



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Independent Testing of a Publicly Available CNN tool for Hippocampal Image Segmentation

Citation for published version:

Snedden, F, Ferguson, K, Muñoz Maniega, S, Valdés Hernández, MC & Wardlaw, JM 2023, Independent Testing of a Publicly Available CNN tool for Hippocampal Image Segmentation. in *27th Conference on Medical Image Understanding and Analysis 2023*. Frontiers in Medical Technology, Frontiers Media SA.

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

27th Conference on Medical Image Understanding and Analysis 2023

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.





27th Conference on Medical Image Understanding and Analysis 2023

Foresterhill, Aberdeen,
Scotland

27th Conference on Medical Image Understanding and Analysis 2023

ISBN

9782832512319

DOI

10.3389/978-2-8325-1231-9

Citation

Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T.,
Gordon, S., McGowan, J., Nicolson, C. (2023)

July 19–21, 2023, Foresterhill, Aberdeen, Scotland

The abstracts in this collection have not been subject to any Frontiers peer review or checks, and are not endorsed by Frontiers. They are made available through the Frontiers publishing platform as a service to conference organizers and presenters. The copyright in the individual abstracts is owned by the author of each abstract or their employer unless otherwise stated. Each abstract, as well as the collection of abstracts, are published under a Creative Commons CC-BY 4.0 (attribution) licence (creativecommons.org/licenses/by/4.0/) and may thus be reproduced, translated, adapted and be the subject of derivative works provided the authors and Frontiers are attributed.

For Frontiers' terms and conditions please see: frontiersin.org/legal/terms-and-conditions.

Table of contents

- 9 **Welcome to the 27th Conference on Medical Image Understanding and Analysis 2023**
- 11 **Predicting glioma tumor growth with denoise diffusion probabilistic models**
Qinghui Liu, Elies Fuster-Garcia, Ivar Thokle Hovden, Donatas Sederevičius, Karoline Skogen, Bradley J MacIntosh, Till Schellhorn, Petter Brandal, Atle Bjørnerud, Kyrre Eeg Emblem
- 16 **Class imbalanced histological datasets segmentation: Contribution of transfer learning and active learning**
Ramzi Hamdi, Ruben Grousset, Thierry Delzescaux, Cédric Clouchoux, Kevin Francois-Bouaou
- 22 **Predicting progression levels of mild cognitive impairment**
Misgina Tsighe Hagos, Niamh Belton, Ronan P. Killeen, Kathleen M. Curran, Brian Mac Namee
- 30 **GAN-GA: A generative model based on genetic algorithm for medical image generation**
Mustafa AbdulRazek, Ghada Khoriba, Mohammed Belal
- 40 **Deformable image registration in the presence of simulated metal artefacts on head and neck CT scans**
Yanni Papastavrou, Tryphon Lambrou, Brian Hutton

- 49 **Masked autofocusing: CNN enabled targeted motion estimation and correction in MRI scans**
Ziad Al-Haj Hemidi, Christian Weihsbach, Mattias P. Heinrich
- 55 **NSCLC radiogenomics, lung nodules segmentation and prediction of EGFR mutation status from CT scans**
Ivo Gollini Navarrete, Mohammad Yaqub
- 62 **Domain-specific interpretable AI for burn wound depth prediction using GPT-4**
Xinwei Zhang, Maxwell Jacobson, Mohamed El Masry, Surya Gnyawali, Yexiang Xue, Gayle Gordillo, Juan Wachs
- 70 **AcquisitionFocus: Slicing optimization for fast cardiac MRI**
Christian Weihsbach, Nora Vogt, Ziad Al-Haj Hemidi, Alexander Bigalke, Lasse Hansen, Mattias Heinrich
- 76 **Automated segmentation of rheumatoid arthritis immunohistochemistry stained synovial tissue**
Amaya Gallagher-Syed, Abbas Khan, Felice Rivellese, Costantino Pitzalis, Myles J. Lewis, Gregory Slabaugh, Michael R. Barnes
- 86 **Deep complex-valued edge attention network for artefact removal in cardiovascular magnetic resonance with undersampling spiral trajectories**
Yaqing Luo, Pedro F. Ferreira, Dudley J. Pennell, Guang Yang, Sonia Nielles-Vallespin, Andrew D. Scott
- 97 **Variation in mammography imaging equipment impacts artificial intelligence performance in breast cancer screening**
Clarisse F. de Vries, Samantha J. Colosimo, Roger T. Staff, Jaroslaw A. Dymiter, Joseph Yearsley, Deirdre Dinneen, Moragh Boyle, David J. Harrison, Lesley A. Anderson, Gerald Lip

- 102 **Ensembles-based active learning for left ventricle segmentation**
Eman Alajrami, Jevgeni Jevsikov, Preshen Naidoo, Sara Adibzadeh, Patricia Fernandes, Nasim Dadashi Serej, Neda Azarmehr, Fateme Dinmohammadi, Massoud Zolgharni
- 108 **Weakly supervised pre-training for brain tumour segmentation using principal axis measurements of tumour burden**
Joshua Mckone, Tryphon Lambrou, Xujiang Ye, James Brown
- 114 **Enhancing generalization of CNN models for breast lesion classification from ultrasound images**
Tahir Hassan, Hongbo Du, Sabah Jassim
- 121 **Investigation of the structural characteristics of the extracellular matrix**
Youssef Arafat, Cristina Cuesta Aposua, Esther Castellano, Constantino Carlos Reyes-Aldasoro
- 127 **Assessing robustness of network-based correlation analysis with preclinical total-body PET data**
Abigail F. Hellman, Paul S. Clegg, Adriana A. S. Tavares
- 132 **Towards automatic scoring of spinal X-ray for ankylosing spondylitis**
Yuanhan Mo, Yao Chen, Aimee Readie, Gregory Ligozio, Thibaud Coroller, Bart-lomiej W. Papiez
- 138 **Scottish Medical Imaging (SMI) – Providing safe and secure access to research-ready, population-scale health and imaging data**
Susan Krueger, Jacqueline Caldwell, Ruairidh MacLeod, Bianca Prodan, Andrew Brooks, Smarti Reel, Laura Moran, Kara Moraw, Guneet Kaur, James Sutherland, Emily Jefferson

- 144 **Text-based medical image classification by body part**
Bianca Prodan, Laura Moran, Susan Krueger, Emily Jefferson
- 149 **Natural language process of radiology reports**
Andrew Brooks, Honghan Wu
- 153 **Medical image anonymisation**
Andrew Brooks, Guneet Kaur
- 158 **Automated segmentation of cerebral small vessel disease from field-cycling MRI**
Nicholas Senn, Vasiliki Mallikourti, P. James Ross, Lionel M. Broche, Gordon D Waiter, Mary-Joan MacLeod
- 163 **^{68}Ga FAPI imaging in cancer diagnosis: A promising approach for targeted molecular imaging**
Sidharth Vinod
- 168 **Cubic Bézier curve approximation for the estimation of Perivascular spaces measurements in MRI brain scans**
Roberto Duarte Coello, Maria Valdés Hernández, José Bernal Moyano, Joanna Wardlaw
- 174 **Independent testing of a publicly available CNN tool for hippocampal image segmentation**
FN Sneden, KJ Ferguson, S Muñoz Maniega, M Valdés Hernández, JM Wardlaw
- 179 **Identifying MRI sequence type from pixel data to enable cohort building from routinely collected brain scans**
Smarti Reel, Esma Mansouri-Benssassi, Kara Moraw, Susan Krueger, Emily Jefferson

- 185 **Improving venous tumour thrombus segmentation in clear cell renal cell cancer MRI Scans with a two-stage 3D nnU-Net**
Robin Haljak, Hanna Wyciszczok, Ines P. Machado, Grant D. Stewart,
James O. Jones, Stephan Ursprung, Ferdia A. Gallagher,
Mireia Crispin-Ortuzar
- 192 **XGBoost classifier-based survival prediction in head and neck cancer patients using pre-treatment PET images**
Mahima Philip, Jessica Watts, Andy Welch, Fergus McKiddie, Mintu Nath
- 199 **Prediction of cystic evolution of tumours in ovarian cancer with CT-derived features**
Maria Delgado-Ortet, Leonardo Rundo, Ramona Woitek, Evis Sala,
Lorena Escudero Sánchez
- 206 **Cross-attention multiple instance learning for interpretable whole slide image classification**
Thomas Allcock, Andy Bulpitt, Andrew Hanby, Rebecca Millican-Slater
- 213 **Investigating the effect of self-supervised contrastive learning on mitosis classification**
Trinh Thi Le Vuong, Mostafa Jahanifar, Neda Zamanitajeddin, Jin Tae Kwak,
Nasir Rajpoot
- 221 **Investigation of topological features of the 3D cellular nuclear envelope as observed with electron microscopy**
Kokeb Dese, Cefa Karabağ, Panos Giannopoulos, Constantino Carlos
Reyes-Aldasoro
- 227 **On generalisability of segment anything model for nuclear instance segmentation in histology images**
Kesi Xu, Lea Goetz, Nasir Rajpoot

- 235 **On enhancing the robustness of vision transformers in medical imaging: Defensive diffusion**
Raza Imam, Muhammad Huzaifa, Mohammed El-Amine Azz
- 242 **Stain-invariant representation for tissue classification in histology images**
Manahil Raza, Saad Bashir, Talha Qaiser, Nasir Rajpoot
- 251 **Radiology report generation using multi-layer visual representation**
Chenyu Wang, Stephen McKenna, Vladimir Janjic
- 257 **Multi-lesion segmentation for diabetic retinopathy**
Mohammed Ali Athar, Kashif Rajpoot
- 263 **The prevalence and association of coronary artery calcification in patients with chronic obstructive pulmonary disease: A systematic review and meta-analysis**
Khalid Hakami, Mohmmad Alghamdi, Abdulmalik Arab, James Chalmers, Faisal Khan

Welcome to the 27th Conference on Medical Image Understanding and Analysis 2023

MIUA 2023 is a UK-based international conference for the communication of image processing and analysis research and its application to medical imaging and biomedicine. This was a rapidly growing subject with ever increasing real world applicability. MIUA 2023 was organised by a broad team at the University of Aberdeen including representatives from Aberdeen Biomedical Imaging Centre; Aberdeen Centre for Health Data Science and the School of Natural and Computing Science.

LIST OF ORGANIZERS

Dr Gordon Waiter

Dr Tryphon Lambrou

Prof Georgios Leontidis

Prof Nir Oren

Teresa Morris

Dr Sharon Gordon

Jade McGowan

Cara Nicolson

Predicting glioma tumor growth with denoise diffusion probabilistic models*

Author

Qinghui Liu – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Elies Fuster–Garcia – Instituto Universitario de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, 46022 Valencia, Spain

Ivar Thokle Hovden – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Donatas Sederevičius – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Karoline Skogen – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Bradley J MacIntosh – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Till Schellhorn – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Petter Brandal – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Atle Bjørnerud – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Kyrre Eeg Emblem – Oslo University Hospital (OUS), Rikshospitalet, 0372 Oslo, Norway

Citation

Liu, Q., Fuster-Garcia, E., Thokle Hovden, I., Sederevičius, D., Skogen, K., MacIntosh, B.J., Schellhorn, T., Brandal, P., Bjørnerud, A., Emblem, K.E. Predicting glioma tumor growth with denoise diffusion probabilistic models.

*We gratefully acknowledge support from South-Eastern Norway Regional Health Authority [2017073, 2013069, 2021057];

Abstract

We propose a new generative model that, given past multi-modal Magnetic Resonance Images (MRI) data with different glioma therapies and exam times, can produce realistic MR images that reflect tumor growth forecasts. We developed this model by extending denoising diffusion probabilistic models (DDPMs) with timing and therapy variables as conditional inputs. We then trained the model on real-world postoperative longitudinal MRI data with treatment information from various exam time series. The model has demonstrated promising performance across a range of tasks, including tumor segmentation, growth prediction, uncertainty estimation, and generation of high-quality synthetic multi-modal MR images. Combined with the synthesized MR images, tumor growth predictions with uncertainty estimates can provide useful information for clinical decision-making.

Dataset

One-hundred and twenty-seven MRI exams from 23 patients with histologically confirmed high-grade glioma treated at our institution were included in this study [1]. Patients received treatment based on standard protocols for high-grade glioma, including surgery, followed by fractionated radiotherapy approximately four weeks after surgery with concomitant and/or adjuvant chemotherapy (CRT) with temozolomide (TMZ) [2].

Methods and results

The proposed conditional DDPM [3] network incorporates a conditional input encoder into each U-Net layer, which involves summing the treatment and day intervals features (between the target day and each reference day, up to three reference exams). Furthermore, the reference MRI exams (e.g., T1/T1c/Flair at Day 0, 15, etc, up to three sessions) were concatenated into the Gaussian noise, while the corresponding tumor labels were added to the noise. Finally, the model directly generates MR images for the target days while generating tumor masks using DDPM sampling algorithms. The overall concept of our conditional DDPM U-Net model is depicted in Figure 1.

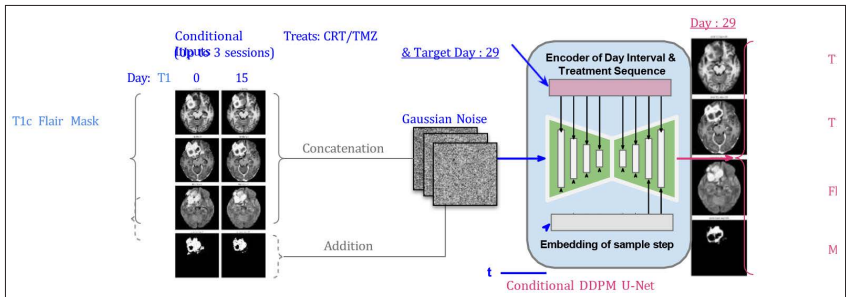


FIGURE 1

The overall concept of our conditional DDPM U-Net model. The purpose of this approach is to generate a set of synthetic MR images for a future time point.

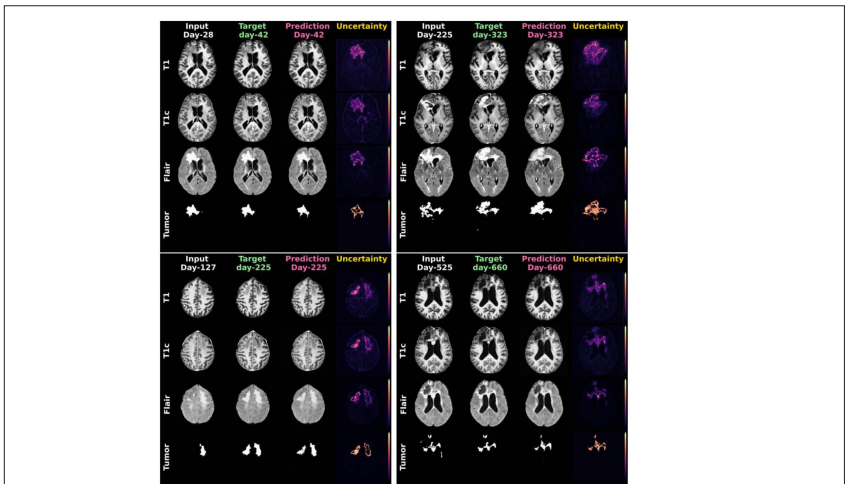


FIGURE 2

Visualization of our model's predictions of tumor masks and generations of 3-modal MR images on the test set, along with the uncertainty maps.

We trained the model on 20 patients (including 190 longitudinal exams) and tested it on 3 patients (including 37 longitudinal exams). Figure 2 shows the qualitative visualization of our model's predictions of tumor masks (Dice score: 0.861 ± 0.05) and generations of 3-modal MR images (SSIM [4] score: 0.853 ± 0.07) on the test set with different target times and therapies. Note that the uncertainty maps were created by computing the standard deviation of our model's predictions on input images with six-time samplings.

Conclusions

We present a novel method for predicting glioma tumor growth and generating multi-modal MR images by incorporating treatment and timing as conditional inputs to DDPM. Using the stochastic sampling process, our model allows for the implicit assembly of various predictions and generations from the same conditional inputs, allowing us to generate uncertainty maps by computing the variance of multiple tumor growth predictions. This approach has the potential in both predicting future tumor growth patterns and quantifying tumor evolution uncertainty - thereby improving treatment planning and ultimately patient outcomes.

References

- [1] Larsson, C., Groote, I., Vardal, J., Kleppestø, M., Odland, A., Brandal, P., Due- Tønnessen, P., Holme, S.S., Hope, T.R., Meling, T.R., et al.: Prediction of Survival and Progression in Glioblastoma Patients Using Temporal Perfusion Changes During Radiochemotherapy. *Magnetic Resonance Imaging* 68, 106–112 (2020)
- [2] Fuster-Garcia, E., Thokle Hovden, I., Fløgstad Svensson, S., Larsson, C., Vardal, J., Bjørnerud, A., Emblem, K.E.: Quantification of Tissue Compression Identifies High-Grade Glioma Patients with Reduced Survival. *Cancers* 14(7), 1725 (2022)

[3] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33, 6840–6851 (2020)

[4] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13(4), 600–612 (2004)

Class imbalanced histological datasets segmentation: Contribution of transfer learning and active learning

Author

Ramzi Hamdi – Witsee, Paris, France

Ruben Grousset – Witsee, Paris, France

Thierry Delzescaux – Université Paris–Saclay, CEA, CNRS, MIRCen, Laboratoire des Maladies Neurodégénératives, Fontenay–aux–Roses, France

Cédric Clouchoux – Witsee, Paris, France

Kevin Francois-Bouaou – Witsee, Paris, France

Citation

Hamdi, R., Grousset, R., Delzescaux, T., Clouchoux, C., Francois, K. Class imbalanced histological datasets segmentation: Contribution of transfer learning and active learning.

Aim

Evaluation of the improvement of histological segmentation performance using transfer learning in the active learning process.

Introduction

With digitized images, histological segmentation is an important topic in medical image analysis. Annotation is usually done manually but is time-consuming and impractical for large datasets. Recently, machine learning, especially deep learning (DL), has been increasingly used to automate

segmentation. In order to properly train complex DL models, a large amount of manually annotated images are required to deal with class balance problems [1, 8]. However, the histological imaging and annotation process make scarce publicly available well-balanced learning datasets [2]. Active learning (AL) [3] and transfer learning (TL) [4] are two approaches that can address these challenges. Active learning can iteratively select informative samples to be segmented to improve model performance with fewer labeled examples, while transfer learning can adapt pre-trained models to new tasks with limited supplementary data.

However, there has been limited research on the efficacy of using both of these techniques on histological datasets with various degrees of class imbalance. This abstract explores the effectiveness of adding AL and TL processes on the histological segmentation architecture Unet [5], and evaluates their impact on the segmentation performance.

Methods

Two annotated datasets were used in our study: The first dataset consisting of Neuronal Nuclei (NeuN) stained macaque brain images which marked neurons (512x512 pixels, x20 magnification, 100 images) [6], and the second dataset extracted from a 13.5-month-old Alzheimer's mouse brain stained with BAM10 (amyloid plaques) and a Bluing Reagent counterstain that blued the nuclear chromatin and cell nucleus membranes (512x512 pixels, x20 magnification, 100 images) [7]. The two dataset were manually segmented by experts and split equally in two subsets (train and test datasets). The class imbalance ratio (CIR) [8] was used to define the imbalance degree of each dataset.

The method used by AL was the uncertainty sampling method [9], which has been employed to reduce annotation costs while maintaining state-of-the-art performance. Six different TL pre-trained models were used to finetune the Unet encoder part, ResNet101V2, ResNet152V2, DenseNet201, DenseNet121, VGG16 and VGG19.

To compare the effects of TL and AL on the histological image segmentation using Unet, a protocol was designed and applied on active learning with and without transfer learning. Six rounds of active learning were performed for each protocol, with 4 initial images and one new proposed image per round. The F-score was calculated on a test dataset independent of the protocol process. Another separate model was trained with the entire train dataset of 50 images. The protocol was launched 20 times per method to account for random variations in the processes.

A Student static test [10] was applied on F-score test for active learning process with / without Transfer and the whole dataset trained model.

Results

The student test showed no significant difference between the AL models without TL and the model with the whole train dataset for both datasets with an F-score test of 0.867 ± 0.009 and 0.865 ± 0.009 respectively (p-value = 0.21) for NeuN dataset and

0.802 ± 0.27 and 0.877 ± 0.013 respectively (p-value = 0.24) for BAM10 dataset.

The CIR was evaluated for both dataset, with a value of 4.37 for the NeuN dataset and

19.4 for the BAM10 dataset. The BAM10 can be defined as the most imbalanced dataset.

For the NeuN dataset, all protocols had converged. The student test showed a significative difference between AL with and without TL for two models, The ResNet101V2 (p-value = 0.004, F-score mean = 0.873 ± 0.005) and DenseNet121 (p-value = 0.0033, Fscore mean= 0.875 ± 0.006) (Fig 1).

For the BAM10 dataset, 11 % of protocols had diverged with null F-score. The student test showed no significant differences between the AL without TL and the AL with TL for all selected TL models (Fig 1).

