardiac findings on cardiac magnetic resonance imaging and unreported rates of these findings in official radiology reports. Pol J Radiol. 2022;87(1):e207–14.

(3) Wickremasinghe DH, Khenkina N, Masci PG, King AP, Puyol-Antón E. Automatic Detection of Extra-Cardiac Findings in Cardiovascular Magnetic Resonance. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Science and Business Media Deutschland GmbH; 2021. p. 98–107.

(4) Defard T, Setkov A, Loesch A, Audigier R. PaDiM: a Patch Distribution Modeling Framework for Anomaly Detection and Localization. 2020 Nov 17; Available from: http://arxiv.org/abs/2011.08785

(5) Deng H, Li X. Anomaly Detection via Reverse Distillation from One-Class Embedding.

(6) Ding C, Pang G, Shen C. Catching Both Gray and Black Swans: Open-set Supervised Anomaly Detection * [Internet]. Available from: https://github.com/choubo/DRA

(7) Gudovskiy D, Ishizaka S, Kozuka K. CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows.

(8) Guo J, Lu S, Jia L, Zhang W, Li H. ReContrast: Domain-Specific Anomaly Detection via Contrastive Reconstruction [Internet]. Available from: https://github.com/guojiajeremy/ReContrast

(9) Liu W, Li R, Zheng M, Karanam S, Wu Z, Bhanu B, et al. Towards Visually Explaining Variational Autoencoders.

(10) Shi Y, Yang J, Qi Z. Unsupervised anomaly segmentation via deep feature reconstruction. Neurocomputing. 2021 Feb 1;424:9–22.

(11) Li CL, Sohn K, Yoon J, Pfister T, Cloud G, Research AI. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization.

(12) Yao X, Li R, Zhang J, Sun J, Zhang C. Explicit Boundary Guided Semi-Push-Pull Contrastive Learning for Supervised Anomaly Detection [Internet]. Available from: https://github.

(13) Baur C, Denner S, Wiestler B, Navab N, Albarqouni S. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. Vol. 69, Medical Image Analysis. Elsevier B.V.; 2021.

(14) Behrendt F, Bhattacharya D, Krüger J, Opfer R, Schlaefer A. Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI. Vol. 227, Proceedings of Machine Learning Research. 2023.

(15) You S, Tezcan KC, Chen X, Konukoglu ENDERKONUKOGLU E. Unsupervised Lesion Detection via Image Restoration with a Normative Prior. Vol. 102, Proceedings of Machine Learning Research. 2019.

(16) Ghorbel A, Aldahdooh A, Albarqouni S, Hamidouche W. Transformer based Models for Unsupervised Anomaly Segmentation in Brain MR Images. 2022 Jul 5; Available from: http://arxiv.org/abs/2207.02059

(17) Kascenas A, Pugeault N, O'neil AQ. Denoising Autoencoders for Unsupervised Anomaly Detection in Brain MRI. Vol. 172, Proceedings of Machine Learning Research. 2022.

(18) Lagogiannis I, Meissen F, Kaissis G, Rueckert D. Unsupervised Pathology Detection: A Deep Dive Into the State of the Art. IEEE Trans Med Imaging. 2024 Jan 1;43(1):241–52.

(19) Meissen F, Paetzold J, Kaissis G, Rueckert D. Unsupervised Anomaly Localization with Structural Feature-Autoencoders. 2022 Aug 23; Available from: http://arxiv.org/abs/2208.10992

(20) Silva-Rodríguez J, Naranjo V, Dolz J. Constrained unsupervised anomaly segmentation. Med Image Anal. 2022 Aug 1;80.

(21) Wyatt J, Leach A, Schmon SM, Willcocks CG. AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise.

(22) Tan J, Hou B, Batten J, Qiu H, Kainz B. Detecting Outliers with Foreign Patch Interpolation. 2020 Nov 9; Available from: http://arxiv.org/abs/2011.04197

(23) Cai Y, Chen H, Yang X, Zhou Y, Cheng KT. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Med Image Anal. 2023 May 1;86.

(24) Tan J, Hou B, Day T, Simpson J, Rueckert D, Kainz B. Detecting Outliers with Poisson Image Interpolation. 2021 Jul 6; Available from: http://arxiv.org/abs/2107.02622

(25) Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2013 Dec 20; Available from: http://arxiv.org/abs/1312.6114

(26) Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med Image Anal. 2019 May 1;54:30–44.

(27) Tan Jeremy and Hou B and DT and SJ and RD and KB. Detecting Outliers with Poisson Image Interpolation. In: de Bruijne Marleen and Cattin PC and CS and PN and SS and ZY and EC, editor. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer International Publishing; 2021. p. 581–91.

Inter-site and inter-scanner reproducibility across four qMRI measurands using SI traceable references

Author

Ben P. Tatman – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK

Robert Hanson – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK

Amy McDowell – UCL, Gower St, London, WC1E 6BT, UK

Elizabeth A. Cooke – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK

Cailean Clarkson – National Measurement Laboratory, LGC, Queen's Road, Teddington, TW11 0LY, UK

Tugba Dispinar – TÜBITAK, Tunus Avenue 80 06100 Ankara, Türkiye

Ilker Un – TÜBITAK, Tunus Avenue 80 06100 Ankara, Türkiye

Sarah Hill – National Measurement Laboratory, LGC, Queen's Road, Teddington, TW11 0LY, UK

Sumiksha Rai – National Measurement Laboratory, LGC, Queen's Road, Teddington, TW11 0LY, UK

Ahmad Abukashabeh – National Measurement Laboratory, LGC, Queen's Road, Teddington, TW11 0LY, UK

Aaron McCann – Belfast Health and Social Care Trust, Royal Victoria Hospital, Grosvenor Road, Belfast, BT12 6BA, UK

Cormac McGrath – Belfast Health and Social Care Trust, Royal Victoria Hospital, Grosvenor Road, Belfast, BT12 6BA, UK

Sian Curtis – University Hospitals Bristol and Weston NHS Foundation Trust (UHBW), Marlborough Street, Bristol, BS1 3NU, UK

Holly Elbert – University Hospitals Bristol and Weston NHS Foundation Trust (UHBW), Marlborough Street, Bristol, BS1 3NU, UK

Jonathon Delve – University Hospitals Bristol and Weston NHS Foundation Trust (UHBW), Marlborough Street, Bristol, BS1 3NU, UK
Cameron Ingham – University Hospitals Bristol and Weston NHS Foundation Trust (UHBW), Marlborough Street, Bristol, BS1 3NU, UK
Simone Busoni – AOU Careggi, Largo Giovanni Alessandro Brambilla, 3, 50134 Firenze, Italy
Jack Clarke – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
John Thornton – UCL, Gower St, London, WC1E 6BT, UK
Nick Zafeiropoulos – UCL, Gower St, London, WC1E 6BT, UK
Stephen Wastling – UCL, Gower St, London, WC1E 6BT, UK
Alessandra Manzin – INRIM Str. Delle Cacce, 91, 10135, Torino, Italy
Riccardo Ferrero – INRIM Str. Delle Cacce, 91, 10135, Torino, Italy
Adriano Troia – INRIM Str. Delle Cacce, 91, 10135, Torino, Italy
Frederic Brochu – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Asha Forde-Scille – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Jessica Goldring – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Asante Ntata – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Katie Obee – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Susan Rhodes – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Merima Smajlhodžić-Deljo – Verifikacioni Laboratorij, Ferhadija 27, Sarajevo 71000, Bosnia & Herzegovina
Amar Deumić – Verifikacioni Laboratorij, Ferhadija 27, Sarajevo 71000, Bosnia & Herzegovina
Alen Bosnjakovic – Institute of Metrology of Bosnia and Herzegovina (IMBIH), 71000 Sarajevo, Bosnia and Herzegovina
Paul Tofts – Brighton and Sussex Medical School, Falmer, Brighton, BN1 9PX, UK
Richard Scott – University Hospitals Bristol and Weston NHS Foundation Trust (UHBW), Marlborough Street, Bristol, BS1 3NU, UK
Matt Cashmore – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK
Matt G. Hall – National Physical Laboratory, Hampton Rd, Teddington, TW11 0LW, UK

Citation
Tatman, B.P., Hanson, R., McDowell, A., Cooke, E.A., Clarkson, C., Dispinar, T., Un, I., Hill, S., Rai, S., Abukashabeh, A., McCann, A., McGrath, C., Curtis, S., Elbert, H., Delve, J., Ingham, C., Busoni, S., Clarke, J., Thornton, J., Zafeiropoulos, N., Wastling, S., Manzin, A., Ferrero, R., Troia, A., Brochu, F., Forde-Scille, A., Goldring, J., Ntata, A., Obee, K., Rhodes, S., Smajlhodžić-Deljo, M., Deumić, A., Bosnjakovic, A., Tofts, P., Scott, R., Cashmore, M., Hall, M.G. Inter-site and inter-scanner reproducibility across four qMRI measurands using SI traceable references.

Introduction

Conventional clinical Magnetic Resonance Imaging (MRI) relies on the formation of qualitative images based on relative differences in tissue response to an MR pulse sequence and thus the signal intensities are not themselves intrinsically meaningful. For high-stakes medical applications to be validated, precise, and accurate, a new culture of imaging based on metrological principles is necessary [1].

Quantitative MRI (qMRI) uses an MRI scanner to perform measurements of physical properties. These measurements can provide additional consistency and clinical specificity as well as the potential for improved consistency between sites but require a metrological foundation. Because qMRI measures physical parameters, the measurements can be used to compare the performance of different scanners [3].

The iMet-MRI project [2] has developed a phantom containing calibrated solutions suitable for qMRI. This study reports on the preliminary results from a multi-site trial of experimental measurements using multiple scanner models from different manufacturers for four qMRI measurands: T_1, T_2, T_2* and apparent diffusion coefficient (ADC).

Methods

MRI experiments were performed on vials containing metrologically calibrated solutions of $NiCl_2$ (T_1, [0.0, 1.0, 2.0, 3.0, 7.0, 13.9 mM]), $MnCl_2$

(T_2, [0.00, 0.04, 0.07, 0.13, 0.20, 0.39mM]), $FeCl_3$ (T_2*, [0.0, 2.5, 5.0, 7.6, 10.1, 12.5mM], measured using ICP-MS) and Polyvinylpyrrolidone (PVP; ADC, [5, 15, 25, 35, 55% m/m]). Nominal values for T_1, T_2 and T_2* were measured using a standardised 3 T NMR spectrometer at 20 °C. The nominal values for ADC were determined using concentration calibrations extrapolated to our measured PVP concentrations [5].