BAM10	AL	AL/ResNet101	AL/ResNet152	AL/DenseNet201	AL/DenseNet121	AL/VGG16	AL/VGG19
AL		0,186	0,577	0,528	0,857	0,175	0,298
AL/ResNet101V2	0,186		0,162	0,191	0,107	0,022	0,096
AL/ResNet152V2	0,577	0,162		0,930	0,484	0,089	0,311
AL/DenseNet201	0,528	0,191	0,930		0,366	0,076	0,283
AL/DenseNet121	0,857	0,107	0,484	0,366		0,124	0,456
AL/VGG16	0,175	0,022	0,089	0,076	0,124		0,395
AL/VGG19	0,298	0,096	0,311	0,283	0,456	0,395	

NeuN	AL	AL/ResNet101	AL/ResNet152	AL/DenseNet201	AL/DenseNet121	AL/VGG16	AL/VGG19
AL		0,004	0,158	0,931	0,003	0,001	0,826
AL/ResNet101	0,004		<0.001	0,101	0,237	<0.001	<0.001
AL/ResNet152	0,158	<0.001		0,433	<0.001	0,002	0,213
AL/DenseNet201	0,931	0,101	0,433		0,040	0,058	0,795
AL/DenseNet121	0,003	0,237	0,000	0,040		<0.001	<0.001
AL/VGG16	0,001	<0.001	0,002	0,058	<0.001		<0.001
AL/VGG19	0,826	<0.001	0,213	0,795	<0.001	<0.001	

FIGURE 1

Matrices of Student test p-value matrix for significance differences between active learning with and without transfer learning. (Top: highly unbalanced BAM10 dataset; Down: moderately unbalanced NeuN dataset).

Conclusion

To resume, active learning can provide significant improvement in histological image segmentation results by reducing the amount of segmented data. The AL general process was validated against the model trained with the whole train dataset, showing no statistical difference on the validation dataset in segmentation quality. Two datasets were used with different degrees of class unbalanced. For the highly unbalanced BAM10 dataset, the adding of TL to the process did not improve the performance

compared to only AL. A possible explanation was that the amount of relevant annotated data was too small, and the TL did not help the selection of relevant images in this case. For the balanced dataset, the ResNet101V2 [11] and DenseNet121 [12] models significantly improved the selection of the final validation score. In case of a large amount of relevant images, the use of TL can help the AL selection process and increase the model performance. Finally, failed training was reported for the highly unbalanced dataset but not for the moderately unbalanced dataset. The reason for this outcome could be attributed to the initialization process. In a dataset with a significant imbalance, certain images with a high CIR may have been chosen for initialization, thereby impacting the overall process.

In conclusion, our study showed active learning can efficiently select informative samples for high segmentation performance with fewer labeled examples, while transfer learning can take advantage of training models efficiently with a large amount of relevant segmentation.

Future works will be led to find the optimum degree (CIR) of data class imbalance to apply AL and TL. Additionally, novel approaches combining the two methods and other techniques such as domain adaptation and data augmentation could improve the performance for imbalance datasets.

References

- [1] Litjens, Geert, et al. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.
- [2] Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." *Journal of Big Data* 6.1 (2019): 1-54.

[3] Yang, Lin, et al. "Suggestive annotation: A deep active learning framework for biomedical image segmentation." *Medical Image Computing and Computer Assisted Intervention– MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III* 20. Springer International Publishing, 2017.

[4] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." *IEEE Transactions on knowledge and data engineering* 22.10 (2010): 1345-1359.

[5] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18. Springer International Publishing, 2015.

[6] You, Zhenzhen, et al. "Automated individualization of size-varying and touching neurons in macaque cerebral microscopic images." *Frontiers in Neuroanatomy* 13 (2019): 98.

[7] Vandenberghe, Michel E., et al. "High-throughput 3D whole-brain quantitative histopathology in rodents." *Scientific reports* 6.1 (2016): 1-12.

[8] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." *IEEE Transactions on knowledge and data engineering* 21.9 (2009): 1263-1284.

[9] Settles, Burr. "Active learning literature survey." (2009).

[10] Student. "The probable error of a mean." *Biometrika* 6.1 (1908): 1-25.

[11] He, Kaiming, et al. "Identity mappings in deep residual networks." *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV* 14. Springer International Publishing, 2016.

[12] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Predicting progression levels of mild cognitive impairment

Author

Misgina Tsighe Hagos – Science Foundation Ireland Centre for Research Training in Machine Learning; School of Computer Science

Niamh Belton – Science Foundation Ireland Centre for Research Training in Machine Learning; School of Medicine, University College Dublin

Ronan P. Killeen – School of Medicine, University College Dublin; Department of Radiology; UCD–SVUH PET CT Research Centre, St Vincent’s University Hospital, Dublin 4, Ireland

Kathleen M. Curran – Science Foundation Ireland Centre for Research Training in Machine Learning; School of Medicine, University College Dublin

Brian Mac Namee – Science Foundation Ireland Centre for Research Training in Machine Learning; School of Computer Science

Citation

Hagos, M.T., Belton, N., Killeen, R.P., Curran, K.M., Namee, B.M. Predicting progression levels of mild cognitive impairment.

Abstract

Early detection of Alzheimer’s Disease (AD), which is preceded by Mild Cognitive Impairment (MCI), is crucial for making treatment decisions of the disease. However, most of the literature on its early detection focuses on categorizing it into one of the three stages: Normal, MCI, and AD. Since this could miss to accurately identify the trajectory of patients, we revisit the AD identification task and re-frame it as an ordinal classification task to predict how close a patient is to the severe AD stage. We prepare an ordinal dataset of MCI patients with a prediction target that indicates the time to progression to AD and train two versions of Siamese networks using MRI images. Our

evaluations show that Siamese networks bring considerable performance gain at predicting how close input brain MRI images are to progressing to AD.

Introduction

Methods ranging from those that require intensive expert input (Dashjams, et al. 2012) to more automated solutions that utilize deep learning (Basaia, et al. 2019) have been proposed in the diagnosis of Alzheimer's Disease (AD) literature. These methods usually classify patients into one of three stages: Normal (patients exhibiting no signs of dementia and no memory complaints), Mild Cognitive Impairment (MCI) (an intermediate state in which a patient's cognitive decline is greater than expected for their age), and full AD (Xiao, et al. 2022). However, a patient's progression from one of the stages to the next can take more than five years—meaning existing solutions would identify patients on the verge of progressing from MCI to AD as MCI because existing training datasets do not take the level of progression of patients into account (Roberts, et al. 2014).

While healthy adult controls progress to AD annually at a maximum rate of 2%, MCI patients progress at a rate of 10%-25% (Grand, Caspar and MacDonald 2011). This necessitates research on identifying MCI subjects at risk of progressing to AD. In a longitudinal study period, participants diagnosed with MCI can be categorized into two: (1) Progressive MCI, which represents participants who were diagnosed with MCI at some stage during the study but were later diagnosed with AD, and (2) Stable MCI, patients who stayed as MCI during the whole study period (Hagos, et al. 2022). In this work, we focus on answering the clinical question "how far away is a progressive MCI patient on their trajectory to AD?" To do this, we propose an ordinal categorization of brain images based on participants' level of progression from MCI to AD. We add ordinal labels to MRI scans of patients with progressive MCI indicating how many years they are from progressing to AD and we construct a dataset of 444 MRI scans from 288 participants with these labels.

In addition to constructing a dataset, we develop a computer assisted approach to identifying an MRI image's progression level. Accurately identifying how far a patient is from progressing to full AD is of paramount importance as this information may enable earlier intervention with medical treatments (Albright 2019). We use Siamese networks due to their ability to handle the class imbalance in the employed dataset (Yang, et al. 2020). We use typical Siamese networks and their weighted variety that uses a loss function tailored to learning to predict input MRI image's likelihood of progression.

Related work

Machine learning techniques such as random forest (Moradi, et al. 2015) and Convolutional Neural Networks (CNNs) have been used to classify between the progressive and stable MCI (Hagos, et al. 2022). 3D brain images are also deployed with 3D CNNs to reduce false positives (Basaia, et al. 2019). However, these approaches can only tell if an MCI patient has a chance of progressing to AD but not how far they are from progression.

(Li, et al. 2020) report that the distance outputs of Siamese networks could be translated to predict disease positions on a severity scale. Although this approach takes the output as severity scale without any prior training on disease severity and does not interpolate ordinal categories within, it does suggest Siamese networks as a promising approach for predicting ordinal progression levels.

Approach

Dataset Preparation

The data used in the experiments described here was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Mueller, et al. 2005). We collected MRI images of 288 progressive MCI participants. We labeled the participants based on their progression levels towards AD with $p \in [0.1, 1.0]$ with a step size = 0.1, where for a single participant, P , $\min(p) = 0.1$ represents the first time P was diagnosed with MCI and P transitions to stage AD at $\max(p) = 1.0$. The distribution of the constructed dataset with corresponding

TABLE 1: Image distribution across progression levels, ρ

ρ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Number of images	4	4	6	10	24	56	172	273	467

levels is shown in Table 1 (where $\rho = 1.0$ represents AD) where the imbalance between the different classes is clear. Within the collected dataset, the maximum number of MRI scans that the progressive MCI participants have had until they progressed to AD ($\rho = 1.0$) is 9, which means that the smallest ρ is 0.2. By sub-sampling from the majority classes, we selected 444 3D MRI images (shape = 160x192x192) for the negative, anchor, and positive subsets (each holding 148 images) required when training a Siamese network using triplet loss. We used 80% of the images for training and the rest for testing. AD images were randomly separated to the anchor and positive dataset.

Weighted Siamese Networks

While typical Siamese networks, which we refer to as Unweighted Siamese, use Eq. 1 as a loss to train a model that distances between embedding of anchor and positive instances should be much smaller than distances between embedding of anchor and negative instances, the Weighted variant of Siamese takes progression levels into consideration and teaches a learner to distance negative instances from an anchor based on their progression levels. Weighted Siamese achieves this using a weighting factor α , where $\alpha = 1.9 - \rho$, excluding $\rho = 1.0$ (Eq. 2).

$$L_w = \max(d_{ap} - d_{an} + \text{margin}, 0) \quad \text{Equation 1}$$

$$L_w = \max(d_{ap} - \alpha * d_{an} + \text{margin}, 0) \quad \text{Equation 2}$$

Training and Evaluation

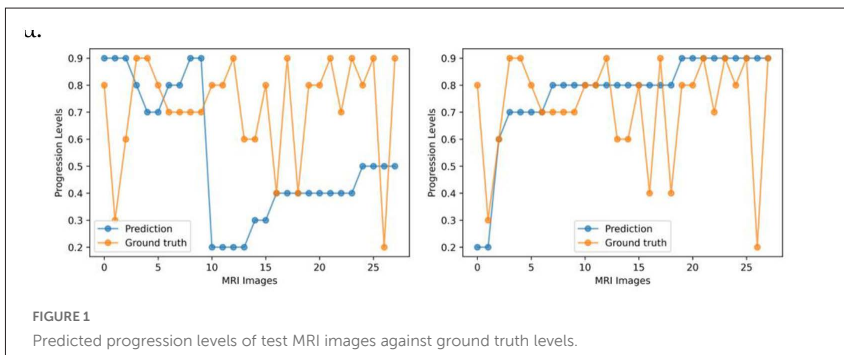
We implemented all of our experiments using TensorFlow and Keras. After comparing performance between different architectures and feature embedding size, we chose to train a 3D ResNet-50 model from scratch by

adding three fully connected layers of sizes 64, 32, and 8 nodes with ReLU activations, taking the last layer of size 8 as the embedding space. We used an Adam optimizer with a decaying learning rate of $1e-3$.

Results and Conclusion

We found that the Weighted Siamese was better at fitting to all the progression levels. While the average training and testing losses of the Weighted Siamese are 2.92 and 2.79, the Unweighted Siamese achieves 10.02 and 17.53, respectively. In addition, while the Weighted Siamese scored a Root Mean Square Error (RMSE) of 2.40, the Unweighted Siamese achieved RMSE of 2.94. We credit this to the effects of adding a weighing factor using α .

A plot of predicted vs. ground truth progression levels of progressive MCI participants is presented in Fig. 1. Our proposed Weighted Siamese network outperforms the Unweighted Siamese network at predicting progression levels (Fig. 1 right vs Fig. 1 left). We observed that the simple modification of factoring the distance between an embedding of anchor and negative instances by a function of the progression level brought considerable performance gain in separating between the interpolated progression levels.



Future work should include model explainability to interpret if the learned embeddings accurately represent the ground truth dataset.

Acknowledgments

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

References

- Albright, Jack. 2019. "Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm." *Alzheimer's & Dementia: Translational Research & Clinical Interventions* (Elsevier) 5: 483–491.
- Basaia, Silvia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, Alzheimer's Disease Neuroimaging Initiative, and others. 2019. "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks." *NeuroImage: Clinical* (Elsevier) 21: 101645.
- Billones, Ciprian D., Olivia Jan Louville D. Demetria, David Earl D. Hostallero, and Prospero C. Naval. 2016. "DemNet: a convolutional neural network for the detection of Alzheimer's disease and mild cognitive impairment." *2016 IEEE Region 10 Conference (TENCON)*. 3724–3727.

Dashjamts, Tuvshinjargal, Takashi Yoshiura, Akio Hiwatashi, Osamu Togao, Koji Yamashita, Yasumasa Ohyagi, Akira Monji, et al. 2012. "Alzheimer's disease: diagnosis by different methods of voxel-based morphometry." *Fukuoka igaku zasshi= Hukuoka acta medica* 103: 59–69.

Grand, Jacob H. G., Sienna Caspar, and Stuart W. S. MacDonald. 2011. "Clinical features and multidisciplinary approaches to dementia care." *Journal of Multidisciplinary Healthcare* (Dove Press) 4: 125.

Hagos, Misgina Tsighe, Ronan P. Killeen, Kathleen M. Curran, Brian Mac Namee, Alzheimer's Disease Neuroimaging Initiative, and others. 2022. "Interpretable Identification of Mild Cognitive Impairment Progression Using Stereotactic Surface Projections." In *PAIS 2022*, 153–156. IOS Press.

Li, Matthew D., Ken Chang, Ben Bearce, Connie Y. Chang, Ambrose J. Huang, J. Peter Campbell, James M. Brown, et al. 2020. "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging." *NPJ Digital Medicine* (Nature Publishing Group UK London) 3: 48.

Moradi, Elaheh, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, and others. 2015. "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects." *Neuroimage* (Elsevier) 104: 398–412.

Roberts, Rosebud O., David S. Knopman, Michelle M. Mielke, Ruth H. Cha, V. Shane Pankratz, Teresa J. H. Christianson, Yonas E. Geda, et al. 2014. "Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal." *Neurology* (AAN Enterprises) 82: 317–325.

Xiao, Tingsong, Lu Zeng, Xiaoshuang Shi, Xiaofeng Zhu, and Guorong Wu. 2022. "Dual-Graph Learning Convolutional Networks for Interpretable Alzheimer's Disease Diagnosis." *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. 406–415.

Yang, Wenshuo, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2020. "HSCNN: a hybrid-siamese convolutional neural network for extremely imbalanced multi-label text classification." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6716–6722.

GAN-GA: A generative model based on genetic algorithm for medical image generation

Author

Mustafa AbdulRazek – Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt; TachyHealth, Lewes, US

Ghada Khoriba – Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt; Information Technology and Computer Science School, Nile University, Giza, Egypt

Mohammed Belal – Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

Citation

AbdulRazek, M., Khoriba, G., Belal, M. GAN-GA: A generative model based on genetic algorithm for medical image generation.

Abstract

Medical imaging is an essential tool for diagnosing and treating diseases. However, lacking medical images can lead to inaccurate diagnoses and ineffective treatments. Generative models offer a promising solution for addressing medical image shortage problems due to their ability to generate new data from existing datasets and detect anomalies in this data. This paper proposes the GAN-GA, a generative model optimised by embedding a genetic algorithm. The proposed model enhances image fidelity and diversity while preserving distinctive features, an essential aspect of image interpretation. To evaluate synthesised images, Frechet Inception Distance (FID) is used. The proposed GAN-GA model is tested by generating acute lymphoblastic leukaemia (ALL) medical images from an image dataset, which

is the first time it has been used in generative models. Our results were compared to those of InfoGAN as a baseline model. The experimental results show that the proposed optimised GAN-GA enhances FID scores by about 6.8%, especially in earlier training epochs.

Introduction

Several challenges arise when working with medical image datasets using deep learning approaches. The most known challenges are limited dataset size, imbalanced datasets, variability in image acquisition, annotation, and labelling, and patient privacy and security. Medical image datasets are often smaller than others, making it challenging to train deep-learning models effectively. Generating realistic medical images through generative adversarial networks (GANs) can be used to train deep-learning models for medical image analysis, ultimately leading to more accurate and effective diagnoses and treatment plans. Medical image synthesis is an active area of research, with applications in medical image analysis, simulation, and the training of machine learning models.

Generative models can create new data from scratch or be comparable to an already-existing data distribution. Generative models, such as variational autoencoders (VAEs) and GANs, have shown promising results in synthesising medical images. One of the main categories of techniques used to learn generative models from challenging real-world data is GANs [4]. However, these models have shown promising results, and careful evaluation of the synthetic data is necessary to ensure its clinical relevance and quality. To boost GAN's performance, several measures have been included, including Kullback-Leibler divergence [6], absolute deviation [9], and Wasserstein distance [2].

The main contributions of this paper are:

- Optimising GANs for medical image synthesis using a genetic algorithm: Integrating generative models with genetic algorithms produced a more exact and effective solution than baseline generative methods.

- Our proposed model was tested on acute lymphoblastic leukaemia, a small medical image dataset. The results show our model enhances generational performance as well as training stability. We create a generative adversarial network that addresses adversarial training as an evolutionary problem (GAEmbedded- with-InfoGAN). The embedding process of the genetic algorithm added another value by accelerating the learning and enhancement processes of model training.

Proposed InfoGAN-GA: Generative model embedded with genetic algorithm

Although deep learning-based methods have produced promising outcomes, they are occasionally limited by a lack of diversity in generated images, which can lead to biased or misleading diagnoses. We propose an InfoGAN-GA model that integrates a genetic algorithm within the Information Maximising GAN (Info-GAN) architecture originally introduced by Chen et al. [3]. This integration enables the InfoGAN-GA to leverage the strengths of the InfoGAN and the genetic algorithm, resulting in an advanced generative model that performs better on image generation tasks in terms of fidelity and converges in fewer epochs than the vanilla InfoGAN. In this section, we first review the InfoGAN and genetic algorithm formulations. Then, we introduce the proposed generative model embedded with the genetic algorithm, InfoGAN-GA.

Generative Adversarial Networks (GANs)

A two-player mini-max game between a discriminator network D and a generative network G is studied by GAN, which was first proposed in [4]. The generative network G outputs new data $G(z)$, whose distribution p_g is meant to be similar to that of the data distribution p_{data} given the noisy sample $z \sim p(z)$ (sampled from a uniform or normal distribution) as the input. In the meantime, the generated sample and the ground-truth data sample $p_{data}(x)$ are separated using the discriminator network D . In the original GAN, the adversarial training process was formulated as follows:

$$\min_G \max_D E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p(z)} [1 - \log D(G(z))] \quad (1)$$

Most existing GANs perform a similar adversarial procedure for different adversarial objective functions.

Information Maximizing Generative Adversarial Nets (InfoGAN)

InfoGAN, originally proposed in [3], can learn interpretable representations that are competitive with those acquired by existing supervised learning techniques. Instead of using a single unstructured noise vector, InfoGAN decomposes the input noise vector into two parts: (i) z , which is treated as a source of incompressible noise; and (ii) c , the latent code that targets the salient structured semantic features of the data distribution. InfoGAN can learn disentangled representations in an entirely unsupervised manner and maximise mutual information. A further term, $-\lambda I(c; g(z, c))$, separates InfoGAN's loss functions from GANs. λ is a tiny positive constant. Maximising the mutual information $I(c; g(z, c))$ and minimising the loss function of an InfoGAN (equations: 2 and 3) equate to minimising the loss of the original GAN.

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{data}} \log D(x) - \mathbb{E}_{z,c} \log [1 - D(g(z, c))] - \lambda I(c; g(z, c)) \quad (2)$$

$$\mathcal{L}_G = -\mathbb{E}_{z,c} \log D(g(z, c)) - \lambda I(c; g(z, c)) \quad (3)$$

Genetic Algorithm

In a genetic algorithm (GA) [7], successor states are generated by combining two parent states instead of modifying a single state. This is an example of stochastic beam search variation. Each state or individual is represented as a string across a finite alphabet and ranked by the objective or fitness function. The likelihood of being picked for reproduction is related to the fitness score, and the corresponding percentages are displayed alongside the raw numbers. GAs begin with a population, which consists of k randomly generated states. A random crossover point is chosen from the string positions to couple each crossover pair. At the crossover point, the parent strings are crossed over to generate the offspring, which can result in a state that is significantly distinct from either parent state. At the early stages of the search phase, when the population is often very diverse, crossover often occurs in large steps within the state space, followed by smaller steps when

the bulk of individuals are similar. The resulting children are then introduced to a new population, and the selection, crossover, and mutation processes are repeated until the new population is complete. Lastly, each position has a small possibility of being subject to random mutation independently. Genetic algorithms incorporate an incline, random exploration, and information sharing across parallel search threads. Genetic algorithms are primarily characterised by their crossover operations. Intuitively, the effect results from crossover’s ability to combine large chunks that have independently evolved to do meaningful tasks, enhancing the search’s granularity.