Experiments were performed across nine 1.5 T scanners comprising five distinct models made by three manufacturers (Siemens, Philips, GE) across four sites. Data were acquired using a set of experimental parameters and conditions outlined in standardised operating procedures (SOPs), with measurements taken at 20–25 °C. Analysis of the resulting datasets was performed by fitting standard mono-exponential decay. Uncertainties were estimated for the fitting as 1.96x the standard deviation (corresponding to a coverage factor of k=2) of all pixels within a selected region of interest (ROI).

Results and Discussion

Scanner limitations meant some data acquired deviated from the prescribed SOPs. For example, some scanner set ups were unable to obtain echo times > 50 ms for T_2 measurements. This meant that these datasets could not be directly compared to other datasets which followed the SOPs. We remove the datasets with deviations from further analysis.

Figure 1(a) shows the T_1 relaxation time measurements compared to the nominal values. The measured values show good agreement within uncertainty between different scanners, however there is a bias observed between parameter estimates from GE and Philips compared to those from Siemens scanners. Calculated T_1 values from different manufacturers were compared using a Mann-Whitney U-test (MWU) [4]. This gives p>0.05 for all nominal values of T_1, indicating no significant difference between results obtained from different manufacturers. However, we note that our sample size is small, making it difficult to confirm the observed bias. Further data is required to determine if systematic differences between scanners exists. There is larger deviation from nominal at higher values of T_1.

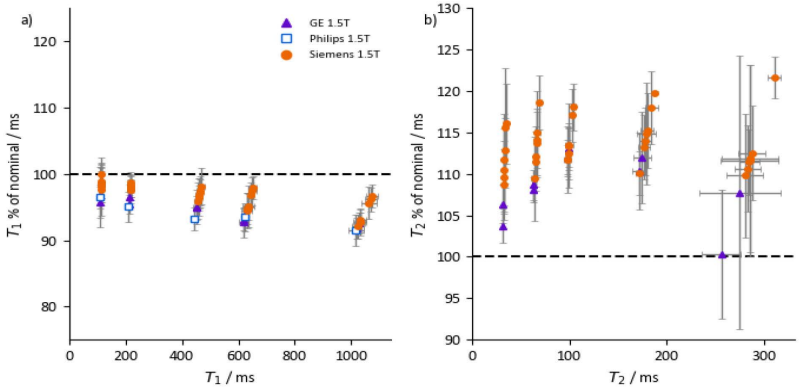


FIGURE 1
Comparison of T_1 and T_2 values for different scanners to nominal values obtained by calibration solutions in a phantom.

Figure 1(b) shows a similar comparison for T_2 acquisitions. We find no significant difference between Siemens and GE scanners (MWU p>0.05) for all nominal values of T_2. We find more variation between scanners at higher values of T_2 and poor agreement with the nominal values, however this is likely due to the higher temperature used for the nominal values.

Figure 2(a) shows a comparison of the T_2* values to nominal. The measured values show good agreement within uncertainty between scanners, with no significant difference found between different manufacturers (MWU p>0.05). We find no trend in measured T_2* with nominal T_2* value. Figure 2(b) shows a similar comparison for ADC. We find better agreement with the nominal values at higher ADC, and good agreement between different scanners for the few scanned here.

Although here we find no significant difference in results obtained on scanners of different make, there are indications that differences may exist, and further work with more scanners is needed. We emphasise that our

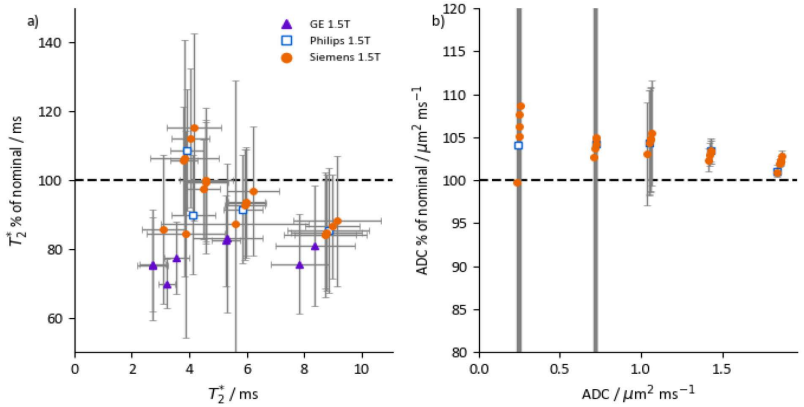


FIGURE 2
Comparison of T_2^* and ADC values for different scanners to nominal values obtained by calibration solutions in a phantom.

results have been obtained using the same set up, following SOPs, with the same calibrated solutions in our phantom, and processed using the same fitting routines. Any differences between results are therefore likely due to subtle differences between scanners, emphasising the need for standardisation in qMRI measurements.

Conclusions

Our results measuring qMRI parameters across different sites and scanner manufacturers highlight the difficulties facing clinical MR physicists when implementing qMRI measurements. Robustly traceable references allow the quantification of the difference in measurement performance between MRI scanners. The iMet-MRI program continues to work towards establishing acquisition and processing protocols to inform future standards and allow quantitative measurements to be taken and metrologically compared across different sites and scanner models.

We find no significant differences between results obtained from different manufacturers and from other parameters/sequences. We highlight difficulties in comparison between manufacturers due to restrictions on standard settings. Our results highlight the difficulties when implementing qMRI measurements and emphasise the need for standardisation.

Acknowledgements
This project 20NRM05 iMet-MRI has received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme.

References

[1] Cashmore MT, McCann AJ, Wastling SJ, McGrath C, Thornton J, Hall MG. Clinical quantitative MRI and the need for metrology. The British Journal of Radiology 2021 94:1120 https://doi.org/10.1259/bjr.20201215

[2] iMet-MRI Homepage https://empir.npl.co.uk/imet-mri/

[3] Keenan KE, Gimbutas Z, Dienstfrey A, Stupic KF, Boss MA, Russek SE, et al. Multi-site, multi-platform comparison of MRI T_1 measurement using the system phantom. PLoS One 2021;16. https://doi.org/10.1371/journal.pone.0252966.

[4] Mann H. B., Whitney D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Statist. 1947;18(1):50–60. https://doi.org/10.1214/aoms/1177730491

[5] F. Wagner *et al*., Temperature and concentration calibration of aqueous polyvinylpyrrolidone (PVP) solutions for isotropic diffusion MRI phantoms', *PLoS One*, vol. 12, no. 6, Jun. 2017, doi: 10.1371/journal.pone.0179276.

How many spin echoes are enough? Sensitivity of T_2 estimation to image noise and B_1 penetration effects

Author

Asante Ntata – National Physical Laboratory, Teddington, United Kingdom
Zeinab Al-Siddiqui – National Physical Laboratory, Teddington, United Kingdom
Nadia Smith – National Physical Laboratory, Teddington, United Kingdom
Elizabeth Cooke – National Physical Laboratory, Teddington, United Kingdom
Paul Tofts – Brighton and Sussex Medical School, Brighton, United Kingdom
Matt Cashmore – National Physical Laboratory, Teddington, United Kingdom
Matt Hall – National Physical Laboratory, Teddington, United Kingdom

Citation

Ntata, A., Al-Siddiqui, Z., Smith, N., Cooke, E., Tofts, P., Cashmore, M., Hall, M. How many spin echoes are enough? Sensitivity of T_2 estimation to image noise and B_1 penetration effects.

Abstract

There is a lack of consensus on the optimal number of echoes for accurate T_2 estimation in MRI measurements. Estimation accuracy is affected by the presence of noise in spin echo intensities and B1 penetration effects. Our goal is to determine the optimal number of spin echoes and explore how flip angle bias away from 180° affects the accuracy of T_2 estimation. We present a simulation-based approach based on the extended phase graphs framework to determine optimal parameters. We have found that 6 echoes are optimal

for accurate T_2 estimation and deviations of flip angle greater than $\pm 5°$ will affect estimation accuracy. We present an approach based on the extended phase graph framework to determine optimal parameters for accurate T_2 estimation. Our results help reach consensus on optimal parameters and provide an efficient way to determine optimal parameters for bespoke pulse sequences.

Introduction

T_2 relaxation times, or simply "T_2", are determined using a Carr-Purcell-Meiboom-Gill (CPMG) sequence[1]. This consists of a 90° excitation pulse followed by a series of 180° pulses with alternating polarity. A spin echo is the refocusing of magnetisation after each 180° pulse. The echo intensities form an exponential decay curve where T_2 is given by the decay constant. Challenges in accurately extracting T_2 include the presence of image noise[2] and B1 penetration effects[3], which lead to reduced flip angles (α) [4]. Without noise, 2 echoes are enough to sufficiently extract T_2 (this is the smallest number of points needed to fit an exponential curve). When noise is considered, however, 2 echoes alone will result in an unreliable T_2 measurement and additional echoes are required.

The number of spin echoes or echo train length (ETL) required for a reliable T_2 estimate is an open question. This is evidenced by the various ETLs used by scanner vendors and differences in common practice between sites[5,6]. Here we present a simulation-based approach for determining the optimal ETL. We also explore the effects of deviations of α away from 180°. The simulations are performed using the extended phase graph (EPG) approach [7] which includes the effects of gradients, RF pulses, relaxation, and dephasing as matrix operations. In this abstract we a) demonstrate the use of the EPG approach to model a CPMG sequence for T_2 mapping, b) determine the optimal ETL for accurate T_2 estimation in the presence of noise and c) demonstrate the sensitivity of T_2 estimation to B1 penetration effects. To the best of our knowledge, this is the first time the EPG framework has been used to investigate measurement uncertainty in MRI.

Methods

We used the EPG approach to generate spin echoes from a from a CPMG sequence. Initially the RF operator with $\alpha = 90°$ is used to model the rotation of spins into the transverse plane. This is followed by the shift operator which models dephasing and evolution of magnetisation over time. The relaxation operator is then used to model longitudinal and transverse relaxation. Then the sequence repeats (according to the desired ETL), beginning with an RF operator with $\alpha = 180°$ to model refocusing.

To investigate the effect of image noise, we generated echoes for 2 scenarios: a) A noiseless CPMG sequence which we used as a gold standard to extract the "True" T_2 and b) a spin echo with the addition of noise to represent what happens in reality. We performed this for a number of echoes, in each case evaluating the mean squared error (MSE) in the estimated T_2 with reference to the "true T_2". We repeated this for varying degrees of noise. Each calculation was repeated 1000 times with independent realisations of noise.

B1 penetration effects manifest themselves as deviations in α away from 180°. This is modelled straightforwardly in the EPG framework by modifying α in the RF operator. We generated spin echoes (noiseless) for α ranging from 150°-180° using 20 echoes. In each case we evaluated the MSE in estimated T_2 with reference to the "True T_2".

Results

Figure 1A shows idealised spin echoes with noise and a noiseless T_2 curve and spin echoes with the addition of noise at an SNR of 15.8. It's clear that fitting a curve through these points with noise would lead to errors in estimated T_2.

Figure 1B shows the dependence of MSE in estimated T_2 on the ETL for SNRs 15.8, 22.4 & 31.6. After 6 echoes there are no benefits to increasing the ETL regardless of the noise level.

Figure 2 shows that deviations in α within $\pm$ 10° will lead to accurate estimations of T2 whereas deviations greater than this will increasingly lead to bias in T_2 values.

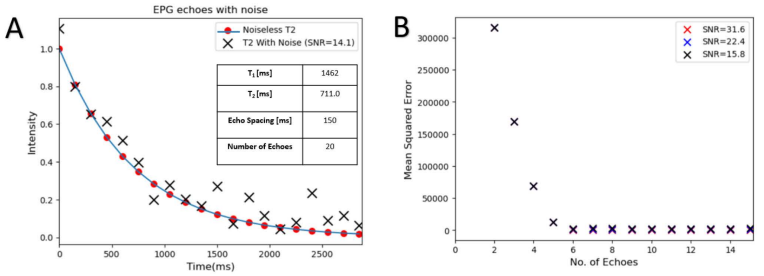


FIGURE 1
A) Plot showing ideal spin echo intensity (red points) with True T_2 (blue line) and spin echo with noise during the time course of a spin echo sequence. B) Plot showing how the accuracy in estimated T_2 varies depending on the number of acquired echoes. 6 echoes is the optimal number of echoes to acquire. There are no benefits to increasing the number of echoes beyond 6 regardless of the noise level.

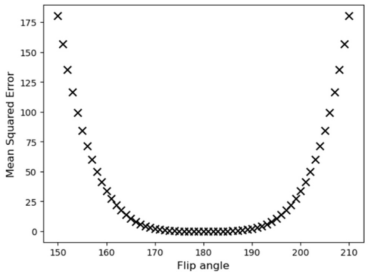


FIGURE 2
Plot showing how the accuracy in estimated T_2 varies depending on the flip angle. Deviations in α within $\pm$ 10° lead to accurate estimations of T_2 whereas deviations greater than this increasingly lead to bias in T_2 values.

We have presented a simulation-based approach for determining the optimal ETL and explored how deviations of α away from 180° affect the accuracy of T_2 estimation. We have found that 6 echoes are optimal and deviations in α greater than $\pm$ 10° will affect estimation accuracy. We note that including Rician noise in echoes tends to increase estimated T_2 whereas biasing α tends

to decrease estimated T_2. This suggests it's possible to compensate for noise effects by changing α, but more work is needed to investigate feasibility.

We have not varied T_2 or the echo time in these results thus further investigation will consist assessing the T_2 dependence for a constant echo-time-to- T_2 ratio.

We have presented an approach using the EPG framework to determine optimal parameters for accurate T_2 estimation. Our results help reach consensus on optimal parameters and provide an efficient way to determine optimal parameters for bespoke sequences.

References

[1] Meiboom S, Gill D. Modified spin-echo method for measuring nuclear relaxation times. Review of Scientific Instruments. 1958;29(8):688–91.

[2] Does MD, Olesen JL, Harkins KD, Serradas-Duarte T, Gochberg DF, Jespersen SN, et al. Evaluation of principal component analysis image denoising on multi-exponential MRI relaxometry. Magn Reson Med. 2019 Jun 1;81(6):3503–14.

[3] Hubbard Cristinacce PL, Keaveney S, Aboagye EO, Hall MG, Little RA, O'Connor JPB, et al. Clinical translation of quantitative magnetic resonance imaging biomarkers – An overview and gap analysis of current practice. Physica Medica. 2022 Sep 1;101:165–82.

[4] De Deene Y, De Wagter C, De Neve W, Achten E. Artefacts in multi-echo T 2 imaging for high-precision gel dosimetry: II. Analysis of B 1-field inhomogeneity. Vol. 45, Phys. Med. Biol. 2000.

[5] Chhetri G, McPhee KC, Wilman AH. Bloch modelling enables robust T2 mapping using retrospective proton density and T2-weighted images from different vendors and sites. Neuroimage. 2021 Aug 15;237.

[6] Karakuzu A, Biswas L, Cohen-Adad J, Stikov N. Vendor-neutral sequences and fully transparent workflows improve inter-vendor reproducibility of quantitative MRI. Magn Reson Med. 2022 Sep 1;88(3): 1212–28.

[7] Weigel M. Extended phase graphs: Dephasing, RF pulses, and echoes - Pure and simple. Vol. 41, Journal of Magnetic Resonance Imaging. John Wiley and Sons Inc; 2015. p. 266–95.

Parameter-free bio-inspired channel attention for enhanced cardiac MRI reconstruction

Author

Anam Hashmi – ML-Labs, Dublin City University, Dublin, Ireland
Julia Dietlmeier – Insight SFI Research Centre for Data Analytics, Dublin City University, Dublin, Ireland
Kathleen M. Curran – School of Medicine, University College Dublin, Dublin, Ireland
Noel E. O'Connor – Insight SFI Research Centre for Data Analytics, Dublin City University, Dublin, Ireland

Citation

Hashmi, A., Dietlmeier, J., Curran, K.M., O'Connor, N.E. Parameter-free bio-inspired channel attention for enhanced cardiac MRI reconstruction.

Abstract

Attention is a fundamental component of the human visual recognition system. The inclusion of attention in a convolutional neural network amplifies relevant visual features and suppresses the less important ones. Integrating attention mechanisms into convolutional neural networks enhances model performance and interpretability. Spatial and channel attention mechanisms have shown significant advantages across many downstream tasks in medical imaging. While existing attention modules have proven to be effective, their design often lacks a robust theoretical underpinning. In this study, we address this gap by proposing a non-linear attention architecture for cardiac MRI reconstruction and hypothesize that insights from ecological principles can guide the development of effective and efficient attention mechanisms. Specifically,

we investigate a non-linear ecological difference equation that describes single-species population growth to devise a parameter-free attention module surpassing current state-of-the-art parameter-free methods.

Background and Motivation

Attention mechanisms play a pivotal role in enhancing the efficacy of convolutional neural networks (CNNs) by focusing on relevant visual features. Attention mechanisms have been successfully integrated as plug-and-play modules into CNNs for various applications such as medical image classification, segmentation, explainability, and most recently Cardiac Magnetic Resonance (CMR) reconstruction [1], [2]. Several prominent architectures include the squeeze-and-excitation networks (SE) [3], linear context transform block (LCT) [4], and convolutional block attention module (CBAM) [5]. These and other state-of-the-art attention architectures have demonstrated success in different applications, each tailored to specific tasks and requirements. and other state-of-the-art attention architectures [6], [7], [8], [9] have demonstrated success in different applications, each tailored to specific tasks and requirements. While these attention architectures, along with others, exhibit promising results and some are even parameter-free [7], [9], their design often lacks a solid theoretical foundation. Thus, despite their efficacy, there remains a need for a more rigorous theoretical basis to underpin their design choices.

Biologically-Inspired Attention

In this work, we draw inspiration from non-linear ecological difference equations used in population biology [10]. These mathematical equations describe the dynamical system of growth of a given population depending on several environmental factors. Specifically, the following Equation (1) [10] (Fig.1 shows implementation in the blue box) describes single-species population growth:

$$N_{t+1}=\lambda[1+\alpha N_t]^{-b}N_t \quad (1)$$

$$N_{t+1}=\lambda[1+\alpha N_{t1}]^{-b}N_{t2} \quad (2)$$

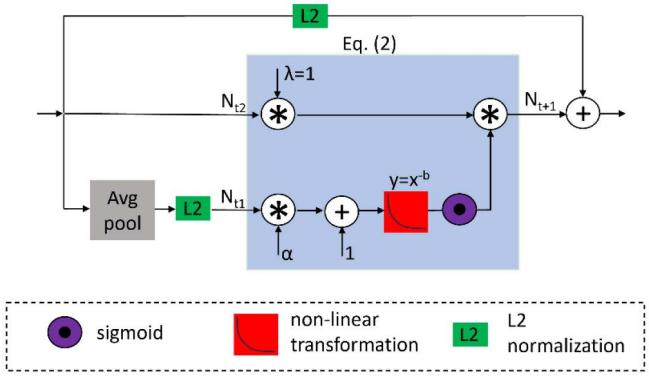


FIGURE 1
A schematic of the proposed biologically-inspired attention block. Our systematic design is based on modelling a non-linear ecological difference equation [10] in addition to average pooling, sigmoid activation function and L2-normalization layers. This follows the original Eq. (1) but with two separate branches N_{t1} and N_{t2} as defined in Eq. (2). Note that α=2 and b=2. Optimizing the hyperparameter values could yield better results, but is beyond the scope of this work.

Our proposed attention block is shown in Figure 1. The logic behind using non-linear ecological equations to model attention draws upon the principles of dynamic adaptation observed in natural systems. Biological populations dynamically adapt to environmental changes, emphasizing relevant factors and suppressing less important ones, similar to the objective of attention mechanisms in neural networks. Moreover, ecological systems, like population dynamics, often show complex non-linear behavior and integrating these dynamics allows to capture the real-world complexity more accurately. Thus, we hypothesize that non-linear ecological equations provide a framework for capturing the complex dynamic interactions between different visual features.

Methodology and Experiments

We experiment with the recently released CMRxRecon dataset [11] and obtain quantitative validation metrics such as PSNR, MSE and SSIM [12]. For the reconstruction task, we utilize a CMR reconstruction network from [1] and term the configuration without attention layers as the *baseline* model.

Our methodology follows [1], [2]. The preprocessing and training details mirror those in [1]. Our computing pipeline was implemented using Python 3.11.5 and PyTorch 2.1.1. Table 1 shows quantitative results. Figure 2 shows selected qualitative results and normalized error maps.

TABLE 1: Comparison of quantitative results. Wilcoxon signed-rank test was conducted to compare the SSIM scores of our method and the competing method SimAM. The results indicated a significant difference between the two methods (p-value < 0.001), demonstrating that our method significantly outperforms SimAM.