InfoGAN-GA: The Proposed Approach

An InfoGAN embedded with a genetic algorithm is proposed, as shown in figure 1. By applying the GA to each training phase, we aimed to improve the generator’s performance and speed up the generation of diverse images (as fake for the GAN) in earlier epochs. The initial population was first populated

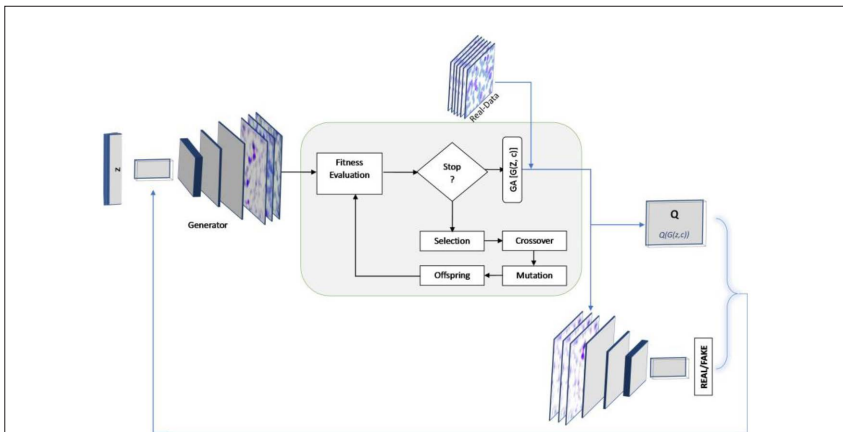


FIGURE 1 Generative Model embedded with Genetic Algorithm (InfoGAN-GA) Architecture. Latent vector z as input to Generator. $GA(G(z, c))$ is the output of genetic algorithm. Q is an auxiliary network attached to the second to last layer of the discriminator inspired from original InfoGAN.

with the fake individuals $G(z)$ (chromosome) generated by the G , and the fitness of each fake individual was determined as the D 's discrimination value $D(G(z))$. The selection type, a roulette wheel selection, and the arithmetic recombination crossover type were employed. The mutation was created by randomly adding a number to each gene with a 10% probability of doing so. The population was evolved using the aforementioned parameters, utilising the high-fitness samples (recognised as real by the discriminator). The least fit member of the conventional population was replaced by the child when the fitness of the offspring was higher than that of the parent.

Results and Discussions

Experimental Settings

Aria et al.[1] created the acute lymphoblastic leukaemia (ALL) [table 1] dataset and made it available to the public. The bone marrow laboratory at Taleqani Hospital created the images for this dataset (Tehran). This dataset comprised 3256 PBS pictures from 89 patients who were thought to have ALL and whose blood samples were skillfully processed and stained by laboratory personnel. The classes of benign and malignant data are separated in this data collection. The former includes hematogenous cells that are quite similar to ALL instances, but this hematopoietic precursor cell is benign, doesn't require chemotherapy, and typically goes away on its own. The latter category includes the malignant lymphoblasts known as ALL and the early pre-B, pre-B ALL, and pro-B ALL subtypes.

TABLE 1: Acute Lymphoblastic Leukemia (ALL) image dataset description

Type	Subtype	No. of samples	No. patients
Benign	Hematogones	504	25
Malignant	Early pre-B ALL	985	20
	Pre-B ALL	963	21
	Pro-B ALL	804	23
	Total	3256	89

Evaluation results

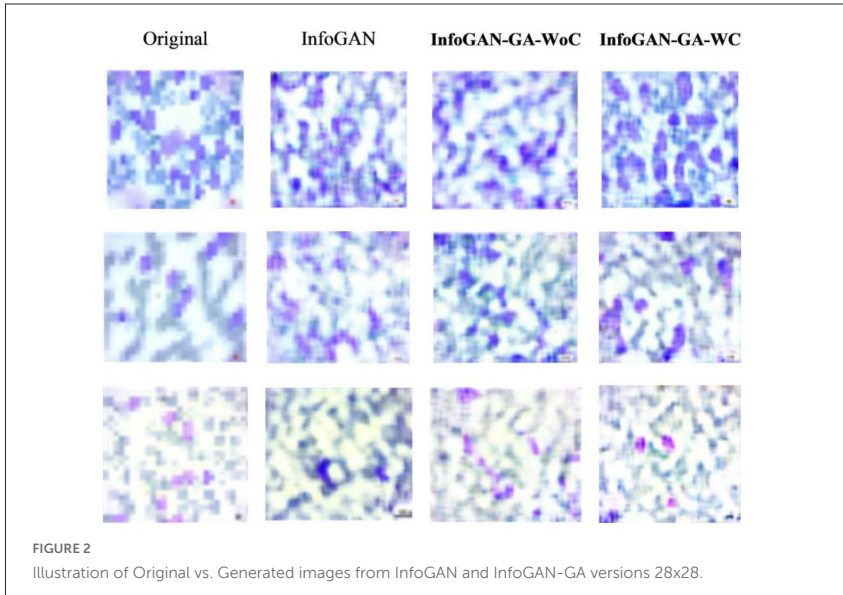
The generated images are evaluated based on their realism or fidelity. Discriminators in GANs are incapable of achieving perfection and frequently overfit to distinguish real from false images for their particular generator, making it difficult to compare and rank different models. Lately, the most popular assessment metrics for GANs have been the Inception Score (IS) and Fréchet Inception Distance (FID), both of which rely on an ImageNet-trained classifier (InceptionNet). FID (Heusel et al.[5]) calculates the Wasserstein-2 distance between multivariate Gaussians fitted to the embedding space of the Inception-v3 network of generated and real images as an improvement to the Inception Score (IS) [8]. We evaluated our model on the "ALL" dataset [1] at 28x28 resolutions. It's the first time to use a similar dataset with generative models, and the big challenge with the obtained dataset is the data size. The lack of sufficiently large datasets in the medical sector is one of the most common problems when dealing with medical AI solutions using ML and deep learning models. The samples generated from our model are presented in Figure 2, columns 3 and 4.

We show FID in Table 2. The network architectures are based on InfoGAN [3] and are briefly introduced here. We use the default hyper-parameter values listed in the default InfoGAN [3] for all experiments. Our experiments show that InfoGAN-GA already outperforms the InfoGAN model. Furthermore, all experiments were trained on the NVIDIA GeForce RTX 2070 with Max-Q Design GPU and input/output 28x28 images.

Table 2 lists trained models, generator losses, discriminator losses, FID, starting FID, and convergence epoch (CE) values, respectively. It's obvious that the proposed model outperforms the original InfoGAN in the final with a FID score of 123 compared with a FID score of 141, respectively.

TABLE 2: Final results and relative error reduction

Model	G Loss	D Loss	FID	Starting FID	CE
InfoGAN	0.195848	0.218547	141	328	2700
InfoGAN-GA-WoC	0.248143	0.221726	127	226	2500
InfoGAN-GA-WC	0.091876	0.224233	123	223	1750



The embedding process of the Genetic Algorithm accelerated the enhancement of generated image quality, synthesised images after the first epoch; it's evident that synthesised images in earlier epochs in the proposed model are better quality than the InfoGAN base model) faster than other previously mentioned base models and achieved diversity by feeding only the discriminator with real images but feeding the Genetic Algorithm with the population generated through Generator, removing the process of feeding the Genetic with target images to generate identical images. The embedding process of the genetic algorithm accelerated the enhancement of the generated image quality. Synthesised images after the first epoch prove that synthesised images in earlier epochs in the proposed model are of higher quality than the InfoGAN base model. The proposed InfoGAN-GA achieved diversity by feeding only the discriminator with real images and the gene pool of the genetic algorithm with synthetic images.

Conclusion

This paper proposes a novel approach for medical image synthesis using a generative model optimised by embedding a genetic algorithm, referred to as GANGA. The proposed model demonstrated enhanced generating performance and training stability compared to the baseline model, InfoGAN, on the acute lymphoblastic leukaemia image dataset. The GAN-GA model also improved image fidelity and diversity while preserving distinctive features. Integrating genetic algorithms with generative models proved to be an effective solution for medical image synthesis, producing a more exact outcome than traditional generative methods. Future research could explore applying the proposed GAN-GA model in other medical imaging domains and evaluating its performance against other state-of-the-art models.

References

- [1] Aria, M., Ghaderzadeh, M., Bashash, D., Abolghasemi, H., Asadi, F., Hosseini, A.: Acute lymphoblastic leukemia (all) image dataset (2021). <https://doi.org/10.34740/KAGGLE/DSV/2175623>, <https://www.kaggle.com/dsv/2175623>
- [2] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 214–223. PMLR (06–11 Aug 2017), <https://proceedings.mlr.press/v70/arjovsky17a.html>
- [3] Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain. pp. 2172–2180 (2016)

- [4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*. vol. 27. Curran Associates, Inc. (2014)
- [5] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017)
- [6] Nguyen, T., Le, T., Vu, H., Phung, D.: Dual discriminator generative adversarial nets (09 2017)
- [7] Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edn. (2010)
- [8] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 29. Curran Associates, Inc. (2016)
- [9] Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial networks. In: *International Conference on Learning Representations* (2017), <https://openreview.net/forum?id=ryh9pmcee>

Deformable image registration in the presence of simulated metal artefacts on head and neck CT scans

Author

Yanni Papastavrou – Department of Physics, Barking, Havering and Redbridge University Hospital NHS Trust

Tryphon Lambrou – School of Natural and Computing Sciences, University of Aberdeen, Scotland

Brian Hutton – Institute of Nuclear Medicine, University College London

Citation

Papastavrou, Y., Lambrou, T., Hutton, B. Deformable image registration in the presence of simulated metal artefacts on head and neck CT scans.

Abstract**Introduction**

Deformable image registration (DIR) is an important tool in radiotherapy. Dental metallic artefacts in head and neck CT images are a strong feature within the transaxial plane and can cause intensity-based DIR algorithms to register these artifacts preferentially over the patient's anatomy. This problem was investigated on artefact-free serial CT images by creating simulated dental metal artefacts and then comparing the DIR to the artefact-free results.

Methods

The baseline artefact-free DIR was performed using a rigid followed by a B-splines based DIR using control point spacings (CPS) of $40mm$, $20mm$ and $10mm$.

Dental artefacts were simulated in the fixed image. The floating image was rotated by small angles in the sagittal plane and artefacts were then simulated on this image in the same anatomical location. Registration was performed using the same parameters as in the artefact-free case. The transformation was compared to the artefact-free transformation by computing the Euclidean error at multiple points within two volumes of interest (VOIs) containing the artefact.

The errors were determined at points in these VOIs in regions containing the artefact and also within regions superior to, and inferior to, the artefact.

Results

The DIR errors only became significant in the vicinity of the artefact at the $10mm$ CPS level. The average errors were $3.0mm$ and $3.8mm$ in the simulated artefact regions and ranged from $0.8mm$ to $1.0mm$ elsewhere.

Discussion and Conclusion

As the registration becomes increasingly local, the ratio of artefact voxels to non-artefact voxels becomes increasingly large at the control points in the artefact's neighbourhood which then dominates the local registration results. This DIR has local support, therefore corruptions to one part of the image have minimal adverse effect on the registration performance on distant regions.

Introduction

Deformable image registration (DIR) is an important technique with many applications in radiotherapy [1]. When head and neck (HAN) patients lose weight requiring a rescan and replan, postural changes can occur despite employing immobilisation devices, resulting in set-up errors [2,3,4].

Such postural changes often result in small rotations about the sagittal plane, the dosimetric effect of these rotations have been studied [5]. Dental filling artefacts in HAN CT images [6] are a strong feature within the transaxial plane causing intensity-based DIR methods to register these metal artifacts preferentially over the patient's anatomy. This was investigated on artefact-free serial CT images by creating simulated artefacts and then registering using a B-splines based free-form deformation (FFD) image registration algorithm [7] and comparing the result to the artefact-free registration.

Simulating dental artefacts on CT images

Four slices were selected into which high intensity values were inserted into small regions of interest (ROI) to simulate the dental artefacts during reconstruction.

Forward projections from each slice at multiple angles $0^\circ \leq \theta \leq 179^\circ$ were constructed using radon transforms using MATLAB [8]. For each projection line, the shadow cast by the ROIs was determined. This raytracing was done by forward-projecting the points inside each region onto the projection line whilst noting the locus of points where high values appeared. Finally, for each sector of each projection line, a high numerical value was inserted. The image was then re-created from these modified projections using an inverse radon transform for back-projection, creating the simulated artefacts.

Assessing how the simulated artefacts affect registration

CT Image Data

Patients recruited into a phase 1 clinical trial to assess the response radiotherapy had PET-CT scans wearing their head-shell immobilisation device. Scans were acquired prior to treatment and following fractions 5, 20 and on treatment completion.

Simulating Artefacts on a Pair of Serial Images

CT scans were selected containing no metal artefacts. Metal artefacts were simulated in the pre-treatment image, $I(\vec{r}, t_1)$ - the fixed image - using the method described above. The floating image $I(\vec{r}, t_0)$ (following fraction 5)

was rotated by a small angle θ in the sagittal plane, known as the *artefact angle*. Simulated artefacts were then created on this image in the same location. These operations on the floating image may be represented by the following mapping:

$$I(\bar{r}, t_1) \Rightarrow I^*(\bar{r}, t_1) = A\{I(\bar{T}(\bar{r}), t_1)\} \quad (1)$$

Where $I(\bar{T}(\bar{r}), t_1)$ is the result of rotating the image $I(\bar{r}, t_1)$ through an angle θ in the sagittal plane (denoted by transformation T_θ) and $A\{I(\bar{r}, t_1)\}$ represents the process of creating the simulated artefact onto $I(\bar{r}, t)$. The operation applied to the fixed image was simply to create the artefact:

$$I(\bar{r}, t_0) \Rightarrow I^*(\bar{r}, t_0) = A\{I(\bar{r}, t_0)\} \quad (2)$$

Effect of the Simulated Artefacts on Registration

The histogram equalized [9] fixed and floating images onto which the artefacts had been simulated, $I^*(\bar{r}, t_0)$ and $I^*(\bar{r}, t_1)$, were registered using a B-splines FFD algorithm driven by normalised mutual information [8]. The registration consisted of a rigid, followed by a multiscale B-splines-based DIR using control point spacings of 40mm, 20mm and 10mm FFD control point spacing to give the following combined transformation:

$$\bar{T}(\bar{r}) = \bar{T}_{rigid,40,20,10}(\bar{r}) = \bar{T}_0(\bar{T}_0(\bar{T}_0(\bar{T}_{rigid}(\bar{r})))) \quad (3)$$

The results were visualised [10] to assess the registration performance in the vicinity of the artefact, with the deformation vectors were projected onto the orthogonal views [10,11].

A quantitative assessment was performed as follows. A registration was performed using the artefact-free images $I^*(\bar{r}, t_0)$ and $I^*(\bar{r}, t_1)$, to yield a baseline transformation. Then, the transformation captured from the registered simulated artefact images was compared to this baseline. The Euclidean distance between these two transformations was evaluated at points within the image. Suppose T is the artefact-free transformation and \bar{T}

is the (artefact) transformation, then the Euclidean distance between these two transformations at point \vec{r} is:

$$\Delta r = \|\mathcal{T}(\vec{r}) - T^*(\vec{r})\| \quad (4)$$

With $\|\vec{v}\|$ denoting vector magnitude.

A DIR, using a rigid transformation followed by a multi-level FFD transformation to give the transform $T_{rigid,40,20,10}(\vec{r})$ was performed for two cases: when the artefact angle was 5° and -3° .

For each case, a rectangular cuboid sub-volume (VOI) was defined where the registration performance was poor (the spine VOI and the jaw VOI) with the artefact approximately mid-way along the volume in the superior-inferior axis. Each VOI was further subdivided into three regions: a small range of transaxial slices containing the artefact; a region superior to the artefact and a region inferior to the artefact. Within these regions, points were randomly sampled and the Euclidean distance error Δr , was calculated using equation 4 and the mean, maximum and minimum Euclidean distance Δr was determined.

Results

For both cases, the rigid registration recovered the inverse applied rotation exactly. There was no detectable adverse effect caused by the artefact for the coarser-scale registrations, $T_{rigid,40}(\vec{r})$ and $T_{rigid,40,20}(\vec{r})$. The effects caused by the artefacts were observed at the fine-scale registration $T_{rigid,40,20,10}(\vec{r})$.

Figure 1 shows the transaxial and coronal CT slices using inverted colour scales, displaying the alignment following the fine-scale registration for the -3° case superimposed with the deformation vector field. Figure 2 shows the distance errors within each region.

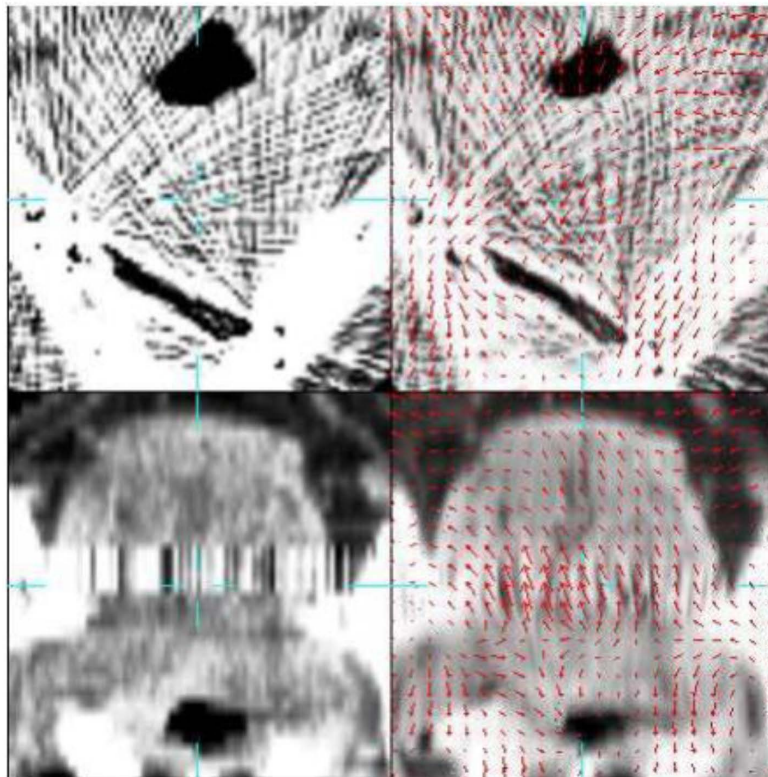
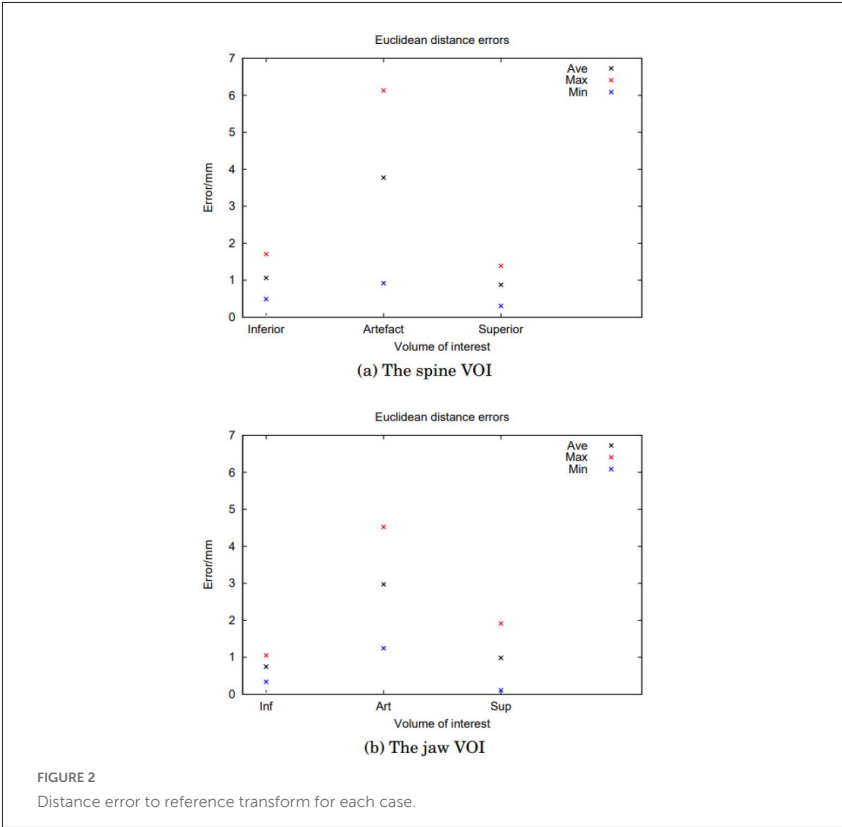


FIGURE 1

Vicinity of artefact following 10mm CPS DIR. Fixed image (left), floating image (right), the projected displacement vectors are displayed.

Discussion and Conclusion

The artefacts had no or negligible effect on the rigid and coarse-scale DIR due to the unaffected distal image regions driving the registration; since the B-splines FFD algorithm has local support. However, at the 10mm CPS



level, the registration becomes locally influenced by the artefact voxels preferentially registering compared to the local anatomy. The adverse effect on the registration caused by the simulated artefact can be seen in the deformations (Figure 1), with the deformation field partially aligning the two artefacts together resulting in the larger errors in the vicinity of the artefact (Figure 2). This preliminary study highlighted an issue for one particular DIR

algorithm. This effect should be more comprehensively studied for more cases and using a range of DIR algorithms, and expanded to deep-learning-based auto-contouring used in radio-therapy [12-14].

References

- [1] Rigaud, Bastien, et al. "Deformable image registration for radiation therapy: principle, methods, applications and evaluation." *Acta Oncologica* 58.9 (2019): 1225-1237.
- [2] Figen, Metin, et al. "Radiotherapy for head and neck cancer: evaluation of triggered adaptive replanning in routine practice." *Frontiers in Oncology* 10 (2020): 2427.
- [3] Sarwar, Asma, Papastavrou, Yanni et al. "Impact of brachial plexus movement during radical radiotherapy for head and neck cancers: the case for a larger planning organ at risk volume margin." *Journal of Radiotherapy in Practice* 19.2 (2020): 103-107.
- [4] Mencarelli, A., et al. "Automatic detection system for multiple region of interest registration to account for posture changes in head and neck radiotherapy." *Physics in Medicine & Biology* 59.8 (2014): 2005.