Attention	Computational overhead	PSNR ↑	MSE ↓	SSIM ↑
Baseline	0	36.2068	0.0002878	0.9245
+ SE [3]	119,936	37.2863	0.0002355	0.9429
+ LCT [4]	6,848	36.7562	0.0002618	0.9450
+ AB [6]	3,424	34.3633	0.0004299	0.9357
+ ECA [8]	90	37.9982	0.0002262	0.9527
+ SimAM [7]	0	37.0492	0.0002583	0.9443
+ GCT [9]	0	36.5874	0.0002695	0.9408
+ Proposed	0	**37.7724**	**0.0002231**	**0.9496**

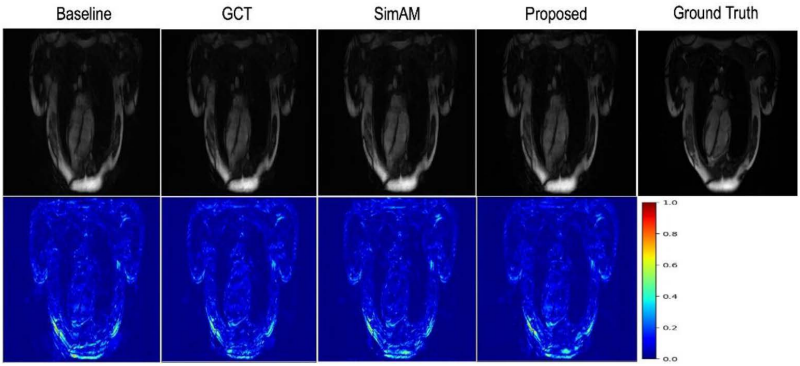


FIGURE 2
The reconstruction results and normalized absolute error maps for a long-axis cine image (case P1) with a ×10 acceleration factor, across different methods.

Conclusion

Our approach, based on non-linear ecological difference equations drawn from established ecological principles, outperformed existing parameter-free methods underscoring its effectiveness in enhancing cardiac MRI reconstruction.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant numbers 18/CRT/6183 and 12/RC/2289_P2.

References

[1] Hashmi A., Cardiac MRI Reconstruction with CMRatt: An Attention-Driven Approach. https://arxiv.org/abs/2404.06941 (2024)

[2] Dietlmeier J., Cardiac MRI Reconstruction from Undersampled K-Space Using Double-Stream IFFT and a Denoising GNA-UNET Pipeline. *In: Camara, O., et al. Statistical Atlases and Computational Models of the Heart. Regular and CMRxRecon Challenge Papers, Lecture Notes in Computer Science*, vol 14507, pp. 326--338. Springer, Cham. (2023)

[3] Hu J., Squeeze-and-Excitation Networks. *CVPR* (2018)

[4] Ruan D., Linear Context Transform Block. *AAAI* (2020)

[5] Woo S., Convolutional Block Attention Module. *ECCV* (2018)

[6] Klomp SR, Performance-Efficiency Comparisons of Channel Attention Modules for ResNets. *Neural Processing Letters* 55, 6797--6813 (2023)

[7] Yang L, SimAM: A simple, parameter-free attention module for convolutional neural networks. *ICML*, 11863–-11874 (2021)

[8] Wang Q., ECA-Net: Efficient channel attention for deep convolutional neural networks. *CVPR* (2020)

[9] Ruan D., Gaussian context transformer. *CVPR* (2021)

[10] May RM, Biological Populations Obeying Difference Equations : Stable Points, Stable Cycles, and Chaos. *Journal of Theoretical Biology* 51, 511--524 (1975)

[11] Wang C., CMR×Recon: An open cardiac MRI dataset for the competition of accelerated image reconstruction. https://arxiv.org/pdf/2309.10836.pdf (2023)

[12] Sara U., Image quality assessment through FSIM, SSIM, MSE and PSNR — a comparative study. *Journal of Computer and Communications* 7(3), (2019)

Set 4: Machine Learning for Endoscopy (EndoML)

Automatic assessment of the degree of cleanliness in esophagogastroduodenoscopy images using EfficientNet-V2 network

Author

Neil de la Fuente – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Spain

Mireia Majó – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Spain

Yael Tudela – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Spain

Irina Luzko – Endoscopy Unit, Hospital Clínic, IDIBAPS, CIBEREHD, University of Barcelona, Barcelona, Spain

Henry Córdova – Endoscopy Unit, Hospital Clínic, IDIBAPS, CIBEREHD, University of Barcelona, Barcelona, Spain

Gloria Fernández-Esparrach – Endoscopy Unit, Hospital Clínic, IDIBAPS, CIBEREHD, University of Barcelona, Barcelona, Spain

Jorge Bernal – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Spain

Citation
Fuente, N.d.l., Majó, M., Tudela, Y., Luzko, I., Córdova, H., Fernández-Esparrach, G., Bernal, J. Automatic assessment of the degree of cleanliness in esophagogastroduodenoscopy images using EfficientNet-V2 network.

Abstract

Gastric cancer is one of the most prevalent worldwide and its mortality rate depends on its early diagnosis, being esophagogastroduodenoscopy the gold standard tool for lesion detection. Its performance depends on the degree of cleanliness of the stomach. In this study an automatic method to assess the degree of cleanliness of the stomach is presented and evaluated in a challenging dataset showing promising results along with indicating future lines of research.

Introduction

Gastric cancer (GC) ranks as the fifth most prevalent cancer globally, with over 1 million new cases reported in 2020. Esophagogastroduodenoscopy (EGD) serves as the established diagnostic method for GC, with studies indicating that early detection significantly reduces mortality rate, as it is shown in Ezoe et al. (1) However, up to 10% of cancers are missed during the exploration, negatively impacting patient survival rates, as mentioned in Tsukuma et al. (2)

The assessment of cleanliness and mucosal visibility holds critical importance in EGD procedures, representing a key factor in cancer diagnosis accuracy. Two scales have been recently published: POLPREP proposed by Romańczyk et al. (3) and the Barcelona scale proposed by Córdova et al. (4). These scales evaluate cleanliness levels across the esophagus, stomach, and duodenum, differing in the number of levels and detail of evaluation. While POLPREP comprises four levels, the Barcelona scale includes three, further segmenting the stomach into fundus, corpus, and antrum.

The objective of this study is to validate the potential of Artificial Intelligence (AI) to automatically determine the quality of cleanliness during esophagogastroduodenoscopy. Such a method would be of importance as it

would allow clinicians to have an in-vivo and real-time objective information which can be used to determine whether the outcome of the exploration is valid or that new procedures should be carried out in case of a poorly prepared stomach.

Recent studies have denoted the potential of AI to automatically assess the level of cleanliness of endoluminal structures such as small bowel or the colon. Ribeiro et al. (5) applied a convolutional neural network for automatic classification of small bowel cleansing in capsule endoscopy, highlighting the model's capability in diverse gastrointestinal conditions. Similarly, Manh et al. (6) employed a U-Net and classification model combination to assess cleaning levels in esophageal images, setting a precedent for machine learning applications in gastroenterological image evaluation. The work of Haithami et al. (7) deals with bowel preparation in colonoscopy and proposes the novel, in this context, use of a Gated Recurring Unit (GRU) as part of their processing pipeline aiming at providing a temporal coherence in method's output.

Development

The foundation of this study is the EfficientNetv2 neural network proposed by Tan et al. (8), chosen for its scalability and state-of-the-art performance in handling image data. EfficientNetV2 employs a compound scaling method, optimizing the model's depth, width, and resolution simultaneously to enhance accuracy and efficiency. This balanced scaling approach allows EfficientNetV2 to achieve superior performance compared to previous models. Furthermore, EfficientNetV2 incorporates Fused-MBConv layers, combining the benefits of MobileNetV2's inverted bottleneck blocks with conventional convolutional layers, resulting in improved speed and accuracy.

EfficientNetv2 is particularly beneficial in medical imaging due to its fast-training speeds and efficient use of parameters, allowing for high accuracy even with limited labelled data. This makes it ideal for evaluating the cleanliness of esophagogastroduodenoscopy images, where precision is crucial.

In this study, EfficientNetv2 neural network was fine-tuned to process high-definition white light endoscopy images. These images were categorized into

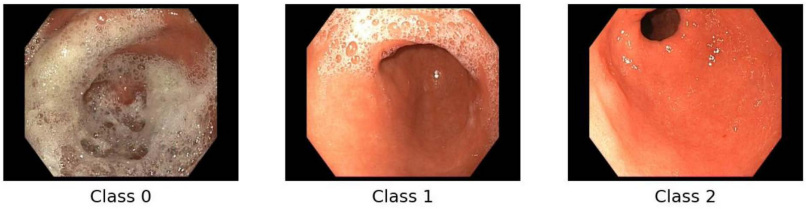


FIGURE 1
Examples of the content of the dataset.

three cleanliness levels based on the Barcelona scale. Additionally, various hyperparameters such as the learning rate and regularization techniques were fine-tuned to enhance the model's sensitivity and specificity. These modifications were critical in enabling the model to detect subtle differences in cleanliness, which is essential for accurate clinical assessment and diagnosis.

Experimental Setup

The dataset used in this study contains 125 High-Definition white light endoscopy images using OLYMPUS EXERA. Images were selected by experts and their degree of cleanliness was determined using Barcelona scale. Out of all the images, 43 (34.44%) were of class 0, 36 of class 1 (28.8%) and 46 (36.8%) of class 2, see sample images in figure 1. The system was built adapting EfficientNetv2 neural network architecture; 94 images (75%) were used for the training stage and the rest (31, 25%) for validation. Special attention was put to ensure an adequate distribution of the different classes in the train and validation sets.

EfficientNetV2 was trained on a machine with 3 NVIDIA RTX 2080Ti GPU. Adam optimizer was used for dynamic learning rate adjustments, and dropout regularization was implemented to prevent overfitting.

Due to the small size of the dataset, 4-Fold Cross Validation was used in the training and validation stages. It must be noted that in each fold an image cannot be in both train and validation sets. Regarding metrics, usual metrics

such as Precision, Recall and Accuracy were used to assess the performance of the AI system.

Results

The system was able to accurately determine the quality of cleanliness in 92 of 125 images (global accuracy of 72.38%). Regarding performance per class, the AI system obtained better results for class 0 (Precision: 76.93%, Recall: 76.22% and Accuracy: 77.12%) and class 2 (Precision: 88.26%, Recall: 83.54% and Accuracy: 88.72%) than for class 1 (Precision: 51.72%, Recall: 56.43% and Accuracy: 51.31%).

Class	Precision (PPV)	Recall (Sensitivity)	Specificity	Accuracy
0	0.7693	0.7622	0.7766	0.7712
1	0.5172	0.5643	0.4698	0.5131
2	0.8826	0. 8354	0.9176	0.8872

After analysing the results, a high level of overlap between classes 0 and 1 was observed. Considering this, an additional experiment was run to assess whether the AI system could separate between the aggregate of class 0 and class 1 against class 2. In this context, the system was able to classify correctly 121 of 125 images (Accuracy: 96.97%). In all cases, the system took less than 20 milliseconds to provide its output, achieving real-time performance.

Conclusions

The results obtained in this study show the potential of AI to assist clinicians to assess the quality of cleanliness of esophagogastroduodenoscopy during in-vivo explorations, providing an accurate output in real-time. Nevertheless, the size and distribution of the samples in the dataset limits the performance that the system can achieve. This performance is also damaged by the great visual similarities between certain instances of classes 0 and 1 which is proven by the better results obtained in the two-class experiment. Future work should involve the acquisition and annotation of new images to improve the performance of AI systems.

References

(1) Ezoe Y, Muto M, Uedo N, Doyama H, Yao K, Oda I, et al. Magnifying narrowband imaging is more accurate than conventional white-light imaging in diagnosis of gastric mucosal cancer. Gastroenterology. 2011;141(6).

(2) Tsukuma H, Oshima A, Narahara H, Morii T. Natural history of early gastric cancer: A non-concurrent, long term, follow up study. Gut. 2000;47(5).

(3) Romańczyk M, Ostrowski B, Kozłowska-Petriczko K, Pawlak KM, Kurek K, Zatorski H, et al. Scoring system assessing mucosal visibility of upper gastrointestinal tract: The POLPREP scale. Journal of Gastroenterology and Hepatology (Australia). 2022;37(1).

(4) Córdova H, Barreiro-Alonso E, Castillo-Regalado E, Cubiella J, Delgado-Guillena P, Díez Redondo P, et al. Applicability of the Barcelona scale to assess the quality of cleanliness of mucosa at esophagogastroduodenoscopy. Gastroenterol Hepatol. 2024;47(3).

(5) Ribeiro T, Mascarenhas Saraiva MJ, Afonso J, Cardoso P, Mendes F, Martins M, et al. Design of a Convolutional Neural Network as a Deep Learning Tool for the Automatic Classification of Small-Bowel Cleansing in Capsule Endoscopy. Medicina (Lithuania). 2023;59(4).

(6) Ha VK, Manh XH, Dao VH, Nguyen PB, Hoang BL, Vu H. Cleaning assessment in endoscopic esophageal images using U-Net and a classification model. In: 2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2020. 2020.

(7) Haithami MS, Ahmed A, Liao IY, Altulea HJ. Automatic Bowel Preparation Assessment Using Deep Learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2023.

(8) Tan M, Le Q V. EfficientNetV2: Smaller Models and Faster Training. In: Proceedings of Machine Learning Research. 2021.

Counterfactuals: The impact of image properties on the quality of generated explanations in XAI

Author
Daniel Nguyen – Department of Computing, Imperial College London, UK
Ahmed E. Fetit – Department of Computing, Imperial College London, UK; UKRI CDT in Artificial Intelligence for Healthcare, Imperial College London, UK
Kanwal Bhatia – Aival, London, UK

Citation
Nguyen, D., Fetit, A.E., Bhatia, K. Counterfactuals: The impact of image properties on the quality of generated explanations in XAI.

Contemporary image classification methods based on deep-learning architectures having proven highly effective, but the "thought process" of these algorithms remains opaque. To help understand their decisions, model-agnostic explanation methods, such as counterfactuals, can be used. In this work, we evaluate using cycleGANs to generate counterfactual explanations of classification tasks, and study the effects of different architectures and image properties on the quality of the results. We illustrate our findings using ophthalmic and artificial datasets, demonstrating that both the classification model's architecture and the images' textural and shape properties strongly impact the quality of the generated counterfactuals.

Introduction

Despite its promise, integrating AI into clinical practise comes with risks as incorrect diagnoses can have devastating consequences. To gain insight into

an AI's decision, and therefore trust, there is increasing interest in explainable AI (XAI). Techniques such as Gradient-weighted Class Activation Mapping [1], attempt to explain an AI's underlying "thought process". However, despite providing valuable information, these techniques tend to be model-specific and require access to the model's architecture. Furthermore, while they show which regions are important in the prediction, there is ambiguity about what these features represent. This motivates the need for post-hoc XAI methods which are agnostic to model architecture. One such method is counterfactual explanations which constructs an artificial image closely resembling the original, but with minimal alterations applied, and which produce a different decision by the AI [2]. One method of generating counterfactual images uses cycle generative adversarial networks (GANs) which focuses on training distinct classes [3]. In this approach, each GAN is responsible for translating an image from one domain to another, or, in the context of the classifier, from one prediction to an alternative. The visual disparities between the original image and its counterfactual counterpart serve as the explanation for understanding the classifier's decision-making. Mertes et al. proposed a method to incorporate the classifier into the training of a cycleGAN [4], as without this, the counterfactuals generated can only be explanations of the GAN, and not tuned to the classifier itself. The authors evaluated their method using a dataset where relevant information is only textural. However, questions remain on how the method would generalise to other data such as structural information (e.g., occurrence of objects or changes in the structure of an object). The aim of our paper is to expand on existing experiments to explore how the quality of explanations is affected by image properties. We investigate this approach on retinal fundus photographs and OCT scans, using a variety of classifier architectures, and with varying textural and structural image properties. By studying and comparing the generated counterfactuals, we report unique insights into how the classifier's architecture and the images' textural and structural properties impact the quality of the generated counterfactuals.

Experimental Setup

Methodology. We follow Mertes et al. [4], testing VGG-19 and ResNet-50 models in addition to the AlexNet model used there. Each model was trained on each dataset to 90% accuracy before being used to train the adapted cycleGAN. Compared to AlexNet, VGG uses significantly smaller but more numerous convolutional layers resulting in a smaller but deeper network. Contrastingly, ResNet uses a series of residual blocks and connections to form a more complex model.

Data. Several datasets with varying combinations of image properties were used:

- RSNA Pneumonia [7]: the original dataset and preprocessing used in Mertes et al. [4], representing textural variation between classification groups

- Diabetic Retinopathy (DR) [5]: a simplified binary dataset consisting of no DR and severe-proliferate DR classes (720 clear images each) with similar optical disc orientations. We use only the green channels for training (spatial information).

- Retinal OCT [6]: two binary datasets: a) Diabetic Macular Edema (DME) (10,948 images) and no DME (10,675 images), and b) Drusen (8,616 images) and no Drusen (10,675 images). DME appears on an OCT scan as "tearing" in the retina (i.e., spatial and structural information), while drusen manifests as bumps in-between the membrane and the epithelium (i.e. structural information).

- Toy-Box: a synthetic dataset created as proof-of-concept: 1300 images with a black background and either a randomly placed grey box, or a grey box with a smaller white box inside for the two classes (i.e., spatial + textural information) (Figure 1).

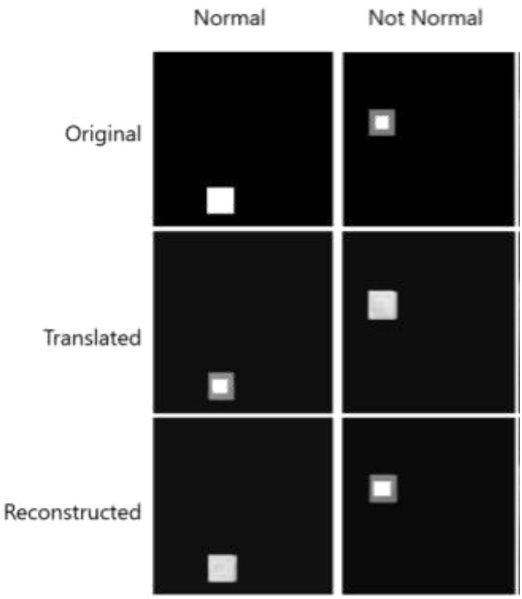


FIGURE 1
A Successfully generated counterfactual pair for both classes for toy box dataset.

Results

For the toy-box dataset, both VGG-19 and AlexNet successfully produced convincing counterfactuals for both classes with VGG creating the most realistic results which also preserved the square shape. ResNet was not used as it resulted in extreme overfitting on this simple dataset. However, for the DR dataset, none of the models was able to produce meaningful counterfactuals.

These counterfactuals were either slightly blurred compared to the original and, in some cases, had introduced and/or exaggerated artifacts that were already present (e.g., due to lighting variations). For the Drusen dataset, the

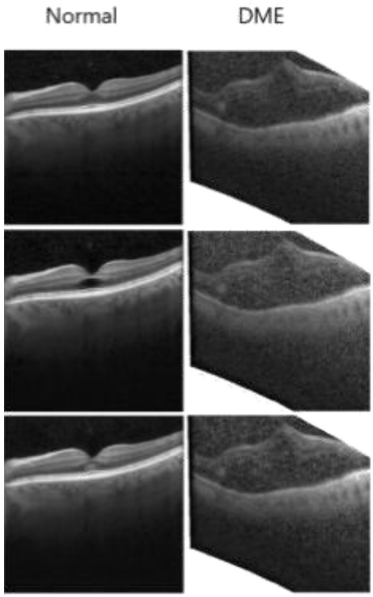


FIGURE 2
A meaningful generated counterfactual for the normal class but a failed explanation for the DME class.

generated counterfactuals were identical to the original images and not meaningful as explanations. For DME, all three models were able to produce an explanation for normal class images with AlexNet producing the most informative explanation. However, counterfactuals generated from DME images did not yield relevant images (Figure 2).

Discussion

We have examined how the properties of a dataset and a model's architecture can impact the quality of its generated outputs. For datasets that express relevant information texturally or spatially as simple general transformations (i.e. the counterfactual class can be represented as two layers consisting of the original image as one layer, and the transformation

that the cycleGAN applies as the second layer), the cycleGAN produces meaningful results. This can be seen in the toy-box, pneumonia and DME datasets. The differences in performance between these datasets is likely due to being able to construct the pneumonia and toy-box examples through one simple consistent transformation across all images, whereas DME can be described as a singular but varying transformation. The DR and Drusen datasets both require several complex transformations to be applied or the modification of the original first layer, respectively, which the cycleGAN fails to learn. In future work, these methods will be extended to 3D medical images.

Acknowledgements
AEF was supported by the UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare in his role as Senior Teaching Fellow (Grant Number: EP/S023283/1). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

References

(1) Ramprasaath R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. Dec. 2019. url: https : / / arxiv.org/abs/1610.02391.

(2) Ruth M. Byrne. "Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning". In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (2019). doi: 10.24963/ijcai.2019/876.

(3) Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: arXiv.org (Aug. 2020), pp. 1–2. url: https://arxiv.org/abs/1703.10593

(4) Silvan Mertes et al. Ganterfactual-counterfactual explanations for medical non-experts using generative adversarial learning. Mar. 2022. url: https://www.frontiersin.org/articles/10.3389/frai.2022.825565/full.