- [5] Stieb, Sonja, et al. "Dosimetric influence of pitch in patient positioning for radiotherapy of long treatment volumes; the usefulness of six degree of freedom couch." *The British Journal of Radiology* 91.1091 (2018): 20170704.
- [6] Boas, F. Edward, and Dominik Fleischmann. "CT artifacts: causes and reduction techniques." *Imaging Med* 4.2 (2012): 229-240.
- [7] Rueckert, Daniel, et al. "Nonrigid registration using free-form deformations: application to breast MR images." *IEEE transactions on medical imaging* 18.8 (1999): 712-721.
- [8] Mathworks website: <https://www.mathworks.com/products/matlab.html> last accessed 10/6/2023
- [9] Zhang, Wannan, and Yuqian Zhao. "Hierarchical registration of brain images based on B-splines and Laplacian commutators." *Optik* 241 (2021): 167022.
- [10] Hartkens, Thomas, et al. "VTK CISC registration toolkit an open source software package for affine and non-rigid registration of single-and multimodal 3D images." *Bildverarbeitung für die Medizin 2002*. Springer, Berlin, Heidelberg, 2002. 409-412.
- [11] Hartkens, Thomas. *Measuring, analysing and visualising brain deformation using non-rigid registration*. PhD Dissertation, King's College London (University of London), 2003.
- [12] Nikolov, Stanislav, et al. "Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study." *Journal of Medical Internet Research* 23.7 (2021): e26151.
- [13] Vrtovec, Tomaž, et al. "Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods." *Medical physics* 47.9 (2020): e929-e950.
- [14] Harrison, K., et al. "Machine learning for auto-segmentation in radiotherapy planning." *Clinical Oncology* 34.2 (2022): 74-88.

Masked autofocusing: CNN enabled targeted motion estimation and correction in MRI scans

Author

Ziad Al-Haj Hemidi, Christian Wehsbach, Mattias P. Heinrich – Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany

Citation

Hemidi, Z.A., Wehsbach, C., Heinrich, M.P. Masked autofocusing: CNN enabled targeted motion estimation and correction in MRI scans.

Introduction

Motivation Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that produces highly detailed images and is widely used for diagnosing various diseases located in soft tissues [5]. However, due to long scan times, MRI suffers from motion artifacts that degrade image quality [9]. Despite major advances in MR pulse sequence design and image reconstruction, motion remains a significant problem, with high clinical and public health costs.

Related work Motion avoidance and MRI acceleration techniques are used to reduce motion artifacts in MRI scans [4, 8], but still allow minor movements to occur which can degrade the quality of the scans and motion correction should be applied: Prospective motion correction involves continuous or repeated position determination and device adjustments [10]. For retrospective motion correction, e.g., autofocusing methods [2] are used to revert motion artifacts as well as CNN-based motion correction which

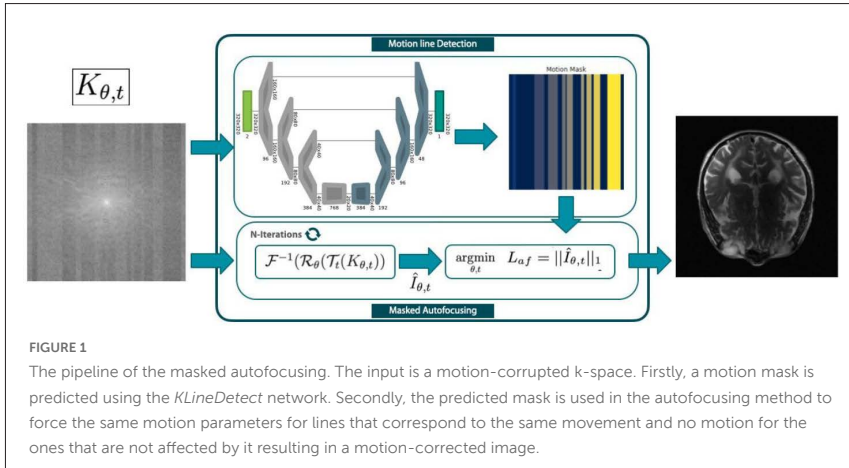
utilizes convolutional neural networks to remove motion artifacts directly from corrupted images [1]. In [3], i.e., CNNs are used to identify and mask motion corrupted k-space lines for a motion-corrected reconstruction.

Contributions In this work, we aim to correct motion retrospectively. We achieve this in two steps: First, we use a CNN-based motion detection network considering the acquisition physics. This way motion-corrupted k-space lines are detected. Secondly, we extend existing retrospective motion correction methods by combining effective CNN-based motion detection and autofocusing methods [2].

Methods

Firstly, motion is simulated using a technique based on [12]. The process involves generating a random movement model by sampling from normally distributed affine motion parameters. An artifact-free image is then resampled based on the movement model. A composite k-space is reconstructed from the k-spaces of multiple resampled images. Finally, the combined k-space is transformed back to the image domain, resulting in the final artifact sample. The application of this motion model yields a motion-corrupted image, the corresponding motion parameters, and a motion mask.

Secondly, we present a K-space Line Detection Network (*KLineDetect*) to identify motion-corrupted lines in raw k-space data. The network architecture is built on a U-Net [11] with ConvNextV2 [13] blocks. The network contains a downsampling path with depth 4 and a corresponding upsampling path. It takes complex-valued k-spaces and outputs the predicted motion masks. We use a binary cross-entropy loss and supervise the training with binary motion masks generated in the previous simulation step. Instance masks are created to differentiate between the different movements, the binary prediction is summed up over its rows and divided by the number of rows yielding a binary 1D vector. The line classified as motion corrupted is instanced by analyzing its predecessor when it's 0 the line will be categorized as a new instance and if 1 it will be categorized as its predecessor.



Lastly, for proof of concept, we deploy the predicted masks in an autofocusing method based on [2] where the 3 motion parameters, namely two translations and one rotation in the 2D case, are optimized for each motion group separately (see Fig. 1). The objective here is to minimize the absolute error in the motion-corrected image to enforce sharp edges and hence decrease blurring caused by motion artifacts.

Result

The FastMRI T_2 -weighted brain dataset [14] was employed for both training of the network and testing of our concept. We trained and evaluated the network on a split of 200/40 samples. We report a Dice overlap of 0.99 for the binary mask and an F1-score of 0.98 for the classification of lines on 20 test cases. For training the number of epochs was set to 4200. As the optimizer, we used AdamW [7] with a learning rate of $1e^{-4}$ and used cosine annealing with warm resets as a learning rate scheduler. In each epoch, we randomly selected a batch of 4 images and applied the motion simulation on the fly. The motion specifications included several movements between

TABLE 1: This table illustrates the motion correction performance of baseline auto-focusing, with ground truth motion mask, and with the predicted masks. The arrows indicate better scores

Method	SSIM \uparrow	PSNR (dB) \uparrow	NMSE \downarrow
Corrupted	0.67 \pm 0.08	24.73 \pm 1.84	0.13 \pm 0.05
autofocusing	0.66 \pm 0.06	24.89 \pm 1.70	0.13 \pm 0.13
Masked autofocusing GT	0.69 \pm 0.07	24.42 \pm 1.77	0.11 \pm 0.12
Masked autofocusing (ours)	0.69 \pm 0.07	24.40 \pm 1.81	0.11 \pm 0.11

5-20, rotations between -5° and 5° , and translations between $-5mm$ and $-5mm$ assuming head movements like nodding and small twists.

For proof of concept, we tested against the baseline autofocusing and the masked autofocusing with a ground truth motion mask while keeping the same test split. The number of iterations for the autofocusing was set empirically to 100 and Adam [6] was used for the gradient descent with a learning rate of $3e^{-4}$,

$\beta_1 = 0.899$, and $\beta_2 = 0.89$. In Table 1 we report the structural similarity index measure (SSIM), the peak signal-to-noise ratio (PSNR), and the normalized mean squared errors (NMSE).

Conclusion

We showed that our pipeline can detect and reduce motion artifacts and even outperforms the autofocusing baseline. Future work will employ different correction and reconstruction methods with the predicted masks. Tests were performed on generated motion data from T_2 -weighted Brain images, future work will include testing on other sequences and anatomies as well as real motion images.

The methods were implemented for 2D rigid motion only, future work will investigate 3D and more complex types of motion.

References

- [1] Al-Masni, M.A., Lee, S., Yi, J., Kim, S., Gho, S.M., Choi, Y.H., Kim, D.H.: Stacked u-nets with self-assisted priors towards robust correction of rigid motion artifact in brain mri. *NeuroImage* **259**, 119411 (2022)
- [2] Atkinson, D., Hill, D.L., Stoye, P.N., Summers, P.E., Clare, S., Bowtell, R., Keevil, S.F.: Automatic compensation of motion artifacts in mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **41**(1), 163–170 (1999)
- [3] Eichhorn, H., Hammernik, K., Spieker, V., Epp, S.M., Rueckert, D., Preibisch, C., Schnabel, J.A.: Deep learning-based detection of motion-affected k-space lines for t2*-weighted mri. *arXiv* (2023). <https://doi.org/10.48550/arxiv.2303.10987>
- [4] Huang, S.Y., Seethamraju, R.T., Patel, P., Hahn, P.F., Kirsch, J.E., Guimaraes, A.R.: Body mr imaging: artifacts, k-space, and solutions. *Radiographics* **35**(5), 1439 (2015)
- [5] Katti, G., Ara, S., Shireen, D.: Magnetic resonance imaging (mri) - a review. *Intl J Dental Clin* **3** (03 2011)
- [6] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
- [7] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)

[8] Lustig, M., Donoho, D., Pauly, J.M.: Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **58**(6), 1182–1195 (2007)

[9] Patel, P., Seethamraju, R., Kirsch, J., Hahn, P., Guimaraes, A.: Body mr imaging: Artifacts, k-space, and solutions. vol. 35 (12 2013). <https://doi.org/10.1148/rg.2015140289>

[10] Pipe, J.G.: Motion correction with propeller mri: application to head motion and free-breathing cardiac imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **42**(5), 963–969 (1999)

[11] Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)

[12] Shaw, R., Sudre, C., Varsavsky, T., Ourselin, S., Cardoso, M.J.: A k-space model of movement artifacts: Application to segmentation augmentation and artifact removal. *IEEE Transactions on Medical Imaging* **PP**, 1–1 (03 2020). <https://doi.org/10.1109/TMI.2020.2972547>

[13] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. *arXiv* (2023)

[14] Zbontar, J., Knoll, F., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdal, M., Romero, A., Rabbat, M.G., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C.L., Recht, M.P., Sodickson, D.K., Lui, Y.W.: fastmri: An open dataset and benchmarks for accelerated MRI. *CoRR* **abs/1811.08839** (2018), <http://arxiv.org/abs/1811.08839>

NSCLC radiogenomics, lung nodules segmentation and prediction of EGFR mutation status from CT scans

Author

Ivo Gollini Navarrete, Mohammad Yaqub – Mohamed Bin Zayed University of Artificial Intelligence. Abu Dhabi, UAE

Citation

Navarrete, I.G., Yaqub, M. NSCLC radiogenomics, lung nodules segmentation and prediction of EGFR mutation status from CT scans.

Introduction

Lung cancer, primarily non-small cell lung carcinoma (NSCLC), continues to be a leading cause of cancer-related deaths worldwide (1.8 million deaths in 2020) [1]. The therapeutic approach for NSCLC has shifted to biomarker-driven treatment, involving the use of small-molecule inhibitors, such as epidermal growth factor receptor (EGFR) or anaplastic lymphoma kinase (ALK), that have shown more promise than traditional chemotherapy. However, these require tissue samples for analysis, which are still expensive and carry risks [2, 3].

Radiogenomics, an interdisciplinary field that correlates radiomic features with genomic profiles, utilizes advanced computational approaches to overcome the shortcomings of molecular techniques [4]. Recent studies have investigated the potential of CT imaging and Machine Learning (ML) to determine the EGFR mutation status of lung cancer, thereby aiding in

diagnosis, treatment, and prognosis [5, 6]. In particular, Deep Learning (DL) has outperformed traditional ML approaches in predicting mutation status, offering improved patient data analysis and tumor segmentation [7, 8].

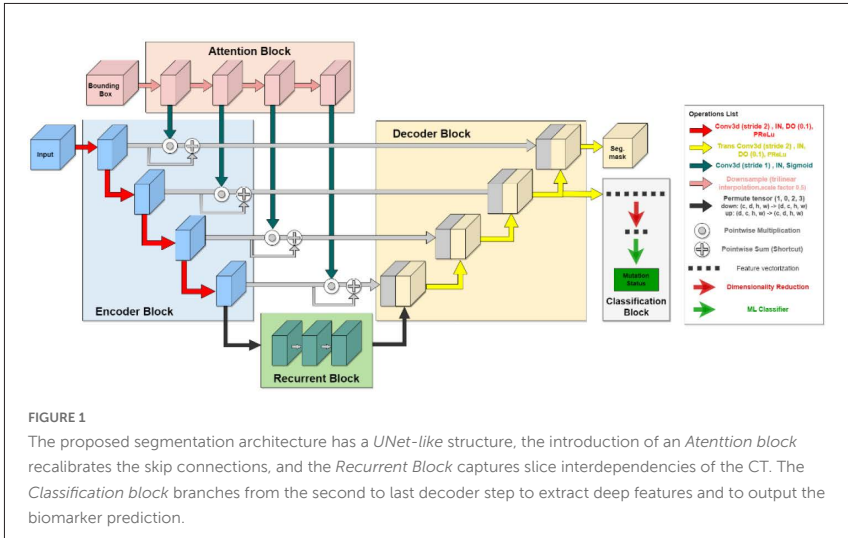
This work proposes an end-to-end radiogenomic pipeline with a DL mixed-supervised architecture for lung nodule segmentation and EGFR mutation prediction. The novel method addresses the data scarcity issue by introducing a 3D attention module to improve the localization of the nodule and a recurrent spatio-temporal block to capture slice interdependencies. The method is on par with SOTA methodologies. A hybrid system extracts deep features using segmentation as an auxiliary task to predict the EGFR mutation status using ML algorithms. The best-performing classifier outperforms SOTA methodologies with an AUC of 0.93.

Materials and Methods

The segmentation framework was trained on two datasets: NSCLC-Radiomics (RAD) [9] with 422 CT images and corresponding nodule manual delineation, 42 of which were excluded due to quality issues, and the Medical Decathlon Lung (MSD) [10], including 63 patients' CT images and lung tumor segmentations. The framework was tested using the NSCLC-Radiogenomics (RADGEN) dataset [11] comprising 211 patients, with tumor segmentation available for 144. The EGFR mutation status was known for 117 patients, making it useful for classification tasks.

All datasets underwent voxel intensity clipping $[-200, 250]$ and normalization of image voxel spacing by resampling to an anisotropic resolution of $1 \times 1 \times 1.5 \text{ mm}^3$ for preprocessing. The lung region of interest was isolated using UNet(R231) [12] pretrained weights to remove background and irrelevant components, after which images were resized to $256 \times 256 \times 256$, and 3D bounding boxes (binarized volume) were generated around tumor segmentation masks.

The segmentation model (Fig. 1) utilized a volumetric UNet architecture [13], capturing slice interdependencies through 3D recurrent layers proposed



in [14] and extending to a 3D setting of the organ-to-lesion (O2L) module presented in [15].

Several augmentations were used to improve performance, evaluated by Dice Similarity Coefficient (DSC) in a 5-fold cross-validation scheme. The model was optimized using ADAM optimizer and MONAI's Dice-Cross-Entropy Loss (DCE).

Segmentation was used as an auxiliary task to extract features for binary classification of EGFR mutation status. The data was first subjected to principal component analysis (PCA) and linear discriminant analysis (LDA) dimensionality reduction to maintain only meaningful features. Four machine learning methods were used: Quadratic Discriminant Analysis (QDA), Decision Tree (DT), Random Forest (RF), and C-Support Vector Classification (SVC). Parameter optimization was performed by maximizing the F1 score.

Results and Discussion

The presented lung nodule segmentation method achieves a DSC validation performance of 67.25 ± 3.12 on the MSD dataset and 72.36 ± 1.69 on RAD. The model outperforms the 3D nnUNET baseline by 11.41 and 9.68, respectively. It also surpasses the performance of existing models like Fredriksen et al. [16] by 2.98 and 19.44, respectively, and Kamal et al. [14] by a negligible difference of 0.08. Noticeably, our model capitalizes on exploiting fine-grained details by learning slice interdependencies and localizing tumors of high-quality labels, suggesting its superiority over increasing the training dataset size with lower-quality pseudolabels.

The presented model also excels when performing inference on the RADGEN test set, achieving a mean DSC of 66.73 ± 2.43 for the MSD dataset and 66.11 ± 1.31 for the RAD. An explanation is that RADGEN tumors' mean volume (8.22 cm^3) is closer in size to MSD (21.92 cm^3) compared to RAD tumor size (75.37 cm^3). Combining the training datasets (RAD+MSD) yields a significant performance improvement (73.54 ± 1.63 DSC). Furthermore, it outperforms Nishio et al.'s [17] method by 0.53, indicating the potential of data diversity over dataset size.

Finally, we test the effect of transfer learning by finetuning the RADGEN dataset using pretrained weights from models trained on RAD, MSD, and combined datasets. The best performance was achieved when utilizing the pretrained weights of the combined dataset (75.26 ± 4.37 DSC), highlighting the extent of transfer learning and further solidifying the impact of data diversity over data size.

In the classification domain, LDA dimensionality reduction outperforms PCA across all calcification algorithms. For example, RF achieves an F1 score of 0.613 and 0.948 when using PCA and LDA, respectively. Contrary to Reddy et al. [18], where PCA outperforms LDA only when the number of samples increases. This suggests the effect of RADGEN data size on our results. Table 1 shows the metrics after LDA dimensionality reduction and hyperparameter search.

TABLE 1: Accuracy, F1-macro, and ROC-AUC scores across multiple classifiers utilizing LDA dimensionality reduction

Classifier	Accuracy	F1 score	Precision	ROC-AUC
QDA	0.96 ± 0.02	0.93 ± 0.03	0.97 ± 0.01	0.90 ± 0.05
DT	0.94 ± 0.03	0.91 ± 0.05	0.91 ± 0.06	0.91 ± 0.06
RF	0.97 ± 0.02	0.95 ± 0.04	0.98 ± 0.01	0.93 ± 0.06
SVC	0.90 ± 0.11	0.83 ± 0.21	0.84 ± 0.22	0.83 ± 0.21

The proposed hybrid method outperforms several ML approaches [7, 19, 20] and current SOTA approaches. Such as the 0.846 reported by Moreno utilizing CNNs [7] or the implementation by Gui [8], which achieves an AUC of 0.86 using image reconstruction as an auxiliary task.

Conclusion

This study advances lung cancer segmentation and radiogenomic characteristics classification. Thanks to the aggregated modules, the proposed segmentation architecture reaches SOTA results. The proposed classification methodology for EGFR mutation status outperforms existing approaches without needing hand-crafted features. Our findings can enhance lung cancer treatment precision, though standardized methods are needed for clinical significance.

References

- [1] Chhikara, B.S., Parang, K.: Global cancer statistics 2022: the trends projection analysis. *Chemical Biology Letters* 10(1), 451–451 (2023).
- [2] Li, T., Kung, H.J., Mack, P.C., Gandara, D.R.: Genotyping and genomic profiling of non–small-cell lung cancer: implications for current and future therapies. *Journal of Clinical Oncology* 31(8), 1039 (2013).

[3] Chen, Z., Fillmore, C.M., Hammerman, P.S., Kim, C.F., Wong, K.K.: Non-small-cell lung cancers: a heterogeneous set of diseases. *Nature Reviews Cancer* 14(8), 535–546 (2014).

[4] Gevaert, O., Echegaray, S., Khuong, A., Hoang, C.D., Shrager, J.B., Jensen, K.C., Berry, G.J., Guo, H.H., Lau, C., Plevritis, S.K., et al.: Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific reports* 7(1), 1–8 (2017).

[5] Rios Velazquez, et al.: Somatic mutations drive distinct imaging phenotypes in lung cancersomatic mutations and radiomic phenotypes. *Cancer research* 77(14), 3922–3930 (2017).

[6] Nair, J.K.R., et al.: Radiogenomic models using machine learning techniques to predict egfr mutations in non-small cell lung cancer. *Canadian Association of Radiologists Journal* 72(1), 109–119 (2021).

[7] Moreno, S., Bonfante, M., Zurek, E., Cherezov, D., Goldgof, D., Hall, L., Schabath, M.: A radiogenomics ensemble to predict egfr and kras mutations in nsclc. *Tomography* 7(2), 154–168 (2021).

[8] Gui, D., Song, Q., Song, B., Li, H., Wang, M., Min, X., Li, A.: Air-net: A novel multi-task learning method with auxiliary image reconstruction for predicting egfr mutation status on ct images of nsclc patients. *Computers in Biology and Medicine* 141, 105157 (2022).

[9] Aerts, H., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature* 5(1), 1–9 (2014).

[10] Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms (2019).

[11] Bakr, S., et al.: A radiogenomic dataset of non-small cell lung cancer. *Scientific data* 5(1), 1–9 (2018).

[12] Hofmanninger, J., Prayer, F., Pan, J., Röhricht, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental* 4(1), 1–13 (2020).

[13] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016).

[14] Kamal, U., Rafi, A.M., Hoque, R., Wu, J., Hasan, M.K.: Lung cancer tumor region segmentation using recurrent 3d-denseunet. In: Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2. pp. 36–47. Springer (2020).

[15] Sun, L., Wu, J., Ding, X., Huang, Y., Wang, G., Yu, Y.: A teacher-student framework for semi-supervised medical image segmentation from mixed supervision. arXiv preprint arXiv:2010.12219 (2020).

[16] Fredriksen, V., Sevle, S.O.M., Pedersen, A., Langø, T., Kiss, G., Lindseth, F.: Teacher-student approach for lung tumor segmentation from mixed-supervised datasets. Plos one 17(4), e0266147 (2022).

[17] Nishio, M., Fujimoto, K., Matsuo, H., Muramatsu, C., Sakamoto, R., Fujita, H.: Lung cancer segmentation with transfer learning: usefulness of a pretrained model constructed from an artificial dataset generated using a generative adversarial network. Frontiers in Artificial Intelligence 4, 694815 (2021).