(5) Emma Dugas, Jared, Jorge, Will Cukierski. (2015). Diabetic Retinopathy Detection. Kaggle. https://kaggle.com/competitions/diabetic-retinopathy-detection

(6) Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2

(7) Anouk Stein, MD, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, kalpathy, Leon Chen, Luciano Prevedello, Marc Kohli, MD, Mark McDonald, Peter, Phil Culliton, Safwan Halabi MD, Tian Xia. (2018). RSNA Pneumonia Detection Challenge. Kaggle. https://kaggle.com/competitions/rsna-pneumonia-detection-challenge

Multi-task SwinV2 transformer for polyp classification and segmentation

Author
Kerr Fitzgerald – Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, United Kingdom
Jorge Bernal – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain
Yael Tudela – Computer Vision Center and Computer Science Department, Universitat Autònoma de Barcelona, Barcelona, Spain
Bogdan J. Matuszewski – Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, United Kingdom

Citation
Fitzgerald, K., Bernal, J., Tudela, Y., Matuszewski, B.J. Multi-task SwinV2 transformer for polyp classification and segmentation.

Abstract
Motivated by the aim of improving polyp classification performance on the CVC-HDClassif dataset, joint classification-segmentation multi-task learning using a SwinV2 Transformer UNet based architecture has been explored.

Introduction
The gold standard of polyp screening and removal procedures is widely considered to be colonoscopy, which allows clinicians to navigate through the colon and visually inspect for abnormalities in real time. However, colonoscopy does have limitations as not all polyps are consistently identified

(A.M. Leufkens et al., 2012). Therefore research has focused on developing computer aided systems to support clinicians in improving the detection rate and characterization of polyps, with deep learning systems achieving current state-of-the-art performance across a variety of polyp imaging tasks.

Whilst polyp segmentation models are showing signs of reaching maturity (K. Fitzgerald et al., 2024) (RG. Dumitru, D. Peteleaza & C. Craciun, 2023), polyp classification has been identified as a critical area for further research. One reason for this is the lack of openly available datasets which contain polyp classification labels. The novel CVC-HDClassif dataset (Y. Tudela et al., 2023) contains 788 training, 113 validation, and 225 testing images, with corresponding ground truth segmentation maps and polyp histology labels (adenomatous vs non adenomatous).

System Design and Methodology

Previous models for medical imaging multi-task learning employ UNet (O. Ronneberger et al., 2015) style architectures with the addition of a classification head at a selected stage of the network (B. Oliveira et al., 2023) (C. Li, J. Liu & J. Tang, 2024). Such multitask learning models lead to improved classification performance, which is hypothesized to occur due to the mixing of detailed spatial information needed for segmentation and global contextual information needed for classification. Motivated by the improved classification performance of these models and the excellent performance of SwinV2 systems when used as encoders in segmentation systems, a SwinV2 UNet system (Z. Liu et al., 2022) with a classification head added the to the final encoder layer was developed for joint polyp classification and segmentation. A description of the SwinV2 UNet style architecture can be found in (K. Fitzgerald et al., 2024). The classification head flattens the tensor from the final encoder stage and then passes this sequentially through two Fully Connected (FC) layers using a dropout rate of 50%. A final FC layer and activation function is used to generate a final classification prediction. The architecture of the SwinV2 UNet segmentation-classification model is shown in Figure 1.

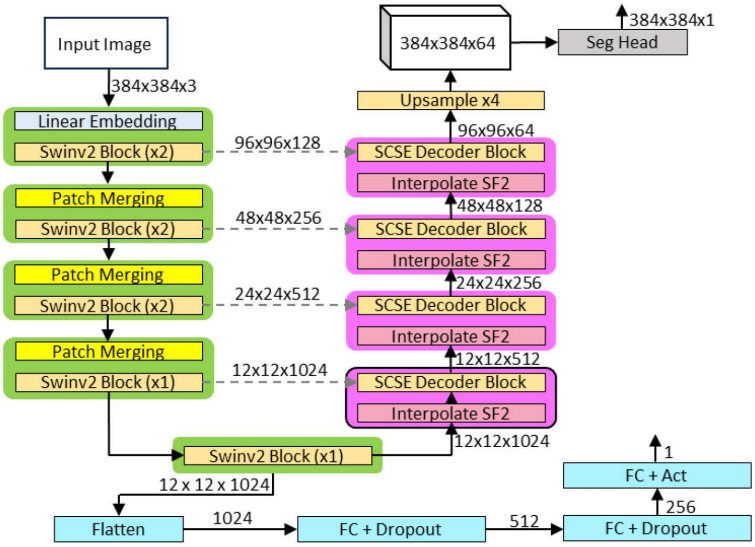


FIGURE 1
SwinV2 UNet system with a classification head added to the final block of the encoder. The decoder block makes use of 'Spatial and Channel Squeeze and Excitation' (SCSE) modules.

The SwinV2 UNet model was implemented using PyTorch and the encoder was initialized using ImageNet-22K (J. Deng et al., 2009) weights available from the PyTorch Image Models Library (R. Wightman, 2019). Since the CVC-HDClassif test split has not been released by the dataset authors, only the training and validation splits were utilized in this study. Due to the relatively small number of images available for training and validation, standard on-the-fly data augmentations (e.g. color variations and geometrical transforms) were applied to the training set using the Albumentations library (A. Buslaev et al., 2018). Static data augmentations were applied to the validation set to stabilize accuracy scores. The AdamW algorithm (I. Loshchilov & F. Hutter, 2019) was used for model optimization alongside a cosine learning rate schedular. To train the SwinV2 UNet model on the CVCHDClassif dataset, a combined segmentationclassification loss function (EQ1) was used.

$$L_{total} = (\gamma_{class} \cdot L_{class}) + (\gamma_{seg} \cdot L_{seg}) \quad \text{EQ1}$$

Where γ_{class} represents a classification weighting factor, L_{class} is the Binary Cross Entropy (BCE) classification loss (E. Bekele & W. Lawson, 2019), γ_{seg} represents a segmentation weighting factor, and L_{seg} represents the segmentation loss which is a combination of the pixel based Binary Cross Entropy (BCE) loss and dice loss (S. Jadon, 2020). The mean training and validation classification losses and accuracies were recorded to examine model performance. Ablation studies showed that setting the classification weight to a very small value (γ_{class} = 1E^{-6}) for the first 15 training epochs allowed the model to achieve strong segmentation performance, before then changing the classification weight to the value of 1 (γ_{class} = 1. The performance of the SwinV2 UNet multitask learning model was also compared to a standard SwinV2 classification model and the fully convolutional EfficienetV2M model (M. Tan & Q. Le, 2021). These models used the same training methodology (excluding the task of segmentation) and required fewer training epochs before signs of overfitting occurred.

Preliminary Results and Discussion

The SwinV2-UNet model shows excellent segmentation performance on the validation set, achieving 90.88 mDice and 85.13 mIoU scores. The training and validation classification losses and accuracies are shown for each model in Figure 2.

For the classification task, the SwinV2-UNet model reached a maximum accuracy of 84.82%, which represents a substantial improvement over the maximum accuracy of 78.86% achieved by the SwinV2 classification model. The SwinV2 classification model is likely overfitting to the training data due to the limited dataset size and network complexity. This highlights the potential for multi-task learning approaches to enhance generalizability performance on classification tasks by leveraging spatial information supplied by segmentation data. The EfficientNetV2 model achieved the highest maximum validation accuracy of 85.42%. The EfficientNetV2 model is likely to offer benefits over Transformer based architectures for small dataset sizes

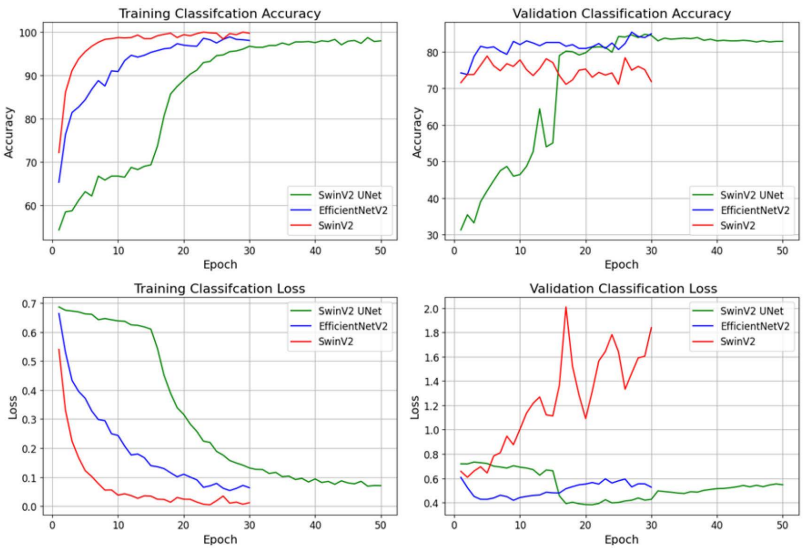


FIGURE 2
Training and validation performance of the models. The increase in classification accuracy at epoch 15 for the SwinV2 UNet model is due to the classification weight change.

due to the inherent inductive biases contained within fully convolutional architectures (A. Dosovitskiy et al., 2021). Further model refinements and larger multitask polyp segmentation-classification datasets will be beneficial to fully investigate and leverage the advantages of multi-task learning frameworks.

References

A. Buslaev et al., 2018. *Albumentations: Fast and Flexible Image Augmentations.* [Online] Available at: ttps://albumentations.ai/[Accessed 16 Janurary 2024].

A. Dosovitskiy et al., 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* Virtual, s.n.

A.G. Roy, N. Navab & C. Wachinger, 2018. *Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks.* s.l., s.n., p. 421–429.

A.M. Leufkens, M.G.H. van Oijen, F.P. Vleggaar & P.D. Siersema, 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy,* 44(5), pp. 470-475.

B. Oliveira et al., 2023. A multi-task convolutional neural network for classification and segmentation of chronic venous disorders. *Nature Scientific Reports.*

D.A. Corley, 2014. Adenoma Detection Rate and Risk of Colorectal Cancer and Death. *New England Journal of Medicine,* Volume 370, pp. 1298-1306.

E. Sanderson & B.J. Matuszewski, 2022. *FCN-Transformer Feature Fusion for Polyp Segmentation.* s.l., s.n., pp. 892-907.

I. Loshchilov & F. Hutter, 2019. *Decoupled Weight Decay Regularization.* New Orleans, s.n.

J. Deng et al., 2009. *ImageNet: A large-scale hierarchical image database.* s.l., s.n.

J. Lee, S.W. Park, Y.S. Kim, K.J. Lee, H.S, P.H. Song, W.J. Yoon & J.S. Moon, 2017. Risk factors of missed colorectal lesions after colonoscopy. *Medicine,* 96(27).

K. Fitzgerald et al., 2024. Polyp Segmentation With the FCB-SwinV2 Transformer. *IEEE Access,* Volume vol. 12, pp., pp. 38927-38943.

M. Abe, 2022. *Swin V2 Unet/Upernet.* [Online]
Available at: https://www.kaggle.com/code/abebe9849/swin-v2-unet-upernet [Accessed January 2023].

M. Tan & Q. Le, 2021. *EfficientNetV2: Smaller Models and Faster Training.* s.l., s.n.

M.J. Whitson, C.A. Bodian, J. Aisenberg & L.B. Cohen, 2012. Is production pressure jeopardizing the quality of colonoscopy? A survey of U.S. endoscopists' practices and perceptions. *Gastrointestinal Endoscopy,* 75(3), pp. 641-648.

N.H. Kim, Y.S. Jung, W.S. Jeong, H.J. Yang, S.K. Park, K. Choi & D.I. Park, 2017. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal Research,* 15(3), pp. 411-418.

O. Ronneberger et al., 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* s.l., s.n., pp. 234-241.

P. Iakubovskii, 2019. Segmentation Models Pytorch. *GitHub Repository.*

R. Wightman, 2019. *PyTorch Image Models.* s.l., s.n.

RG. Dumitru, D. Peteleaza & C. Craciun, 2023. Using DUCK-Net for polyp image segmentation. *Nature Scientific Reports,* Volume 13.

World Health Organization, 2023. *Colorectal cancer.* [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer [Accessed December 2023].

Y. Tudela et al., 2023. *Towards Fine-Grained Polyp Segmentation and Classification.* s.l., Cham: Springer Nature, pp. 32-42.

Z. Liu et al., 2022. *Swin Transformer V2: Scaling Up Capacity and Resolution.* s.l., s.n., pp. 12009-12019.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin & B. Guo, 2021. *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.* s.l., s.n., pp. 10012-10022.

Polyp segmentation generalisability of pretrained backbones

Author
Edward Sanderson, Bogdan J. Matuszewski – Computer Vision and Machine Learning (CVML) Group, School of Engineering, University of Central Lancashire, Preston, UK Institute

Citation
Sanderson, E., Matuszewski, B.J. Polyp segmentation generalisability of pretrained backbones.

Introduction

Due to the low availability of annotated data for training polyp segmentation models, e.g. Sanderson and Matuszewski (2022), which typically take the form of an autoencoder with UNet-style skip connections (Ronneberger et al., 2015), it is common practice to pretrain the encoder, also known as the backbone. This has almost exclusively been done in a supervised manner with ImageNet-1k (Deng et al., 2009). However, we recently demonstrated that pretraining backbones in a self-supervised manner generally provides better fine-tuned performance, and that models with ViT-B (Dosovitskiy et al., 2020) backbones typically perform better than models with ResNet50 (He et al., 2016) backbones (Sanderson and Matuszewski, 2024).

In this paper, we extend this work to consider generalisability. I.e., we assess performance on a different dataset to that used for fine-tuning, accounting for variation in network architecture and pretraining pipeline (algorithm

and dataset). This reveals how well models generalise to a somewhat different distribution to the training data, which arise in deployment as a result of different cameras, demographics of patients, and other factors. Our results provide further insights into the strengths and weaknesses of existing architectures and pretraining pipelines that should inform the future development of polyp segmentation models.

Analysis

We consider 12 polyp segmentation models pretrained and fine-tuned in a previous study (Sanderson and Matuszewski, 2024), specifically those fine-tuned on Kvasir-SEG (Jha et al., 2020). Each model is either a ResNet50 encoder with a DeepLabV3+ (Chen et al., 2018) decoder, or a ViT-B encoder with a DPT (Ranftl et al., 2021) decoder. Additionally, each model was pretrained on either Hyperkvasir-unlabelled (Borgli et al., 2020) or ImageNet-1k in a self-supervised manner using either MoCo v3 (Chen et al., 2021), Barlow Twins (Zbontar et al., 2021) (ResNet50 only), or MAE (He et al., 2022) (ViT-B only); or pretrained in a supervised manner (ImageNet-1k only); or not pretrained at all.

We evaluate performance on the full CVC-ClinicDB dataset (Bernal et al., 2015) with mDice, mIoU, mPrecision, and mRecall. The results are reported in Table 1, where we also specify each model's rank on each metric, as well as any change in rank relative to the model's evaluation on the Kvasir-SEG test set (Sanderson and Matuszewski, 2024). The results show that self-supervised pretraining on ImageNet-1k generally provides the best generalisation, that supervised pretraining on ImageNet-1k is generally better than self-supervised pretraining on Hyperkvasir-unlabelled, and that any considered pretraining is better than no pretraining. These findings are consistent with the evaluation on the Kvasir-SEG test set.

However, the model pretrained with MAE on ImageNet-1k, which performs best on the Kvasir-SEG test set, reduces its rank on every metric, notably dropping from rank 1 to 4 on mDice. In contrast, models with a ResNet50 backbone generally improve their ranking, implying greater generalisability

TABLE 1: Performance of models fine-tuned on the Kvasir-SEG training set and tested on CVC-ClinicDB. In addition to reporting the value of each metric, we also indicate the rank of each model, as well as any change in this rank relative to the model's evaluation on the Kvasir-SEG test set. For conciseness, we abbreviate Hyperkvasir-unlabelled to HK, ImageNet-1k to IN, and Barlow Twins to BT

Backbone	Pretraining		mDice		mIoU		mPrecision		mRecall	
arch.	Data	Algo.	Value	Rank	Value	Rank	Value	Rank	Value	Rank
ResNet50	HK	MoCo v3	0.789	9 (↑1)	0.686	10	0.785	10	0.856	3 (↑7)
		BT	0.801	8 (↑1)	0.709	8 (↑1)	0.831	8 (↑1)	0.848	7 (↓2)
	IN	MoCo v3	0.843	1 (↑3)	0.760	1 (↑3)	0.867	3 (↑5)	0.874	1
		BT	0.826	5	0.735	6 (↑1)	0.858	6	0.854	4 (↑2)
		Supervised	0.822	6	0.735	5	0.899	1 (↑4)	0.811	10 (↓2)
	None	None	0.520	12	0.394	12	0.496	12	0.724	12
ViT-B	HK	MoCo v3	0.789	10 (↓2)	0.696	9 (↓1)	0.812	9 (↓2)	0.848	8 (↓1)
		MAE	0.828	3	0.743	3	0.852	7 (↓4)	0.858	2 (↑1)
	IN	MoCo v3	0.830	2	0.742	4 (↓2)	0.861	4 (↓2)	0.849	5 (↓3)
		MAE	0.827	4 (↓3)	0.746	2 (↓1)	0.868	2 (↓1)	0.848	6 (↓2)
		Supervised	0.809	7	0.717	7 (↓1)	0.860	5 (↓1)	0.832	9
	None	None	0.637	11	0.519	11	0.670	11	0.759	11

than models with a ViT-B backbone, which generally experience a drop in ranking, and the best generalisation is achieved by the model with a ResNet50 backbone that was pretrained on ImageNet-1k using MoCo v3, notably improving from rank 4 to 1 on mDice. To better understand this, we compare the distribution of instance-wise Dice scores from each model's evaluation on the Kvasir-SEG test set against the distribution from its evaluation on CVC-ClinicDB in Fig. 1. This indicates that all models experience a drop in overall performance that primarily arises from a higher variance. However, the portion of each distribution for the highest Dice scores shows that most models with ResNet50 backbones achieve better performance on some instances of CVC-ClinicDB than any in the Kvasir-SEG test set, while models with ViT-B backbones fail to exceed their maximum Dice score across the Kvasir-SEG test set when evaluated on CVC-ClinicDB. We verify that all models experience a drop in performance, and quantify the relative drop, in Fig. 2, which reveals that most models with ResNet50

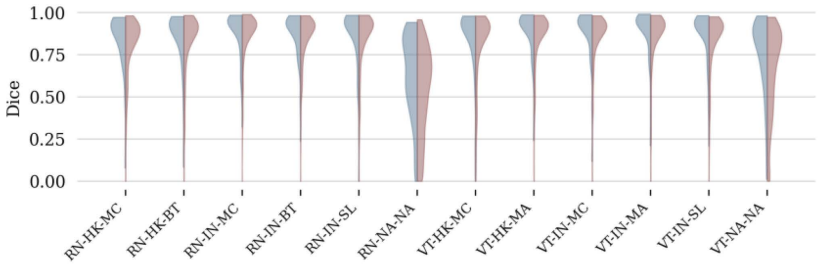


FIGURE 1
Comparison of the distribution of instance-wise Dice score from each model's evaluation on the Kvasir-SEG test set (blue) against the distribution from its evaluation on CVC-ClinicDB (red). For conciseness, we denote ResNet50s with RN, ViT-Bs with VT, Hyperkvasir-unlabelled with HK, ImageNet-1k with IN, MoCo v3 with MC, Barlow Twins with BT, MAE with MA, supervised pretraining with SL, and no pretraining with NA-NA.

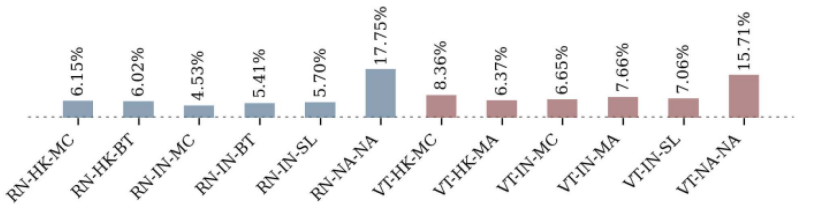


FIGURE 2
Relative drop in mDice from each model's evaluation on the Kvasir-SEG test set to its evaluation on CVC-ClinicDB. For conciseness, we denote ResNet50s with RN, ViT-Bs with VT, Hyperkvasir- unlabelled with HK, ImageNet-1k with IN, MoCo v3 with MC, Barlow Twins with BT, MAE with MA, supervised pretraining with SL, and no pretraining with NA-NA. For clarity, the results for ResNet50 models are coloured blue and the results for ViT-B models are coloured red.

backbones do indeed experience less of a drop, potentially as a result of their improvement in maximum Dice score, explaining the improvement in ranking.

Conclusion

In this paper, we showed that previous findings, regarding pretraining pipelines for polyp segmentation, hold true when considering generalisability. However, our results imply that models with ResNet50 backbones typically generalise better, despite being outperformed by models with ViT-B backbones in evaluation on the test set from the same dataset used for fine-tuning. We expect that this is a result of the larger complexity of the models with ViT-B backbones allowing for overfitting on the distribution underlying the training data. However, this challenges the assumption that the considered pretraining pipelines should help prevent this, and more work is required to better understand the relationships between architecture, pretraining pipeline, and performance on different distributions of data, as well as the amount of training data.