[18] Reddy, G.T., Reddy, M.P.K., Lakshmana, K., Kaluri, R., Rajput, D.S., Srivastava, G., Baker, T.: Analysis of dimensionality reduction techniques on big data. Ieee Access 8, 54776–54788 (2020).

[19] Koyasu, S., et al.: Usefulness of gradient tree boosting for predicting histological subtype and egfr mutation status of non-small cell lung cancer on 18f fdg-pet/ct. Annals of Nuclear Medicine 34 (2020).

[20] Morgado, J., et al.: Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. Applied Sciences 11(7), 3273 (2021).

Domain-specific interpretable AI for burn wound depth prediction using GPT-4

Author

Xinwei Zhang – Purdue University, School of Industrial Engineering, West Lafayette, IN 47907

Maxwell Jacobson – Purdue University, Department of Computer Science, West Lafayette, IN 47907

Mohamed El Masry – Indiana University, School of Medicine, Indianapolis, IN 46202

Surya Gnyawali – Indiana University, School of Medicine, Indianapolis, IN 46202

Yexiang Xue – Purdue University, Department of Computer Science, West Lafayette, IN 47907

Gayle Gordillo – Indiana University, School of Medicine, Indianapolis, IN 46202

Juan Wachs – Purdue University, School of Industrial Engineering, West Lafayette, IN 47907

Citation

Zhang, X., Jacobson, M., Masry, M.E., Gnyawali, S., Xue, Y., Gordillo, G., Wachs, J. Domain-specific interpretable AI for burn wound depth prediction using GPT-4.

Abstract

Our study introduces an interpretable ultrasound burn diagnosis system that balance accuracy and interpretability through domain-specific interpretability. We used an in-vivo porcine ultrasound dataset of burn, extracted key medical features of burn depth via deep learning, and predicted depth with a decision tree. The decision path was textualized using GPT-4 to generate comprehensible diagnoses statement. This reliable and effective

solution for automated burn wound assessment can improve patient outcomes, particularly in settings with limited access to clinical experts.

Introduction

The identification of burn depth remains a challenging aspect in clinical evaluation [1]. Despite innovative research using various technologies [2], automatic burn wound assessment with domain-specific interpretability [3] (DSI) is scarce, leading hurdles in understanding burn pathology and diagnostics when expertise is unavailable. We propose an interpretable ultrasound-based burn diagnosis system with a GPT4 [4] pipeline for explanatory statements, enhancing understanding and diagnostic accuracy.

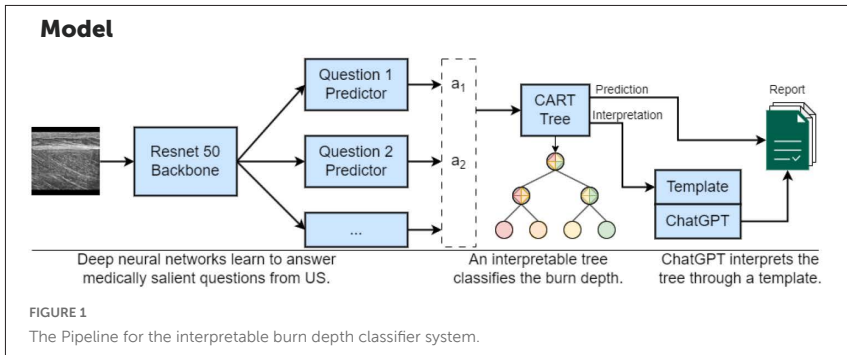
Background

Burns can be categorized into three types: superficial, partial-thickness, and full-thickness burns [5]. The assessment of burns is important for treatment selection and patient outcomes [6], but conventional diagnostic methods for burns pose certain limitations [7]. Consequently, researchers have explored various devices and algorithms to enhance the precision of burn diagnosis [8]. In recent years, the potential of B-mode ultrasound to improve classification accuracy has been recognized and deep learning algorithms have shown promising results in research [9]. Nonetheless, the black-box nature of deep learning remains an obstacle to its complete understanding [10]. Explicit decision tree algorithms are better suited to provide transparent inference processes for human comprehension [11]. Constructing datasets using diagnosis reports [12] and other forms to train models, thus enabling non-expert professionals to interpret medical imaging knowledge, has become a popular research area. Recently, ChatGPT has led breakthrough explorations in improving the method of knowledge instruction [13].

Methodology

Data Collection and Annotation

A burn depth dataset of porcine models was scanned by a B-mode device, including healthy skin and varying burn depths, verified via biopsies [14, 15]. This dataset, consisting of 4461 images from two pigs, was utilized for

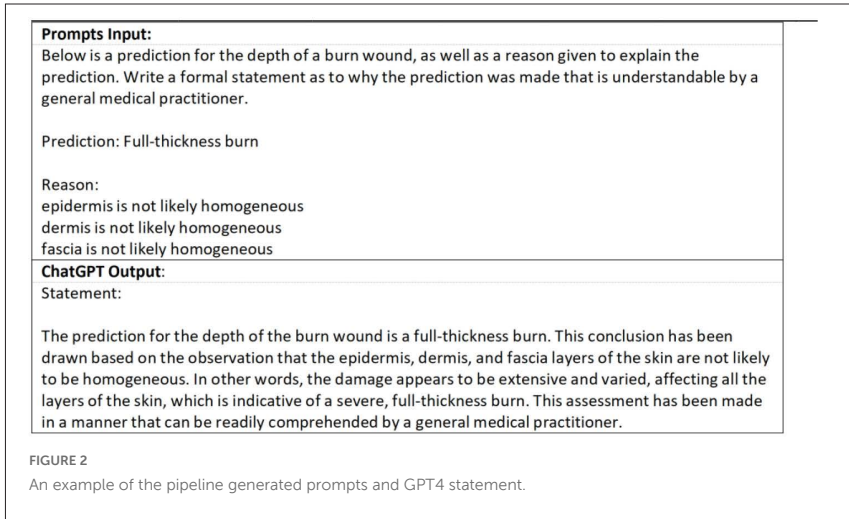


both training and testing. Cross-validation was applied to the dataset for all experiments, with the results averaged and reported subsequently.

We collaborated with burn experts to identify burn parameters through ultrasound, skin anatomy, and burn pathology. Epidermis, dermis, fascia and muscle were identified under ultrasound [16]. Next, each burn category was described based on disruptions to different layers. For instance, normal skin featured an intact epidermis without any damage to the underlying layers. Superficial burns were characterized by non-homogeneous epidermis without dermal damage, while partial thickness burns showed epidermal or dermal damage and spots of absent ultrasound waves. For full-thickness burns, extensive damage to all skin layers is observed, which is indicated by the absence of ultrasound waves in the deeper layers of the skin. Finally, three most representative question-answer pairs for fresh burns were obtained and annotated: "Is the epidermis homogenous?", "Is the dermis homogenous?" and "Is the fascia homogenous?"

Fig. 1. shows the structure of our system. We implemented a multi-target ResNet-50 [17] model for feature detection in ultrasound scans, each feature answering a specific question.

A CART decision tree [18], fine-tuned on our specific dataset, was utilized for burn depth prediction based on tissue prediction results. Baselines were



established using ResNet, SVM [19], and SVM with pretrained ResNet features. To ensure accurate baseline results, the hyperparameter grid search in Sklearn [20] was used for the SVM-based methods.

In our method, the decision path of the CART Tree was textualized and used to build a pipeline for automatic GPT prompt generation. This passed tissue prediction results for burn data into ChatGPT, leading to a comprehensible diagnostic statement (see Fig. 2).

Results

Table 1 shows the metrics of our ResNet-based algorithm for detecting the heterogeneity of different skin layers. Table 2 presents the results of burn depth prediction. It can be observed that the performance of our algorithm is comparable to that of the standard ResNet, with accuracy, precision, recall, and average F1 scores around 84%. The under-estimated rate and over-estimated rate are also provided. The “under-estimated rate” and “over-estimated rate” refer to how often burns are predicted to be less severe or more severe than they actually are, respectively.

TABLE 1: Performance metrics for each tissue feature

Tissue feature	Accuracy%	Precision%	Recall%	F1 Avg%
Heterogeneous epidermis	92.8	93.3	92.8	92.5
Heterogeneous dermis	95.6	95.6	96.5	96.5
Heterogeneous fascia	91.2	91.2	91.2	91.1

TABLE 2: Quantitative results in burn depth prediction

Method	Accuracy%	Precision%	Recall%	F1 Avg%	Under-est%	Over-est%
Ours	84.1	83.6	85.0	83.6	7.7	8.2
ResNet	84.8	87.9	84.8	85.3	7.0	8.3
SVM with ResNet features	26.9	32.5	26.9	25.0	38.6	34.6
SVM	67.9	71.9	70.6	66.5	20.5	11.7

To evaluate the explainability of the system, a survey was designed for three expert burn surgeons. These specialists were presented with a random selection of ten burn images from the testing dataset, accompanied by the corresponding model predictions and explanations. They rated the reasonableness of the explanations and their confidence in the predictions using a Likert scale [21]. The results indicated that the experts found the explanations generally reasonable and expressed moderate confidence in the predictions.

Conclusion

Our method ensures high prediction accuracy with interpretability, enabling medical professionals without extensive burns or machine learning knowledge to understand predictions. We achieved this by extracting key tissue features of burn depth using neural networks, predicting depth with a decision tree, and employing ChatGPT to generate comprehensible statements.

References

- [1] Atiyeh, B.S., Gunn, S.W., Hayek, S.N.: State of the art in burn treatment. *World Journal of Surgery* 29, 131–148 (2005).
- [2] Claes, K.E., Hoeksema, H., Vyncke, T., Verbelen, J., De Coninck, P., De Decker, I., Monstrey, S.: Evidence Based Burn Depth Assessment Using Laser-Based Technologies: Where Do We Stand?. *Journal of Burn Care & Research*, vol. 42, no. 3, pp. 513-525 (2021).
- [3] Habib, A., Karmakar, C., Yearwood, J.: Interpretability and optimisation of convolutional neural networks based on sinc-convolution. *IEEE Journal of Biomedical and Health Informatics* (2022).
- [4] OpenAI: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023).
- [5] Warby, R., Maani, C.V.: *Burn Classification*. (2019).
- [6] Lee, K.C., Joory, K., Moiemmen, N.S.: History of burns: the past, present and the future. *Burns & Trauma* 2(4), 2321-3868 (2014).

- [7] Monstrey, S., Hoeksema, H., Verbelen, J., Pirayesh, A., Blondeel, P.: Assessment of burn depth and burn wound healing potential. *Burns* 34, 761–769 (2008).
- [8] Haller, H.L., Giretzlehner, M., Dirnberger, J., Owen, R.: *Medical Documentation of Burn Injuries*. Springer, Vienna (2012).
- [9] Lee, S., Ye, H., Chittajallu, D., Kruger, U., Boyko, T., Lukan, J.K., Enquobahrie, A., Norfleet, J., De, S.: Real-Time Burn Classification Using Ultrasound Imaging. *Scientific Reports*, vol. 10, no. 1, pp. 1-13 (2020).
- [10] Tjoa, E., Guan, C.: A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *arXiv preprint arXiv:1907.07374* (2019).
- [11] Freitas, A.A.: Comprehensible Classification Models: A Position Paper. *ACM SIGKDD Explorations Newsletter*, 15(1), pp. 1-10 (2014).
- [12] Jing, B., Xie, P., Xing, E.: On the Automatic Generation of Medical Imaging Reports. *arXiv preprint arXiv:1711.08195* (2017).
- [13] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z.: Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *arXiv preprint arXiv:2304.01852* (2023).
- [14] Kim, J. Y., Dunham, D. M., Supp, D. M., Sen, C. K., Powell, H. M.: Novel Burn Device for Rapid, Reproducible Burn Wound Generation. *Burns*. 42(2), 384-391 (2016).
- [15] Sen, C. K., Ghatak, S., Gnyawali, S. C., Roy, S., Gordillo, G. M.: Cutaneous Imaging Technologies in Acute Burn and Chronic Wound Care. *Plastic and Reconstructive Surgery*. 138(3 Suppl), 119S (2016).
- [16] Mlosek, R. K., Malinowska, S.: Ultrasound Image of the Skin, Apparatus and Imaging Basics. *Journal of Ultrasonography*. 13(53), 212 (2013).
- [17] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778 (2016).

[18] Li, B., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees (CART). *Biometrics* 40(3), 358-361 (1984).

[19] Cortes, C., Vapnik, V.: Support-Vector Networks. *Machine Learning* 20, 273-297 (1995).

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825-2830 (2011).

[21] Likert, R.: A Technique for the Measurement of Attitudes. *Archives of Psychology* (1932).

AcquisitionFocus: Slicing optimization for fast cardiac MRI

Author

Christian Weihsbach – Institute of Medical Informatics, University of Luebeck, Luebeck, Germany

Nora Vogt – IADI U1254, Inserm, Université de Lorraine, Nancy

Ziad Al-Haj Hemidi – Institute of Medical Informatics, University of Luebeck, Luebeck, Germany

Alexander Bigalke – Institute of Medical Informatics, University of Luebeck, Luebeck, Germany

Lasse Hansen – EchoScout GmbH, Lübeck

Mattias Heinrich – Institute of Medical Informatics, University of Luebeck, Luebeck, Germany

Citation

Weihsbach, C., Vogt, N., Hemidi, Z.A., Bigalke, A., Hansen, L., Heinrich, M. AcquisitionFocus: Slicing optimization for fast cardiac MRI.

Introduction

Cardiac MRI imaging follows a specific acquisition routine: Firstly, a low resolution scout scan is taken to coarsely localize the heart (step 1). Secondly, the MRI technicians select appropriate imaging planes by examining the scout scan - of- ten this is done manually by referring to standardized protocols [1] (step 2). The scanner is then adjusted to capture the imaging planes of interest with higher resolution (spatially and/or temporally). Thirdly, the structures of interest are acquired (step 3). Lastly, the acquired images are examined by clinical experts or automated postprocessing procedures are applied to the acquired data (step 4).

Optimally, a high-contrast image with high spatial and temporal resolution could be acquired, but due to the physical MRI acquisition principal, contrast, field of view, spatial and temporal resolution are interlinked: Acquiring a larger field of view takes more time which increases the chance for motion artifacts during the scan which lowers the image contrast [1]. Overcoming the aforementioned constraints has been an active field of research: E.g. Parallel imaging shortened the acquisition times while keeping the image contrast [5]. Another way to achieve shortened acquisition times is to limit the field of view during acquisition - this requires the selection of descriptive imaging planes. These planes are mostly selected or defined manually in practice and previous studies [3, 4].

We hypothesize that the examination of medical images with computer-assisted techniques is improved further when the slice selection is also optimally tailored. To demonstrate this, we build upon recent work which explored the challenging task of reconstructing the full cardiac shape from US images [6] and translate this scenario: For MRI, we constrain the acquisition field of view to two sparse slices, which are fast to acquire physically and learn the optimal slice view orientation to achieve a better shape reconstruction. The definition and selection of optimal imaging planes [7, 1] for this task may be beyond human intuition, especially when deep learning methods are involved.

1. In this challenging target scenario, we aim to reconstruct the *full* cardiac shape of five structures replicating the MRI acquisition pipeline
2. We introduce the idea of joint-optimization of MRI slicing parameters of a 3D segmentation model

Methods

Our pipeline mimics the MRI acquisition process: From a low-resolution scout scan, a coarse anatomical shape is extracted by image segmentation. Descriptive shape slices are then selected by an acquisition model (see Fig. 1, left). The selected slices are then used to reconstruct the cardiac shape using a reconstruction model (see Fig. 1, right).

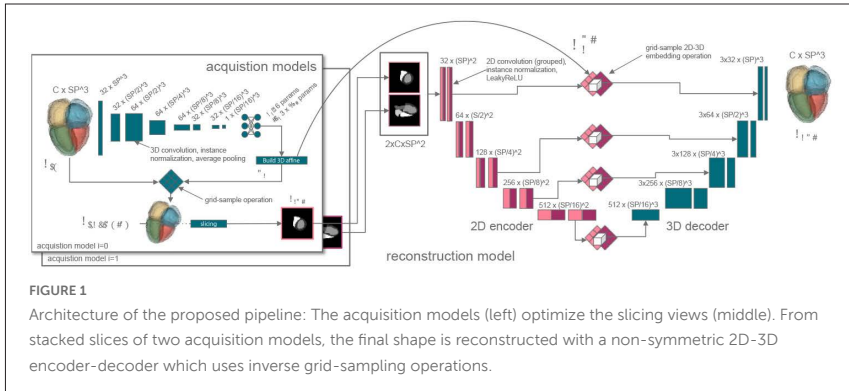


FIGURE 1
 Architecture of the proposed pipeline: The acquisition models (left) optimize the slicing views (middle). From stacked slices of two acquisition models, the final shape is reconstructed with a non-symmetric 2D-3D encoder-decoder which uses inverse grid-sampling operations.

Acquisition model: To obtain optimizable slice orientations, we feed the segmentation of a (low resolution) scout image scan V_{in} into an acquisition model (step 1). The model comprises a reorientation and slicing operators (see Fig. 1). For the acquisition model we take inspiration from Jaderberg et al. [2] and use a spatial transformer network (STN) including a 3D to 2D grid-sample operation. It consists of a CNN localization network with learnable parameters mapping the input volume V_{in} to six rotational and three translational parameters from which a 3D affine matrix A_1 is generated using the continual representation from [8]. The 3D affine matrix is then used to create a grid for the differentiable STN sampling layer.

Reconstruction model: For a given set of N optimized 2D image slices S from N acquisition models, we aim to reconstruct the full volumetric cardiac shape V_{re} . Our task maps data from 2D to 3D space - thus we configure the model to have a 2D encoder and a 3D branch, where the inverse of the affine matrices is used in the skip connections and bottleneck to embed back the 2D slices in 3D space again (see Fig 1).

Joint-optimization: Given the above models we obtain N optimized slices, by jointly training the parameters of N acquisition models and

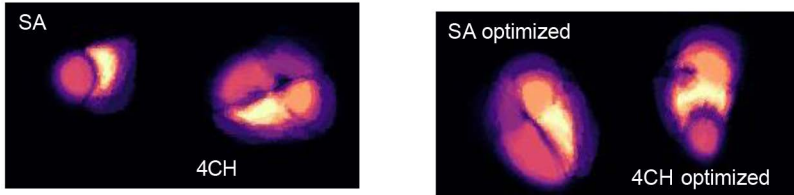


FIGURE 2

Slice optimization result of experiment I: Heatmap of oriented slice views. Slice orientations of standard SA and 4CH view (left). Optimized orientations (right).

one reconstruction model. For an optimal reconstruction, we require $V_{re} \equiv V_{in}$. This mapping could be fulfilled by learning an identity function, but it is restricted, since we feed the data through two bottlenecks: In the first bottleneck the information is reduced by slicing and the second bottleneck stores a compressed shape representation. In our pipeline the slice bottleneck is of special interest, as the reoriented slices $S_{1,\dots,N}$ reveal information which structures are most valuable to the re-construction task. Passing the predicted affine matrix A_i to the MRI control panel, the optimized views can be captured in a higher resolution.

Results and Conclusion

In experiments, we could show that the proposed pipeline can reorient standard clinical views during joint-optimization. Standard and optimized views are depicted in Fig. 2 as heatmap overlay of training samples. Notably, the less-informative SA slice was reoriented to a more informative view corresponding to the preselected 4CH view in the first optimization step. During the second optimization, the second view was then optimized to show a diagonal cut through several cardiac structures. Our presented method has some desirable advantages: The method considers the whole MRI acquisition pipeline and can reorient slice views for arbitrarily oriented low-resolution scout scans to a common space. This is currently still often done manually by experts. Automating this process may further prevent

mistakes of manual intervention. Moreover, optimization could be targeted to specific classes or even more complex downstream tasks such as disease classification.

References

- [1] Ismail, T.F., Strugnell, W., Coletti, C., Božić-Iven, M., Weingärtner, S., Ham-mernik, K., Correia, T., Küstner, T.: Cardiac mr: from theory to practice. *Frontiers in cardiovascular medicine* **9**, 137 (2022)
- [2] Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. *Advances in neural information processing systems* **28** (2015)
- [3] Natalia, F., Young, J.C., Afriliana, N., Meidia, H., Yunus, R.E., Sudirman, S.: Automated selection of mid-height intervertebral disc slice in traverse lumbar spine mri using a combination of deep learning feature and machine learning classifier. *Plos one* **17**(1), e0261659 (2022)
- [4] Nitta, S., Shiodera, T., Sakata, Y., Takeguchi, T., Kuhara, S., Yokoyama, K., Ishimura, R., Kariyasu, T., Imai, M., Nitatori, T.: Automatic 14-plane slice-alignment method for ventricular and valvular analysis in cardiac magnetic resonance imaging. *Journal of Cardiovascular Magnetic Resonance* **16**(Suppl 1), P1 (2014). <https://doi.org/10.1186/1532-429x-16-s1-p1>
- [5] Ridgway, J.P.: Cardiovascular magnetic resonance physics for clinicians: part i. *Journal of cardiovascular magnetic resonance* **12**(1), 1–28 (2010)

[6] Stojanovski, D., Hermida, U., Muffoletto, M., Lamata, P., Beqiri, A., Gomez, A.: Efficient pix2vox++ for 3d cardiac reconstruction from 2d echo views. In: Simplifying Medical Ultrasound: Third International Workshop, ASMUS 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. pp. 86–95. Springer (2022)

[7] Watkins, M.P., Williams, T.A., Caruthers, S.D., Wickline, S.A.: Cardiovascular mr function and coronaries: Cmr 15 minute express. *Journal of Cardiovascular Magnetic Resonance* **15**, 1–3 (2013)

[8] Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019)

Automated segmentation of rheumatoid arthritis immunohistochemistry stained synovial tissue

Author

Amaya Gallagher-Syed - Centre for Translational Bioinformatics, Queen Mary University of London; Digital Environment Research Institute, Queen Mary University of London

Abbas Khan - Digital Environment Research Institute, Queen Mary University of London

Felice Rivellese - Centre for Experimental Medicine and Rheumatology, Queen Mary University of London

Costantino Pitzalis - Centre for Experimental Medicine and Rheumatology, Queen Mary University of London

Myles J. Lewis - Centre for Experimental Medicine and Rheumatology, Queen Mary University of London

Gregory Slabaugh - Digital Environment Research Institute, Queen Mary University of London

Michael R. Barnes - Centre for Translational Bioinformatics, Queen Mary University of London; Digital Environment Research Institute, Queen Mary University of London

Citation

Gallagher-Syed, A., Khan, A., Rivellese, F., Pitzalis, C., Lewis, M.J., Slabaugh, G., Barnes, M.R. Automated segmentation of rheumatoid arthritis immunohistochemistry stained synovial tissue.

Introduction

Rheumatoid Arthritis Rheumatoid Arthritis (RA) is a chronic, autoimmune disease of unknown aetiology, affecting circa 1% of the total population. The joint synovial membrane (or synovium) is the major target of RA, which is characterised by synovial inflammation and hyperplasia [9]. The methods available for the study of synovial tissue have advanced considerably in recent years, from arthroplasty and blind needle biopsy to the use of arthroscopic and ultrasonographic technologies which improve the reliability and quality of synovial biopsies [12]. This has led to rapid progress in the study of disease pathogenesis and patient stratification, with increasingly complex analytical and technological approaches [12].

Digital Pathology One such approach is the histopathological assessment of joint synovium samples using digital Whole Slide Images (WSIs). The analysis of WSIs can lead to patient diagnosis and treatment by enabling the identification and quantification of spatial organisation and cellular features within the joint [3, 12, 14]. Complementary information can be gathered using several stain types, such as Haematoxylin & Eosin (H&E) and Immunohistochemistry (IHC). IHC in particular stains cellular proteins using specialised antibodies and is therefore well suited to highlighting functional organisation [13].

Tissue segmentation Yet much of the pre-processing and analysis of these samples is performed manually and semi-quantitatively by expert pathologists, a labour and knowledge-intensive task which precludes wider access, implementation in clinical practice and research reproducibility [14, 9, 8]. The effectiveness of other widely used medical image segmentation methods such as edge-based techniques [17, 11], active contours [6, 19] or watersheds [5, 4, 10] is limited by the great heterogeneity in stain intensity and colour, the fragmented nature of synovial tissue samples, as well as the presence of many undesirable artefacts present in the WSIs, such as water droplets, pen annotation, folded tissue, blurriness (see Figure 1C for reference) [5]. Furthermore, slides are typically stained with three or more different IHC stains and can originate from a variety of clinical centres, each with their own staining protocol, microscope and digital scanners [8]. There

is therefore a need for a robust, automated segmentation algorithm which can cope with this variability.

Contribution We provide a fully trained UNet segmentation tool for WSI IHC synovial tissue which can be used as the first step in an automated image analysis pipeline. It is robust to common WSIs artefacts, clinical centre/ scanner batch effect and can be used on different types of IHC stains. It can be used as is, or fine-tuned on any IHC musculoskeletal dataset, removing the need for manual segmentation by pathologists and offering a solution to the current image analysis bottleneck. The code is available: https://github.com/AmayaGS/IHC_Synovium_Segmentation

Methods

Data collection A total of 164 patients, fulfilling the 2010 American College of Rheumatology/European Alliance of Associations for Rheumatology (EULAR) classification criteria for RA were recruited to the R4RA clinical trial from 20 European centers [15] [7]. Patients underwent ultrasound-guided synovial biopsy of a clinically active joint.

Briefly, samples were then fixed in formalin, embedded in paraffin, cut with microtome and stained with the relevant IHC stains: IHC CD20 (B cells), IHC CD68 (macrophages) and IHC CD138 (plasma cells) [7]. Samples were then placed on glass slides and scanned into Whole Slide Image (.ndpi format) with digital scanners under 40x or 20x objectives.

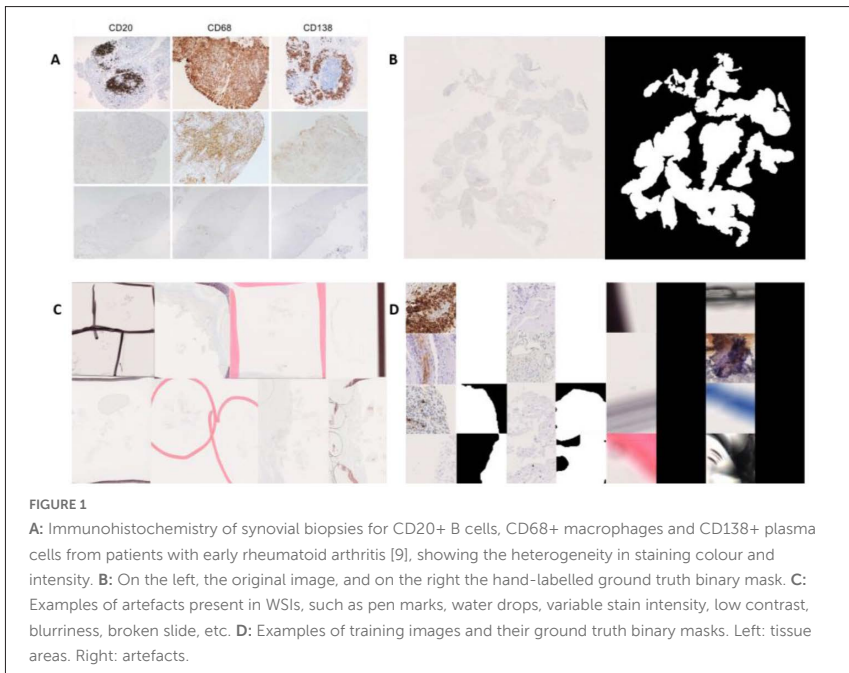
Hand labelling 465 IHC WSIs were manually labelled and full-scale binary Ground Truth (GT) masks were extracted using the QuPath software [2], as shown in Figure 1B.

Training set Patient IDs were used to randomly divide the dataset into Train/ Val/Test sets with 80/10/20 percent of the data in each. For the train and validation set, patches were generated as follows: within tissue areas, the high-resolution WSI was split into non-overlapping patches of 224x224 pixels at 10x magnification. Furthermore, to represent all the artefacts present in the dataset, patches were chosen at random in non-tissue areas totalling approximately

half the training dataset. In total 240,181 patches were created, each with a corresponding GT mask. The 10x magnification was chosen as a compromise showing both the macro/micro-architecture of the tissue, as well as reducing the number of patches for storage and computation purposes.

Testing set 107 Test set WSIs were extracted at magnification 10X. Contrary to the Train/Val set the whole image was patched and reconstructed within the testing pipeline. Segmentation masks were predicted for the whole image and coloured as follows: Yellow for True Positive (TP) segmented tissue, Red for False Negative (FN) and Green for False Positive (FP).

Training schedule Training was conducted using a UNet segmentation architecture as in [16]. All the models were trained using Adam optimizer with



their default beta values; the learning rate was set to 0.0001 with a batch size of 20. A custom early stopping mechanism was employed to terminate the training before the model overfits the data. The loss and Dice score were monitored, and training halted if neither improved for five consecutive epochs. Data augmentation was used to modify Saturation, Hue, Contrast and Brightness values randomly. All networks were trained with the Focal Tversky loss function [1]. The Dice score metric was used to evaluate all results.

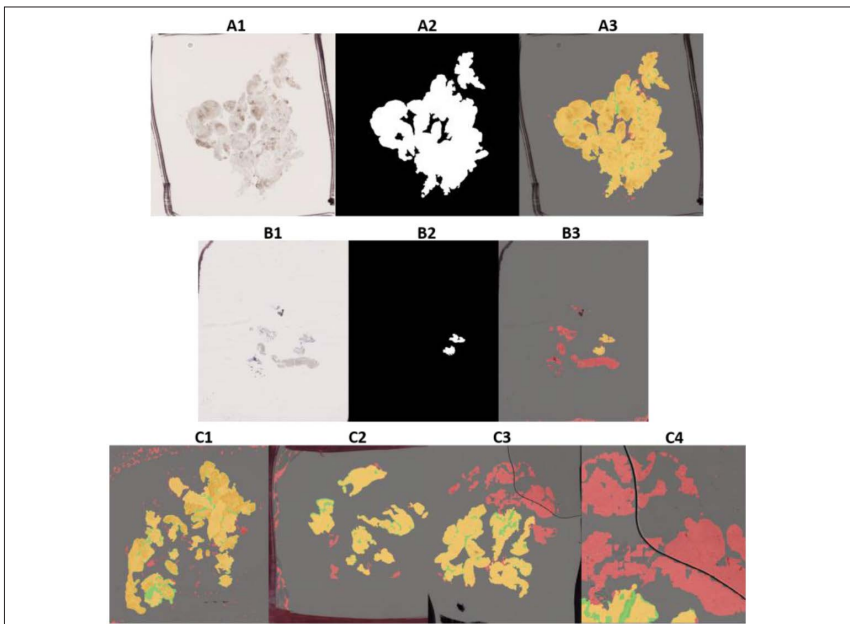


FIGURE 2

A: Example of a high Dice score results (0.97), with original image, hand-labelled segmentation mask and predicted segmentation mask. **B:** Example of a low Dice score result (0.20), with a poorly hand-labelled ground truth mask. **C:** in **C1** speckled dye spread on the slide is recognised as tissue. **C2** iridescent stain on the slide is recognised as tissue. **C3** a large area of tissue is classified in red as False Positive, however, this area corresponds to real tissue which was not annotated fully. In **C4** the algorithm correctly avoids segmenting the borders of a water drop.

Experimental results

The UNet algorithm successfully segments synovial tissue WSIs stained with three different IHC stains and scanned in 20 different clinical centres across Europe obtaining a Dice score of 0.863 ± 0.112 , indicating it is highly robust to different IHC stain types, clinical centre batch effect and data heterogeneity. In Figure 2, we show qualitative results to illustrate some of its strengths and limitations: in A2 we see the model is able to segment areas of low contrast, whilst ignoring strong signals such as pen marks. In B2, we see the model was actually more successful than the human labeller at correctly segmenting tissue areas, highlighting the need for an automated segmentation algorithm which does not suffer from fatigue or moments of inattention. Finally, in C2 we illustrate some of the limitations of the model, such as incorrect segmentation of speckled dye and iridescence stains. Further training of image artefacts could help improve robustness, yet overall the model is able to deal with common WSIs artefacts, such as variations in colour and intensity, blurriness, different colour pen marks, tissue folding, water drops, etc.

Conclusion

We present a fully automated deep learning segmentation algorithm, which is freely available and can be used as a first step in any rheumatoid arthritis or with further finetuning, any musculoskeletal IHC image analysis pipeline, avoiding lengthy manual annotation and helping to improve their speed, repeatability and robustness. This is a key step towards the acceleration of research into the mechanisms involved in rheumatoid arthritis and potentially other forms of inflammatory arthritis.

References

[1] Abraham, N., Khan, N.M.: A novel focal tversky loss function with improved attention u-net for lesion segmentation. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 683–687. IEEE (2019)

[2] Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., James, J.A., Salto-Tellez, M., Hamilton, P.W.: QuPath: Open source software for digital pathology image analysis. *Scientific Reports* 7(1), 16878 (Dec 2017).

<https://doi.org/10.1038/s41598-017-17204-5>

<https://www.nature.com/articles/s41598-017-17204-5>

[3] Dennis, G., Holweg, C.T., Kummerfeld, S.K., Choy, D.F., Setiadi, A.F., Hackney, J.A., Haverty, P.M., Gilbert, H., Lin, W.Y., Diehl, L., Fischer, S., Song, A., Musselman, D., Klearman, M., Gabay, C., Kavanaugh, A., Endres, J., Fox, D.A., Martin, F., Townsend, M.J.: Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. *Arthritis Research & Therapy* 16(2), R90 (2014)

<https://doi.org/10.1186/ar4555>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060385/>

[4] Dhage, P., Phegade, M.R., Shah, S.K.: Watershed segmentation brain tumor detection. In: 2015 International Conference on Pervasive Computing (ICPC). pp. 1–5 (Jan 2015). <https://doi.org/10.1109/PERVASIVE.2015.7086967>

[5] Di Cataldo, S., Ficarra, E., Acquaviva, A., Macii, E.: Automated segmentation of tissue images for computerized IHC analysis. *Computer Methods and Programs in Biomedicine* 100(1), 1–15 (Oct 2010).

<https://doi.org/10.1016/j.cmpb.2010.02.002>

<https://www.sciencedirect.com/science/article/pii/S0169260710000337>

[6] Hemalatha, R.J., Thamizhvani, T.R., Dhivya, A.J.A., Joseph, J.E., Babu, B., Chandrasekaran, R.: Active Contour Based Segmentation Techniques for Medical Image Analysis. *IntechOpen* (Jul 2018).

<https://doi.org/10.5772/intechopen.74576>

<https://www.intechopen.com/chapters/59741>

[7] Humby, F., Durez, P., Buch, M.H., Lewis, M.J., Rizvi, H., Rivellese, F., Nerviani, A., Giorli, G., Mahto, A., Montecucco, C., Lauwerys, B., Ng, N., Ho, P., Bombardieri, M., Romão, V.C., Verschueren, P., Kelly, S., Sainaghi, P.P., Gendi, N., Dasgupta, B., Cauli, A., Reynolds, P., Cañete, J.D., Moots, R., Taylor, P.C., Edwards, C.J., Isaacs, J., Sasieni, P., Choy, E., Pitzalis, C., Thompson, C., Bugatti, S., Bellan, M., Congia, M., Holroyd, C., Pratt,

A., Fonseca, J.E.C.d., White, L., Warren, L., Peel, J., Hands, R., Fossati Jimack, L., Hadfield, G., Thorborn, G., Ramirez, J., Celis, R.: Rituximab versus tocilizumab in anti-TNF inadequate responder patients with rheumatoid arthritis (R4RA): 16-week outcomes of a stratified, biopsy-driven, multicentre, openlabel, phase 4 randomised controlled trial. *The Lancet* 397(10271), 305–317 (Jan 2021)

[https://doi.org/10.1016/S0140-6736\(20\)32341-2](https://doi.org/10.1016/S0140-6736(20)32341-2)

[https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(20\)32341-2/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32341-2/fulltext)

[8] Lahiani, A., Gildenblat, J., Klaman, I., Navab, N., Klaiman, E.: Generalising multistain immunohistochemistry tissue segmentation using end-to-end colour deconvolution deep neural networks. *IET Image Processing* 13(7), 1066–1073 (2019). <https://doi.org/10.1049/ietipr.2018.6513>

<https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-ipr.2018.6513>

[9] Lewis, M.J., Barnes, M.R., Blighe, K., Goldmann, K., Rana, S., Hackney, J.A., Ramamoorthi, N., John, C.R., Watson, D.S., Kummerfeld, S.K., Hands, R., Riahi, S., Rocher-Ros, V., Rivellese, F., Humby, F., Kelly, S., Bombardieri, M., Ng, N., DiCicco, M., van der Heijde, D., Landewé, R., van der Helm-van Mil, A., Cauli, A., McInnes, I.B., Buckley, C.D., Choy, E., Taylor, P.C., Townsend, M.J., Pitzalis, C.: Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes. *Cell Reports* 28(9), 2455–2470.e5 (Aug 2019).

<https://doi.org/10.1016/j.celrep.2019.07.091>

<https://www.sciencedirect.com/science/article/pii/S2211124719310071>

[10] Lu, Y., Jiang, Z., Zhou, T., Fu, S.: An Improved Watershed Segmentation Algorithm of Medical Tumor Image. IOP Conference Series: Materials Science and Engineering 677(4), 042028 (Dec 2019).

<https://doi.org/10.1088/1757-899X/677/4/042028>,

[11] Luo, Y., Liu, L., Huang, Q., Li, X.: A Novel Segmentation Approach Combining Region- and Edge-Based Information for Ultrasound Images. BioMed Research International 2017, 9157341 (2017).

<https://doi.org/10.1155/2017/9157341>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426079/>

[12] Orr, C., Vieira-Sousa, E., Boyle, D.L., Buch, M.H., Buckley, C.D., Cañete, J.D., Catrina, A.I., Choy, E.H.S., Emery, P., Fearon, U., Filer, A., Gerlag, D., Humby, F., Isaacs, J.D., Just, S.A., Lauwerys, B.R., Le Goff, B., Manzo, A., McGarry, T., McInnes, I.B., Najm, A., Pitzalis, C., Pratt, A., Smith, M., Tak, P.P., Tas, S.W., Thurlings, R., Fonseca, J.E., Veale D.J.: Synovial tissue research: a state-of-the-art review. Nature Reviews Rheumatology 13(8), 463–475 (Aug 2017).

<https://doi.org/10.1038/nrrheum.2017.115>

<https://www.nature.com/articles/nrrheum.2017.115>

[13] Pham, B., Gaonkar, B., Whitehead, W., Moran, S., Dai, Q., Macyszyn, L., Edgerton, V.R.: Cell Counting and Segmentation of Immunohistochemical Images in the Spinal Cord: Comparing Deep Learning and Traditional Approaches. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2018, 842–845 (Jul 2018). <https://doi.org/10.1109/EMBC.2018.8512442>

[14] Rivellese, F., Mauro, D., Nerviani, A., Pagani, S., Fossati-Jimack, L., Messemaker, T., Kurreeman, F.A.S., Toes, R.E.M., Ramming, A., Rauber, S., Schett, G., Jones, G.W., Jones, S.A., Rossi, F.W., Paulis, A.d., Marone, G., Shikh, M.E.M.E., Humby, F., Pitzalis, C.: Mast cells in early rheumatoid arthritis associate with disease severity and support B cell autoantibody production. Annals of the Rheumatic Diseases 77(12), 1773–1781 (Dec 2018).

<https://doi.org/10.1136/annrheumdis-2018-213418>, <https://ard.bmj.com/content/77/12/1773>

[15] Rivellese, F., Surace, A.E.A., Goldmann, K., Sciacca, E., Çubuk, C., Giorli, G., John, C.R., Nerviani, A., Fossati-Jimack, L., Thorborn, G., Ahmed, M., Prediletto, E., Church, S.E., Hudson, B.M., Warren, S.E., McKeigue, P.M., Humby, F., Bombardieri, M., Barnes, M.R., Lewis, M.J., Pitzalis, C.: Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nature Medicine* 28(6), 1256–1268 (Jun 2022).

<https://doi.org/10.1038/s41591-022-01789-0>

<https://www.nature.com/articles/s41591-022-01789-0>

[16] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. pp. 234–241. *Lecture Notes in Computer Science*, Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[17] Salman, N.H., Ghafour, B.M., Hadi, G.M.: Medical Image Segmentation Based on Edge Detection Techniques. *European Journal of Applied Sciences* 3(2), 1–1 (May 2015). <https://doi.org/10.14738/aivp.32.1006>

<https://journals.scholarpublishing.org/index.php/AIVP/article/view/1006>

[18] Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access* 9, 82031–82057 (2021).

<https://doi.org/10.1109/ACCESS.2021.3086020>

[19] Yang, Y., Hou, X., Ren, H.: Efficient active contour model for medical image segmentation and correction based on edge and region information. *Expert Systems with Applications* 194, 116436 (May 2022).

<https://doi.org/10.1016/j.eswa.2021.116436>

<https://www.sciencedirect.com/science/article/pii/S0957417421017231>

Deep complex-valued edge attention network for artefact removal in cardiovascular magnetic resonance with undersampling spiral trajectories

Author

Yaqing Luo - National Heart and Lung Institute, Imperial College London, London, UK; CMR Unit, Royal Brompton and Harefield Hospitals, London, UK; EPSRC Centre for Doctoral Training in Smart Medical Imaging, King's College London and Imperial College London, London

Pedro F. Ferreira - National Heart and Lung Institute, Imperial College London, London, UK; CMR Unit, Royal Brompton and Harefield Hospitals, London, UK

Dudley J. Pennell - National Heart and Lung Institute, Imperial College London, London, UK; CMR Unit, Royal Brompton and Harefield Hospitals, London, UK

Guang Yang - National Heart and Lung Institute, Imperial College London, London, UK; CMR Unit, Royal Brompton and Harefield Hospitals, London, UK

Sonia NIELLES-VALLESPIN - National Heart and Lung Institute, Imperial College

London, London, UK; CMR Unit, Royal Brompton and Harefield Hospitals, London, UK

Andrew D. Scott - National Heart and Lung Institute, Imperial College London, London, UK; CMR Unit, Royal Brompton and Harefield Hospitals, London, UK

Citation

Luo, Y., Ferreira, P.F., Pennell, D.J., Yang, G., NIELLES-VALLESPIN, S., Scott, A.D. Deep complex-valued edge attention network for artefact removal in cardiovascular magnetic resonance with undersampling spiral trajectories.

Abstract

Diffusion Tensor Cardiovascular Magnetic Resonance (DT-CMR) reveals in vivo myocardial microstructure but can be time-consuming and prone to artefacts. We propose a complex-valued deep Edge Attention Network for artefact removal in highly undersampled DT-CMR images. Our proposed model outperformed other networks in reducing artefacts, recovering myocardial details, increasing perceptual quality and image sharpness, and it potentially benefits clinical applications and facilitates accelerated imaging.

Introduction

Diffusion Tensor Cardiovascular Magnetic Resonance (DT-CMR) is a powerful non-invasive technique that interrogates the in vivo myocardial microstructure [1], providing measures such as the orientation of cardiomyocytes [2] and sheetlets [3] in healthy subjects and cardiovascular diseases [4-7]. However, DT-CMR has a limited spatial resolution due to its inherent low Signal to Noise Ratio (SNR) [8-11].

Non-Cartesian sampling techniques, such as spiral sampling, can improve the efficiency of data acquisition [12-17] and are well-suited to sub-Nyquist sampling. A key challenge for undersampled spiral DT-CMR is removing the artefacts.

Using the simulated spiral data, we aim to demonstrate effective artefact removal from highly accelerated DT-CMR data acquired with undersampled spiral k-space trajectories, by developing a novel deep complex-valued [18-24] Edge Attention Residual U-Net [25].

Methods

Proposed Model

We introduced the complex-valued Spatial Attention Module (SAM) [26] which operates in each depth concatenation path and emphasises the salient image regions by assigning importance measures. This attention module strengthens the connection between the encoded feature maps

and decoded feature maps. This would allow the informative activations that represent the myocardium to be preserved through propagation, and the features that contain artefacts and background noise to be progressively suppressed during training. This should enhance the network's robustness against undersampling artefacts, and further improve the quality of the predicted images.

The loss of high spatial frequency information can be difficult to recover, particularly the areas corresponding to edge features. We proposed a complex-valued Edge Attention Module (EAM) within each encoding layer to enhance myocardial details and image sharpness. The EAM calculates the horizontal image gradient to highlight object edges by identifying the boundaries and myocardial details that would need further enhancement. Edge images are passed through complex convolution and batch normalisation layers to reduce unwanted noise before the complex sigmoid activation function is applied to create the edge attention maps. This module is implemented in the encoding stage of the network to reduce the loss of high frequency information through convolution and pooling.

We train networks to map a highly undersampled complex image to the corresponding unaliased complex ground truth, outputting complex predicted images. We define the loss function as a weighted sum of the \mathcal{L}^1 norm of the complex data in image space and k-space, with empirical weighting of 1 and 0.1 respectively. The optimiser was Stochastic Gradient Descent (SGD) with a learning rate of 0.0003, momentum of 0.6 and a weight decay of 0.0001. The number of training epochs is 1000 to give sufficient iterations for the network to converge. Here we compare four Residual U-Net based networks: magnitude, magnitude Edge Attention, complex and our proposed method, complex Edge Attention.

Data and Training

Our dataset contains DT-CMR images from 20 controls, scanned with STEAM-EPI [27] in both diastole and systole, at 1.5T and 3T (4 datasets per subject). The in-plane acquisition resolution is $2.8 \times 2.8 \times 8\text{mm}^3$ and field

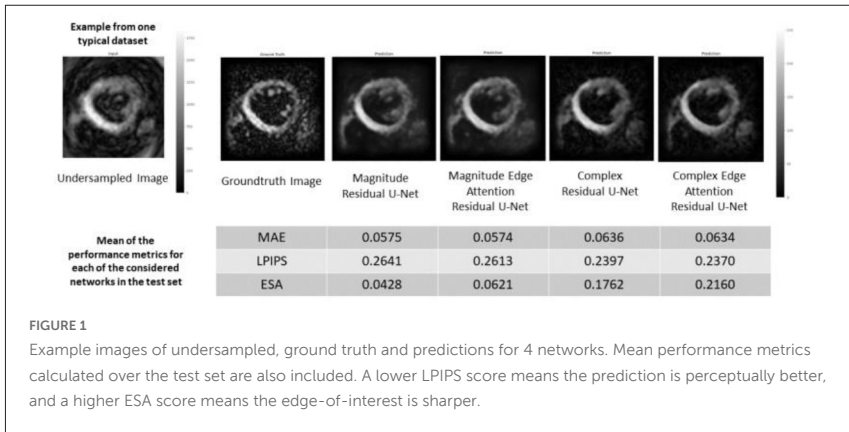
of view is $358 \times 134mm^2$. Multiple averages of six encoding directions and b-values of 0, 150 and $600 s/mm^2$ were acquired.

We resampled the STEAM-EPI acquisitions along variable density spiral trajectories with the central 25% of k-space fully-sampled, then linearly reducing until 55% of the radius, followed by an acceleration factor of 7, which is equivalent to a uniform 4-fold acceleration. This gives the simulated highly undersampled images.

The training-validation-testing split by subjects was 14 (4319 images), 3 (973 images), and 3 (980 images) respectively. Complex weight initialisation [19] and convolution with a kernel size of 3 were implemented. Data augmentation by a factor of 2 was implemented by applying random rotations to the spiral trajectories.

Results

Examples of predictions and test set performance metrics are shown in Fig. 1. Our proposed method outperforms other networks in Learned Perceptual Image Patch Similarity (LPIPS) [28], and is substantially better for Edge Sharpness Assessment (ESA), measured on the free wall epicardium [29],



which are around 10% and over 4 times respectively when compared against the magnitude network, but slightly worse in terms of Mean Absolute Error (MAE). In particular, our proposed model is able to recover the papillary muscles of the left ventricle which appear sharp, and its prediction closely resembles the structural information in the ground truth.

Example DT-CMR maps and pixelwise MAE for FA and MD for the four models are shown in Fig. 2. Our proposed method is able to produce DT-CMR results that most closely resemble the ground truth across FA, MD, and HA. Also, there is a decreasing trend in the medians of the MAE values, from magnitude networks to complex-valued networks.

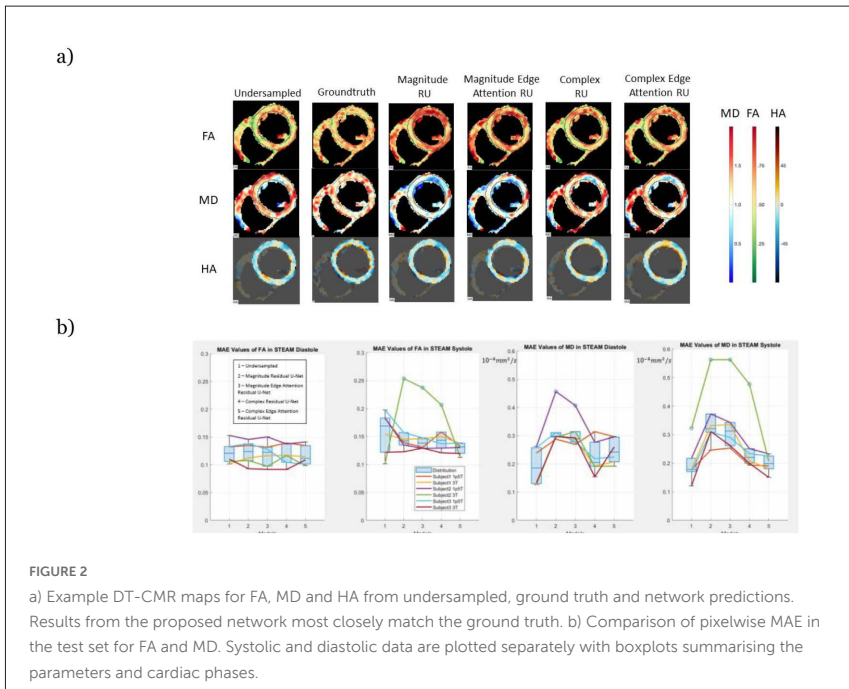


FIGURE 2

a) Example DT-CMR maps for FA, MD and HA from undersampled, ground truth and network predictions. Results from the proposed network most closely match the ground truth. b) Comparison of pixelwise MAE in the test set for FA and MD. Systolic and diastolic data are plotted separately with boxplots summarising the parameters and cardiac phases.

Discussion

Experiments

The proposed complex-valued deep neural network with SAM and EAM effectively reduces artefacts and preserves myocardial edge sharpness, better than magnitude networks or without the edge attention module. The network enhances high frequency components, leading to substantial increases in sharpness.

The medians of the MAE of DT-CMR parameters are in general lower for our proposed model, compared to other networks, suggesting that higher accuracies can be achieved with complex networks.

Limitations and Future Work

One limitation is that, while the MAE in MD is slightly higher for our network than the undersampled data, this is likely to be a consequence of the image scaling and should be resolved with further refinement of the training procedure.

A loss function that is dependent on the complex k-space data, and the magnitude of the prediction and ground truth in image space, was implemented. However, this might not be the best choice as training itself is performed in the complex domain. Complex image loss functions can be investigated, which have the potential to increase network performance and data accuracy, and achieve faster and more efficient training.

In order to strengthen the effectiveness of EAM, we could instead use two-dimensional gradient and adaptive edge detection filters, which might suppress the presence of artefacts and noise, and only keep the edges of the myocardium. This would allow the enhancement of details to be more intelligent.

Conclusion

We have proposed a complex-valued deep learning approach for artefact removal in highly undersampled DT-CMR images, that outperformed other

networks by implementing complex EAM and SAM. Our approach has potential future clinical utility in accelerating high resolution DT-CMR in patient cohorts.

References

- [1] Nielles-Vallespin, S., Khalique, Z., Ferreira, P.F., de Silva, R., Scott, A.D., Kilner, P., McGill, L.A., Giannakidis, A., Gatehouse, P.D., Ennis, D. and Aliotta, E.: Assessment of myocardial microstructural dynamics by in vivo diffusion tensor cardiac magnetic resonance. *Journal of the American College of Cardiology*, 69(6), pp.661-676. (2017)
- [2] Dou, J., Tseng, W.Y.I., Reese, T.G. and Wedeen, V.J.: Combined diffusion and strain MRI reveals structure and function of human myocardial laminar sheets in vivo. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 50(1), pp.107-113. (2003)
- [3] Ferreira, P.F., Kilner, P.J., McGill, L.A., Nielles-Vallespin, S., Scott, A.D., Ho, S.Y., McCarthy, K.P., Haba, M.M., Ismail, T.F., Gatehouse, P.D. and de Silva, R.: In vivo cardiovascular magnetic resonance diffusion tensor imaging shows evidence of abnormal myocardial laminar orientations and mobility in hypertrophic cardiomyopathy. *Journal of Cardiovascular Magnetic Resonance*, 16(1), pp.1-16. (2014)

- [4] Khaliq, Z., Ferreira, P.F., Scott, A.D., Nielles-Vallespin, S., Firmin, D.N. and Pennell, D.J.: Diffusion tensor cardiovascular magnetic resonance imaging: a clinical perspective. *Cardiovascular Imaging*, 13(5), pp.1235-1255. (2020)
- [5] Von Deuster, C., Sammut, E., Asner, L., Nordsletten, D., Lamata, P., Stoeck, C.T., Kozerke, S. and Razavi, R.: Studying dynamic myofiber aggregate reorientation in dilated cardiomyopathy using in vivo magnetic resonance diffusion tensor imaging. *Circulation: Cardiovascular Imaging*, 9(10), p.e005018. (2016)
- [6] Nguyen, C., Lu, M., Fan, Z., Bi, X., Kellman, P., Zhao, S. and Li, D.: Contrast-free detection of myocardial fibrosis in hypertrophic cardiomyopathy patients with diffusion-weighted cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 17(1), pp.1-7. (2015)
- [7] Nguyen, C., Fan, Z., Xie, Y., Dawkins, J., Tseliou, E., Bi, X., Sharif, B., Dharmakumar, R., Marbán, E. and Li, D.: In vivo contrast free chronic myocardial infarction characterization using diffusion-weighted cardiovascular magnetic resonance. *Journal of cardiovascular magnetic resonance*, 16(1), pp.1-10. (2014)
- [8] Nielles-Vallespin, S., Scott, A., Ferreira, P., Khaliq, Z., Pennell, D. and Firmin, D.: Cardiac diffusion: technique and practical applications. *Journal of Magnetic Resonance Imaging*, 52(2), pp.348-368. (2020)
- [9] Ferreira, P.F., Banerjee, A., Scott, A.D., Khaliq, Z., Yang, G., Rajakulasingam, R., Dwornik, M., De Silva, R., Pennell, D.J., Firmin, D.N. and Nielles-Vallespin, S.: Accelerating Cardiac Diffusion Tensor Imaging With a U-Net Based Model: Toward Single Breath-Hold. *Journal of Magnetic Resonance Imaging*, 56(6), pp.1691-1704. (2022)
- [10] Jones, D.K. and Basser, P.J.: "Squashing peanuts and smashing pumpkins": how noise distorts diffusion-weighted MR data. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 52(5), pp.979-993. (2004)

[11] Scott, A.D., Nielles-Vallespin, S., Ferreira, P.F., McGill, L.A., Pennell, D.J. and Firmin, D.N.: The effects of noise in cardiac diffusion tensor imaging and the benefits of averaging complex data. *NMR in Biomedicine*, 29(5), pp.588-599. (2016)

[12] Nayak, K.S., Cunningham, C.H., Santos, J.M. and Pauly, J.M.: Real-time cardiac MRI at 3 Tesla. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 51(4), pp.655-660. (2004)

[13] Tian, Y., Lim, Y., Zhao, Z., Byrd, D., Narayanan, S. and Nayak, K.S.: Aliasing artifact reduction in spiral real-time MRI. *Magnetic resonance in medicine*, 86(2), pp.916-925. (2021)

[14] Gorodetzky, M., Scott, A.D., Ferreira, P.F., Nielles-Vallespin, S., Pennell, D.J. and Firmin, D.N.: Diffusion tensor cardiovascular magnetic resonance with a spiral trajectory: An in vivo comparison of echo planar and spiral stimulated echo sequences. *Magnetic resonance in medicine*, 80(2), pp.648-654. (2018)

[15] Gorodetzky, M., Ferreira, P.F., Nielles-Vallespin, S., Gatehouse, P.D., Pennell, D.J., Scott, A.D. and Firmin, D.N.: High resolution in-vivo DT-CMR using an interleaved variable density spiral STEAM sequence. *Magnetic resonance in medicine*, 81(3), pp.1580-1594. (2019)

[16] van Gorkum, R.J., Guenther, C., Koethe, A., Stoeck, C.T. and Kozerke, S.: Characterization and correction of diffusion gradient-induced eddy currents in second-order motion-compensated echo-planar and spiral cardiac DTI. *Magnetic Resonance in Medicine*, 88(6), pp.2378-2394. (2022)

[17] Delattre, B.M., Heidemann, R.M., Crowe, L.A., Vallée, J.P. and Hyacinthe, J.N.: Spiral demystified. *Magnetic resonance imaging*, 28(6), pp.862-881. (2010)

[18] Hirose, A. and Yoshida, S.: Generalization characteristics of complex-valued feedforward neural networks in relation to signal coherence. *IEEE Transactions on Neural Networks and learning systems*, 23(4), pp.541-551. (2012)

- [19] Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J.F., Mehri, S., Rostamzadeh, N., Bengio, Y. and Pal, C.J.: Deep complex networks. *arXiv preprint arXiv:1705.09792*. (2017)
- [20] Jojoa, M., Garcia-Zapirain, B. and Percybrooks, W.: A Fair Performance Comparison between Complex-Valued and Real-Valued Neural Networks for Disease Detection. *Diagnostics*, 12(8), pp.1893. (2022)
- [21] Cole, E.K., Pauly, J. and Cheng, J.: Complex-valued convolutional neural networks for MRI reconstruction. In *Proc. 27th Annu. Meeting of ISMRM*, pp. 4714. (2019)
- [22] Dedmari, M.A., Conjeti, S., Estrada, S., Ehses, P., Stöcker, T. and Reuter, M.: Complex fully convolutional neural networks for MR image reconstruction. In *Machine Learning for Medical Image Reconstruction: First International Workshop, MLMIR 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 1*, pp. 30-38. Springer International Publishing. (2018)
- [23] Duan, C., Xiong, Y., Cheng, K., Xiao, S., Lyu, J., Wang, C., Bian, X., Zhang, J., Zhang, D., Chen, L. and Zhou, X.: Accelerating susceptibility-weighted imaging with deep learning by complex-valued convolutional neural network (ComplexNet): validation in clinical brain imaging. *European Radiology*, 32(8), pp.5679-5687. (2022)
- [24] Cole, E., Cheng, J., Pauly, J. and Vasanawala, S.: Analysis of deep complex-valued convolutional neural networks for MRI reconstruction and phase-focused applications. *Magnetic resonance in medicine*, 86(2), pp.1093-1109. (2021)
- [25] Hauptmann, A., Arridge, S., Lucka, F., Muthurangu, V. and Steeden, J.A.: Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning—proof of concept in congenital heart disease. *Magnetic resonance in medicine*, 81(2), pp.1143-1156. (2019)

[26] Penso, M., Babbaro, M., Moccia, S., Guglielmo, M., Carerj, M.L.: Giacari, C.M., Chiesa, M., Maragna, R., Rabbat, M.G., Barison, A. and Martini, N., Cardiovascular magnetic resonance images with susceptibility artifacts: artificial intelligence with spatial-attention for ventricular volumes and mass assessment. *Journal of Cardiovascular Magnetic Resonance*, 24(1), p.62. (2022)

[27] Nilles-Vallespin, S., Mekkaoui, C., Gatehouse, P., Reese, T.G., Keegan, J., Ferreira, P.F., Collins, S., Speier, P., Feiweier, T., De Silva, R. and Jackowski, M.P.: In vivo diffusion tensor MRI of the human heart: reproducibility of breath-hold and navigator-based approaches. *Magnetic resonance in medicine*, 70(2), pp.454-465. (2013)

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

[29] Ahmad, R., Ding, Y. and Simonetti, O.P.: Edge sharpness assessment by parametric modeling: application to magnetic resonance imaging. *Concepts in Magnetic Resonance Part A*, 44(3), pp.138-149. (2015)

Variation in mammography imaging equipment impacts artificial intelligence performance in breast cancer screening

Author

Clarisse F. de Vries – Aberdeen Centre for Health Data Science, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland

Samantha J. Colosimo – National Health Service Grampian (NHS), Aberdeen Royal Infirmary, Aberdeen, Scotland; School of Medicine, Medical Science and Nutrition, University of Aberdeen, Aberdeen, Scotland

Roger T. Staff – National Health Service Grampian (NHS), Aberdeen Royal Infirmary, Aberdeen, Scotland; School of Medicine, Medical Science and Nutrition, University of Aberdeen, Aberdeen, Scotland

Jaroslav A. Dymiter – Grampian Data Safe Haven (DaSH), Aberdeen Centre for Health Data Science, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland

Joseph Yearsley – Kheiron Medical Technologies, London, England

Deirdre Dinneen – Kheiron Medical Technologies, London, England

Moragh Boyle – Aberdeen Centre for Health Data Science, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland

David J. Harrison – School of Medicine, University of St Andrews, St Andrews, Scotland

Lesley A. Anderson – Aberdeen Centre for Health Data Science, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland

Gerald Lip – National Health Service Grampian (NHS), Aberdeen Royal Infirmary, Aberdeen, Scotland

on behalf of the iCAIRD Radiology Collaboration.

Citation

de Vries, C.F., Colosimo, S.J., Staff, R.T., Dymiter, J.A., Yearsley, J., Dinneen, D., Boyle, M., Harrison, D.J., Anderson, L.A., Lip, G. Variation in mammography imaging equipment impacts artificial intelligence performance in breast cancer screening.

Background

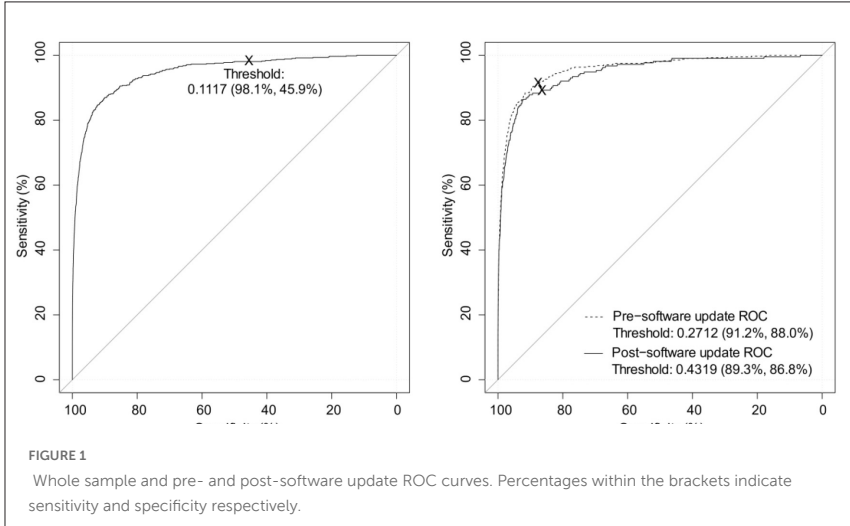
Artificial intelligence (AI) algorithms may assist breast screening mammography programmes. However, the United Kingdom (UK) National Screening Committee currently does not recommend using AI in NHS breast screening due to the insufficient quality and quantity of the supporting evidence (1-3). Previous research has been limited by evaluations performed on cancer-enriched samples in non-real-world settings. Furthermore, evidence on the generalisability of algorithms to new settings and possible sources of bias is limited. To address these limitations, the performance of a commercially available breast screening AI algorithm (4) was assessed using retrospective screening data.

Methods

This study used data from 22/02/2016 to 19/03/2020 from a UK regional screening programme with known clinical outcomes (75,823 cases; NHS Grampian). The stand-alone performance of the AI algorithm was evaluated with both pre-specified and locally calibrated decision thresholds. The pre-specified threshold was evaluated on the whole sample; the locally calibrated thresholds were evaluated on the test dataset, excluding cases used for threshold calibration. AI performance was quantified based on its sensitivity, specificity and recall rate. Positive cases were defined as histologically confirmed cancers detected through routine screening.

Results

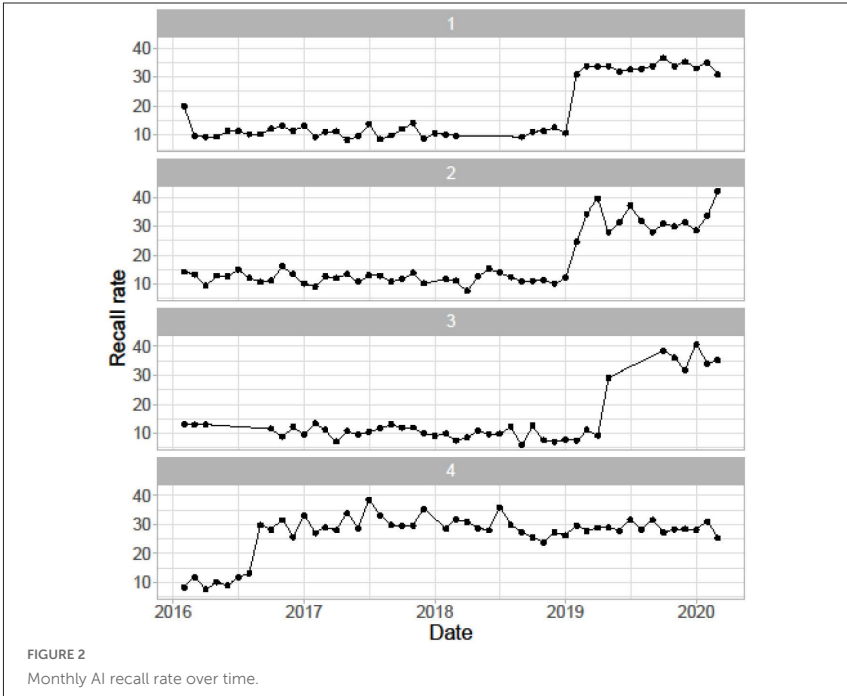
Fig. 1a shows the receiver operating characteristic (ROC) curve for the whole sample. The pre-specified threshold (0.1117) resulted in a sensitivity of 98.1%, specificity of 45.9% and recall rate of 54.5%. In response to the high recall rate, an initial locally calibrated threshold (0.2938) was generated. Overall, the new threshold reduced the recall rate. However, a sudden three-fold increase in recall rate was observed following the application of a software upgrade on the mammography imaging systems (Fig. 2). The software upgrades were applied to each of the imaging systems at different time points. Software-specific thresholds (0.2712 and 0.4319 respectively; Fig. 1b) reduced the overall recall rate to 13.3%, with sensitivity of 91.8% and



specificity of 87.2%. With the recalibrated thresholds, the AI algorithm would have recalled 379/413 (91.8%) of cancers detected through routine screening and 58/159 (36.5%) of cancers diagnosed between screening cycles (interval cancers).

Conclusion

Mammography equipment software updates can substantially worsen AI performance. A potential solution may include mammography equipment running different software for human and AI readers. This could reduce variability in AI performance and avoid disruption to the screening service. Different ways to monitor AI performance in real time should also be explored.



References

(1) Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 2021;374.

(2) Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for mammographic image analysis in breast cancer screening. *Rapid review and evidence map*. 2022 February.

(3) Taylor-Phillips S, Seedat F, Kijauskaite G, Marshall J, Halligan S, Hyde C, et al. UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. *The Lancet Digital Health* 2022;4(7):e558-e565.

(4) Sharma N, Ng AY, James JJ, Khara G, Ambrozay E, Austin CC, et al. Large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. medRxiv 2021.stylefix

Ensembles-based active learning for left ventricle segmentation

Author

Eman Alajrami, Jevgeni Jevsikov, Preshen Naidoo, Sara Adibzadeh, Patricia Fernandes, Nasim Dadashi Serej, Neda Azarmehr, Fateme Dinmohammadi, and Massoud Zolgharni – School of Computing and Engineering, University of West London, London, UK

Citation

Alajrami, E., Jevsikov, J., Naidoo, P., Adibzadeh, S., Fernandes, P., Serej, N.D., Azarmehr, N., Dinmohammadi, F., Zolgharni, M. Ensembles-based active learning for left ventricle segmentation.

Abstract

Left ventricle segmentation is essential in assessing cardiac function from 2D echocardiographic images. Active learning has proven effective in reducing annotation efforts and improving the accuracy of deep learning models. This study explores the use of established loss functions in ensemble-based active learning for left ventricle segmentation. The results show that applying this approach with the U-Net model outperforms existing ensemble-based methods, improving accuracy and reducing annotation efforts.

Introduction

Precise automated left ventricular (LV) assessment is crucial in cardiac diagnostics. The U-Net deep learning (DL) model employs Convolutional Neural Networks (CNNs) and has achieved impressive performance in medical image segmentation [11]. Several adaptations of the U-Net model, such as ResUnet and ResUnet++ [2, 6], have been developed; however, these

models require a large amount of annotated data. In the medical domain, obtaining sufficient labelled datasets can be expensive, time-consuming, and labour-intensive.

Deep active learning (DeepAL) combines active learning (AL) and DL, offering a solution for limited annotated data. AL selects informative samples from an unlabeled pool for annotation, reducing the time and cost of labelling the dataset while improving the DL model's accuracy [13]. AL selection strategies include uncertainty sampling [10,4], diversity sampling [9], and hybrid methods [12,5]. Ensemble-based methods are used as a query-by-committee approach to estimate a model's uncertainty by having multiple predictions on each sample and finding the data they disagree with. Monte Carlo dropout (MCD) is proposed as a Bayesian approximation for AL ensembles [3]. A recent study compared using ensembles of different DL models against MCD for uncertainty estimation [1]. The choice of loss functions is critical in DL [8]. This research explores the influence of loss functions on ensemble-based AL. It compares its performance against two state-of-the-art methods: MCD and ensemble-based using various DL architectures (U-Net, ResUnet and ResUnet++).

Datasets and Methodology

Two 2D echocardiographic datasets were used; CAMUS is a public dataset (450 images) [7], and Unity is obtained and ethically approved from the Imperial College Healthcare NHS Trust database (2094 images). Both datasets are split into 70%, 15%, and 15% for training, validation, and testing, respectively. We selected 10% (35 images in CAMUS) and 4% (82 images in Unity) of the training dataset as the initially labelled data. The size of the next selected batch to label was 5% and 1% of the total CAMUS and Unity datasets, respectively. The MCD U-Net architecture with dropout layers after each encoder and decoder block with a dropout d centre is used as our baseline model. Additionally, ResUnet and ResUnet++ are used for the ensemble-based method with different DL architectures. We customised their implementation by adding a dropout layer with a dropout probability of 0.3 after each encoder to have stochastic models. All models were implemented in TensorFlow and were trained on an Nvidia RTX3090 GPU.

A pool-based AL and different sampling methods were used for selecting the samples, including random and various uncertainty approaches, maximum entropy, variance, and Bayesian active learning with disagreement (BALD) [4, 1].

– Ensemble-based AL approaches

- **MCD ensemble-based MCD:** U-Net was trained on labelled data and used to predict each image from the unlabelled pool over 100 predictions to measure its uncertainty.
- **Multiple architecture ensemble-based:** U-Net, ResUnet, and ResUnet++ were trained on labelled data to predict unlabelled samples, with the predictions used to calculate uncertainty scores for each image.
- **Different loss functions ensemble-based:** Three ensembles of the U-Net with three different loss functions (Binary cross entropy (BCE), Dice, and BCE + Dice) were trained.

In each AL iteration, the labels of the most uncertain images are added to the training data for the next training.

Each approach was evaluated using the baseline model using the Dice Coefficient (DC) metric on the testing datasets at each AL iteration. Each selection strategy was trained three times, and the mean DC score at each iteration was calculated.

Results and Discussions

The U-Net baseline was trained on the whole annotations, achieving a mean DC of 0.946 and 0.914 on the CAMUS and Unity testing datasets, respectively.

The results show that our proposed method (using three loss functions) outperformed the random, MCD, and multiple architecture ensemble-based methods for all the uncertainty sampling techniques on both datasets.

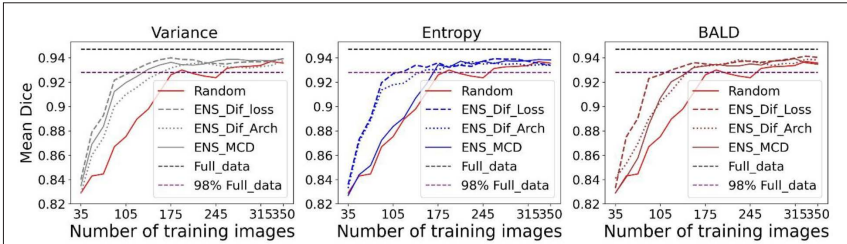


FIGURE 1

CAMUS performance profiles for various uncertainty strategies at each AL iteration using different ensemble-based methods. The black dashed lines represent the mean DC achievable with complete datasets, while the purple represents 98% of this performance.

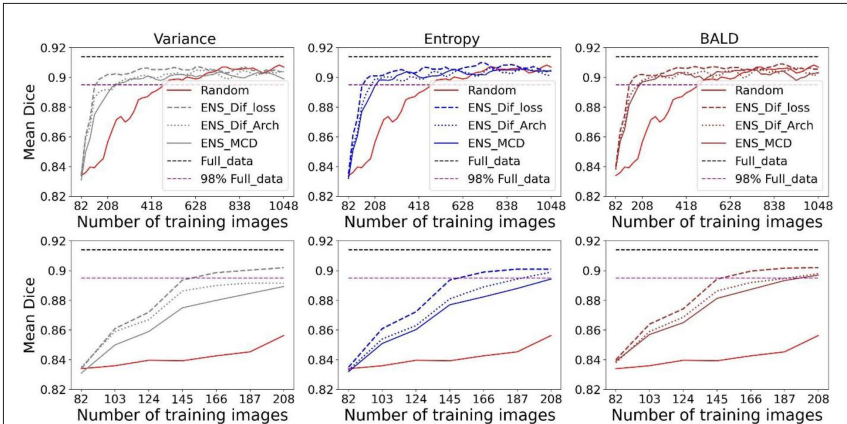


FIGURE 2

Unity performance for different ensemble-based methods; the lower panel shows a magnified version of early stages presented at the upper panel. The black dashed lines represent the mean DC achievable with complete datasets, while the purple represents 98% of this performance.

Fig. 1 demonstrates that our proposed ensemble-based AL approach on CAMUS crossed 98% of the maximum performance using less than 30% of annotations, 105 images, for all uncertainty methods. However, to achieve that performance using other ensemble-based methods, 50% and 40% of annotations are needed for entropy and BALD, respectively.

Fig. 2 shows that the introduced method outperformed the alternatives on Unity, achieving 98% of the maximum performance utilising 7% of annotations, 145 images. However, more than 10% of annotations, 208 images, are used to reach that performance using other approaches for all uncertainty strategies.

References

- [1] Beluch, W.H., Genewein, T., N ürnberger, A., Köhler, JM: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9368–9377 (2018)
- [2] Diakogiannis, F., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing 16, 94–114 (02 2020). <https://doi.org/10.1016/j.isprsjprs.2020.01.013>
- [3] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Proceedings of The 33rd International Conference on Machine Learning (06 2015)
- [4] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data.arXiv (03 2017).<https://doi.org/10.48550/arXiv.1703.02910>

- [5] Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(10), 1936–1949 (2014). <https://doi.org/10.1109/TPAMI.2014.2307881>
- [6] Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: *2019 IEEE International Symposium on Multimedia (ISM)*. pp. 225–2255. IEEE (2019)
- [7] Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging* 38(9), 2198–2210 (2019)
- [8] Naidoo, P., Alajrami, E.I., Lane, E.S., Jevsikov, J., Shun-shin, M.J., Francis, D.P., Zolgharni, M.: Influence of loss function on left ventricular volume and ejection fraction estimation in deep neural networks. In: *Medical Imaging with Deep Learning* (2021)
- [9] Nguyen, H., Smeulders, A.: Active learning using pre-clustering. *ICML* (2004). <https://doi.org/10.1145/1015330.1015349>
- [10] Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C., Goksel, O.: Active learning for segmentation based on Bayesian sample queries. *Knowledge-Based Systems* 214, 106531 (2021). <https://www.sciencedirect.com/science/article/pii/S0950705120306602>
- [11] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI 2015*. pp. 234–241. Springer International Publishing, Cham (2015)
- [12] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489* (2017)
- [13] Settles, B.: Active learning literature survey. *Computer Sciences Technical Report 1648*, University of Wisconsin–Madison (2009), <http://axon.cs.byu.edu/martinez/classes/778/Papers/settles.activelearning.pdf>

Weakly supervised pre-training for brain tumour segmentation using principal axis measurements of tumour burden

Author

Joshua Mckone – School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, UK

Tryphon Lambrou – School of Natural and Computing Sciences, University of Aberdeen, King's College, Aberdeen, AB24 3FX, UK

Xujiong Ye – School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, UK

James Brown – School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, UK

Citation

Mckone, J., Lambrou, T., Ye, X., Brown, J. Weakly supervised pre-training for brain tumour segmentation using principal axis measurements of tumour burden.

Abstract

State-of-the-art methods for multi-modal brain tumour segmentation frequently rely on large quantities of individually annotated data. There are many settings in which such labelled data may be scarce, where instead there may be value in exploiting cheaper to acquire data, such as Response Assessment in Neuro-Oncology (RANO). This work demonstrates the utility of such measurements for multi-modal brain tumour segmentation, whereby an encoder network is first trained to regress synthetic “pseudo-RANO”

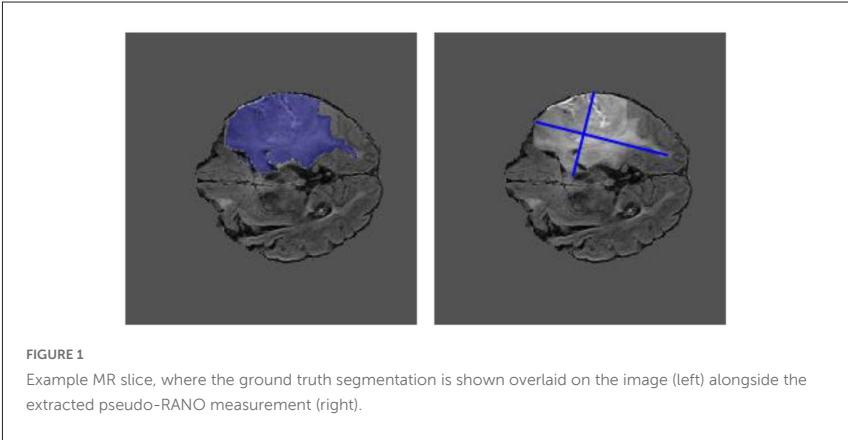
measurements, utilising a dual loss function of mean squared error (MSE) in addition to cosine similarity loss which promotes orthogonality of the principal axes, followed by whole tumour segmentation via the combination of a pre-trained encoder appended to a randomly initialised decoder to form a U-Net type architecture. Our results demonstrate that weakly supervised encoder models converge faster than those trained without any such pre-training, and help to minimise the annotation burden when trained to perform segmentation.

Introduction

Gliomas are one of the most common primary brain malignancies, with a survival rate of 39.7% over the period of a year, and 5% - 9.8% over the period of 5 years [1] [2]. MRI (Magnetic Resonance Imaging) is widely used in glioma assessment, in which quantification of tumour burden remains a key technical and clinical challenge that has prompted the development of automated image analysis techniques. Segmentation is a major step in automated pipelines, a process whereby the tumour region is delineated to provide area or volume estimates to aid in treatment response monitoring and surgical planning. Weak supervision invokes the principle that models trained on noisy or imprecise annotations can still learn meaningful representations, alleviating the burden of manual annotation. To validate the approach in this paper, we derive a pseudo-measure similar to RANO, "pseudo-RANO", from the labels supplied as part of the BraTS (Brain Tumour Segmentation Challenge) dataset [3]. The ultimate aim is to improve both the performance and clinical viability of such models and provide a framework that can be applied in many contexts with only minor changes to the architectures used.

Methodology

The BraTS 2018 dataset is used for model development and evaluation [4] [5] [3]. Pseudo-RANO measurements of whole tumour burden were utilised in training of the encoder network, operating on individual slices (Figure 1). Tumour principal axes were extracted using the approach outlined by [6].



The encoder portion of the U-net architecture [7] is trained independently from the decoder to annotate pseudo-RANO measurements, as a pre-training pathway before whole tumour region segmentation, which fully utilises the whole encoder-decoder architecture. Each pseudo-RANO measure is defined as a vector $v \in R^8$ which represents the principal axes of a tumour region,

$$v_i = \{(x_1, y_1) \dots (\tilde{x}_4, \tilde{y}_4) \in R^8\},$$

(x_1, x_2) represent the co-ordinates of a pixel on the tumour boundary. We formulate pseudo-RANO estimation as a regression problem by minimising a mean squared error (MSE) loss and enforcing orthogonality of the major and minor axis through minimising cosine (dis)-similarity. We refer to the major and minor principal axis as line segments v^{maj} and v^{min} , respectively,

$$v_{1,2}, v_{3,4}: v^{min}, \tilde{v}^{maj},$$

with $\|v^{min} < v^{maj}\|$. In addition to MSE, a cosine similarity loss is computed to enforce a degree of orthogonality between v^{min} and v^{maj} . In

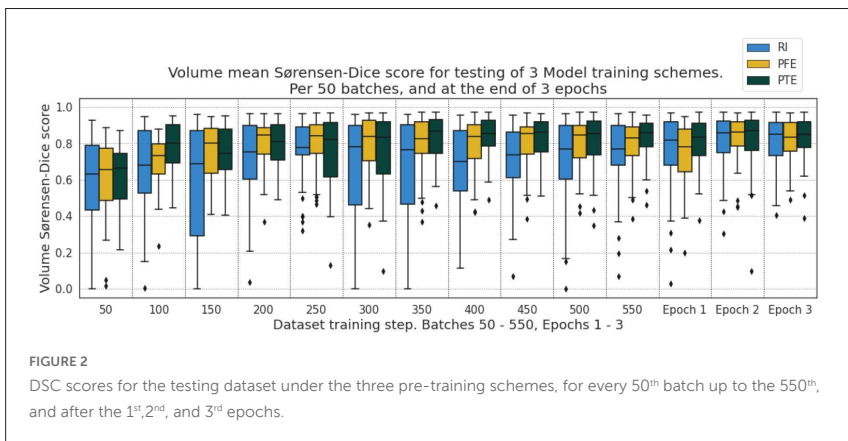
addition to MSE, a cosine similarity loss is computed to enforce a degree of orthogonality between \mathbf{v}^{min} and \mathbf{v}^{maj} , the combined loss is then calculated with a weighting ω to balance the relative contributions of the two terms,

$$\mathcal{L} = \mathcal{L}_{MSE} + \mathcal{L}_{Orth} \cdot \omega$$

Results

In total, 210 HGG (High Grade Glioma) and 75 LGG (Low Grade Glioma) MRI volumes were used, randomly split into 70% for training, 10% for validation and 20% for testing. Augmentation was used via horizontal and vertical flips, rotation, and scaling. Training of the segmentation task using the full U-Net model is then investigated under three different schemes. They are defined as: the baseline RI (Randomly Initialized) scheme with no encoder pre-training, the PFE (Pre-trained Frozen Encoder) scheme where the encoder weights from regression are used but not fine-tuned during further training, and the PTE (Pre-trained Trainable Encoder) scheme where the encoder weights from regression are used and allowed to update during further training.

The overall behaviour of the three training schemes is summarised in Figure 2. In general, we observe tighter distributions of DSC (Sørensen-Dice Score) for



PTE compared to the RI and PFE schemes. Some cases show the PFE slightly outperforming the PTE for some examples early on, particularly between the 200th and 300th batches.

Conclusion

In this paper, we demonstrate a pre-training approach for lesion segmentation based on estimation of the lesion's principal axis. The "pseudo-RANO" measures used in this work are relatively cheap and straightforward to obtain, and could serve as an alternative to dense segmentation masks as training data. We further observe encoder pre-training appearing to facilitate the transfer of features to enable faster convergence and modest improvements in overall segmentation performance. In particular, we observe that a pre-trained, trainable encoder offers optimal performance when compared to randomly initialised models.

Acknowledgements

This work has been supported by the EPSRC Doctoral Training Partnership EP/T518177/1.

References

- [1] U. N. Chukwueke et al. "Use of the Response Assessment in Neuro-Oncology (RANO) criteria in clinical trials and clinical practice," *CNS oncology*, vol. 8, p. CNS28, 2019.
- [2] C. Villa, et al. "The 2016 World Health Organization classification of tumours of the central nervous system," *La Presse Médicale*, vol. 47, p. e187–e200, 2018.

[3] B. H. Menze, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, p. 1993–2024, 2014.

[4] S. Bakas, et al. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 1–13, 2017.

[5] S. Bakas, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[6] K. Chang, et al. "Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement," *Neuro-oncology*, vol. 21, p. 1412–1422, 2019.

[7] O. Ronneberger, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

Enhancing generalization of CNN models for breast lesion classification from ultrasound images

Author

Tahir Hassan, Hongbo Du, Sabah Jassim – School of Computing, The University of Buckingham, Buckingham, MK18 1EG, UK

Citation

Hassan, T., Du, H., Jassim, S. Enhancing generalization of CNN models for breast lesion classification from ultrasound images.

Abstract

This proof-of-concept study is concerned with the frequently occurring problem: CNN models trained and tested on breast ultrasound (BUS) images from one clinical centre, fail to generalize for unseen data from other sources. We propose a new augmentation scheme, based on convolutions by Hadamard-based filters for training CNN models to improve their generalization capability. To form a larger training dataset, the augmentation procedure utilizes image convolution with only six 5x5 Hadamard filters on the images of tightly cropped tumour Region of Interest (RoI). We trained and tested 4 well-known CNN architectures, in fine-tuning mode, on the augmented dataset following a 5-fold-cross-validation. We then determine the generalization errors of the 5 models on images from another dataset, before and after applying the augmentation scheme. The results demonstrate significant enhancement of these CNN models' generalization rates post augmentation.