References

Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. (2015). Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics 43, 99–111

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., et al. (2020). Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data 7, 283

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV). 801–818

Chen, X., Xie, S., and He, K. (2021). An empirical study of training self-supervised vision transformers. CoRR abs/2104.02057

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (Ieee), 248 255

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 16000–16009

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., et al. (2020). Kvasir-seg: A segmented polyp dataset. In MultiMedia Modeling: 26th International Conference, MMM 2020 Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26 (Springer), 451–462

Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF international conference on computer vision. 12179–12188

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (Springer), 234–241

Sanderson, E. and Matuszewski, B. J. (2022). Fcn-transformer feature fusion for polyp segmentation. In Annual conference on medical image understanding and analysis (Springer), 892–907

Sanderson, E. and Matuszewski, B. J. (2024). A study on self-supervised pretraining for vision problems in gastrointestinal endoscopy. IEEE Access 12, 46181–46201. doi:10.1109/ACCESS.2024.3381517

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (PMLR), 12310–12320

Toward automated small bowel capsule endoscopy reporting using a summarizing machine learning algorithm: The sum up study

Author

Charles Houdeville – Sorbonne University, Center for Digestive Endoscopy, Saint-Antoine Hospital, APHP, 75012 Paris, France; Équipes Traitement de l'Information et Systèmes, ETIS UMR 8051, CY Paris Cergy University, ENSEA, CNRS, 95000 Cergy, France

Marc Souchaud – Équipes Traitement de l'Information et Systèmes, ETIS UMR 8051, CY Paris Cergy University, ENSEA, CNRS, 95000 Cergy, France

Romain Leenhardt – Sorbonne University, Center for Digestive Endoscopy, Saint-Antoine Hospital, APHP, 75012 Paris, France

Lia Goltstein – Department of Gastroenterology and Hepatology, Radboud University Medical Center, Nijmegen, Netherlands

Guillaume Velut – Sorbonne University, Center for Digestive Endoscopy, Saint-Antoine Hospital, APHP, 75012 Paris, France

Hanneke Beaumont – Department of Gastroenterology and Hepatology, Amsterdam University Medical Center, , Amsterdam, Netherlands

Xavier Dray – Sorbonne University, Center for Digestive Endoscopy, Saint-Antoine Hospital, APHP, 75012 Paris, France; Équipes Traitement de l'Information et Systèmes, ETIS UMR 8051, CY Paris Cergy University, ENSEA, CNRS, 95000 Cergy, France ; Equal contributors

Aymeric Histace – Équipes Traitement de l'Information et Systèmes, ETIS UMR 8051, CY Paris Cergy University, ENSEA, CNRS, 95000 Cergy, France ; Equal contributors

Citation
Houdeville, C., Souchaud, M., Leenhardt, R., Goltstein, L., Velut, G., Beaumont, H., Dray, X., Histace, A. Toward automated small bowel capsule endoscopy reporting using a summarizing machine learning algorithm: The sum up study.

Background and Objectives

A capsule endoscope (CE) is a medical device that provides a combination of systems that are biocompatible, waterproof, and miniaturized, aiming to explore the gastrointestinal (GI) tract. Once ingested, this capsule-sized intestinal drone captures an average of 50,000 images. Gastrointestinal angiodysplasias or angiectasias (GIAs) is the most common lesion found in patients with a suspected small bowel bleeding. The "typical" aspect of a GIA is strongly correlated with its increased risk of bleeding, whereas other "atypical" or smaller vascular lesions (such as erythematous patches, phlebectasias, red dots) seem to be at decreased risk. Therefore, physicians need to accurately distinguish GIAs from other vascular lesions. Deep learning (DL) algorithms demonstrate excellent diagnostic performance for the detection of vascular lesions via small bowel (SB) CE, including vascular abnormalities with high (P2), intermediate (P1) or low (P0) bleeding potential (classification by Saurin et al. [1]), while dramatically decreasing the reading time. We aimed to improve the performance of a DL algorithm by characterizing vascular abnormalities and selecting the most relevant images for insertion into reports.

Materials and Methods

A training database of 75 SB CE videos was created, containing 401 sequences of interest that encompassed 1,525 images of various vascular lesions. All sequences of interest were then processed by the DL-based algorithm [2] (**Figure 1**) for automated selection of still frames with vascular abnormalities (images of interest), providing a detection frame and a degree of confidence in the proposed diagnosis. All images of interest were examined by an adjudication group, to be sorted according to the P0, the P1, and P2 classifications, to create two datasets. The first dataset, named

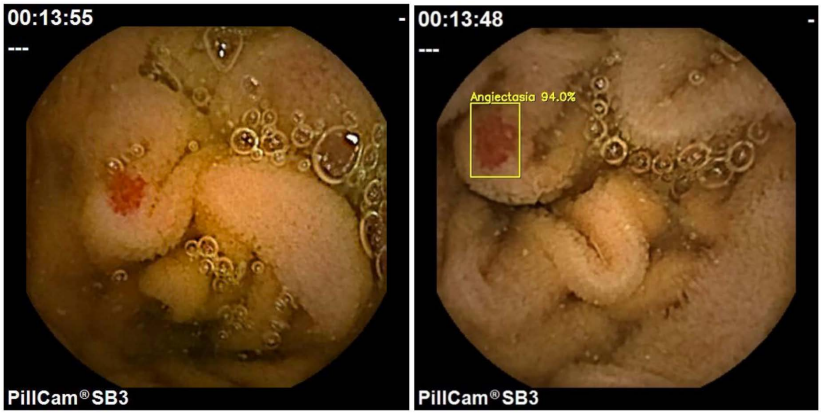


FIGURE 1
Left, capsule endoscopy view of a typical gastrointestinal angiectasia of the small bowel; Right, capsule endoscopy view, with Deep Learning-based detection and delineation of a typical gastrointestinal angiectasia, with a 94.0% of confidence in the diagnosis.

"dataset P2/P1", included all images of interest classified as P1 or P2, and the other group, named "dataset P0", included P0 images. These two datasets were used for training the algorithm.

With this limited in size training dataset, several machine learning image classification algorithms were tested, to discriminate "typical angiodysplasia" (P2/P1) and "other vascular lesion" (P0) and to select the most relevant image within sequences with repetitive images.

The performances of the best-fitting algorithms were subsequently assessed on an independent database of 73 full-length SB CE video recordings (**Figure 2**).

Then, the best-fitting algorithm was enhanced to select the most relevant frame within the sequence of interest where the image originally belonged. The diagnostic performance of it was assessed for selecting the most relevant frame on the testing subset, with the experts' opinions serving as a reference.

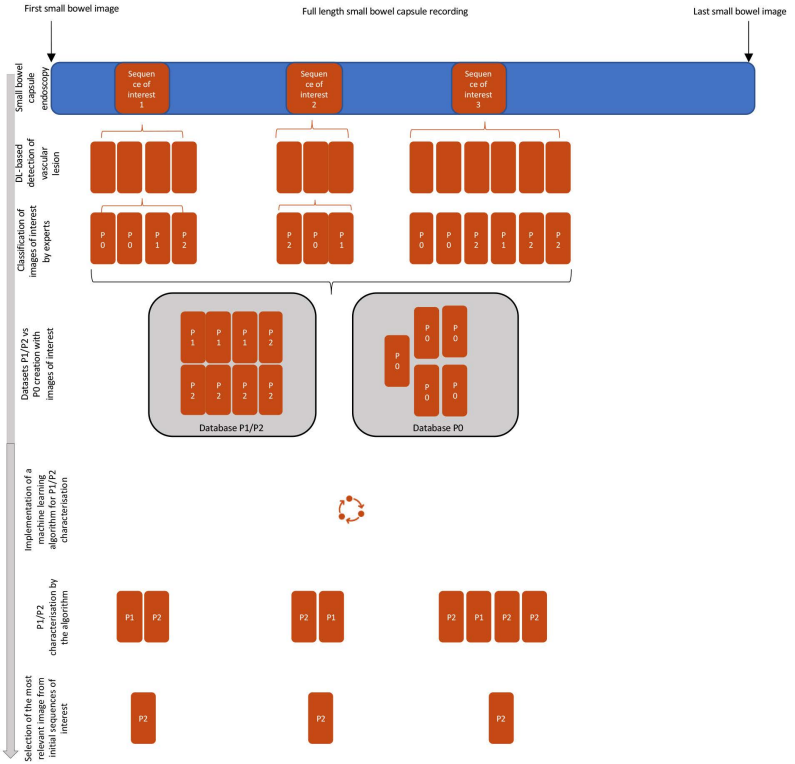


FIGURE 2

Illustration of the methodology used in the study. In this example, three sequences of interest with vascular lesions are extracted from a full-length SB CE recording. A DL solution detects some images of interest (with vascular abnormalities) within those sequences of interest. These images of interest are sorted out by experts to create a "P2/P1" dataset" (containing pattern associated with intermediate or high risk of gastrointestinal bleeding), and a "P0" dataset (containing pattern with low risk for gastrointestinal bleeding). The datasets are used to train and test a ML algorithm which aims to characterise P2/P1 (high risk) only vascular lesions, and to select the most relevant vascular lesion within a sequence of interest.

Results
Following DL detection, a random forest (RF) method demonstrated a specificity of 91.1%, an area under the receiving operating characteristic curve of 0.873, and an accuracy of 84.2% for discriminating P2/P1 from P0 lesions while allowing an 83.2% reduction in the number of reported images. In the independent testing database, after RF was applied, the output number decreased by 91.6%, from 216 (IQR 108-432) to 12 (IQR 5-33). The RF algorithm achieved 98% agreement with initial, conventional (human) reporting.

Conclusion
Following DL detection, the RF method allowed better characterization and accurate selection of images of relevant (P2/P1) SB vascular abnormalities for CE reporting without impairing diagnostic accuracy. As the otherwise detected (but not selected as "most relevant") images are not deleted, the human reader can still review them when appropriate. This better characterization of images makes it possible to obtain a diagnosis more quickly without the reader necessarily having to re-examine a large quantity of images.

This proof-of-concept study opens the path to even faster readings of CE recordings and to semiautomated reporting.

References

[1] Saurin JC, Pioche M. Why should we systematically specify the clinical relevance of images observed at capsule endoscopy? Endosc Int Open 2014;2:E88–9. https://doi.org/10.1055/s-0034-1377264.

[2] Leenhardt R, Vasseur P, Li C, Saurin JC, Rahmi G, Cholet F, et al. A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. Gastrointestinal Endoscopy 2019;89:189–94. https://doi.org/10.1016/j.gie.2018.06.036.

Where scientists empower society

Creating solutions for healthy lives on a healthy planet

frontiersin.org

Why publish with us?

All Frontiers journals are fully open access, meaning every research article we publish is immediately and permanently free to read.

Our collaborative peer review is handled by experts in the field who objectively certify the quality, validity, and rigor of research.

With the latest custom-built technology and artificial intelligence, we're revolutionizing the way research is published, evaluated, and communicated.

We are the third most-cited publisher with 1.4 billion article views and downloads – reflecting the power of research that is open for all.

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Swizerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact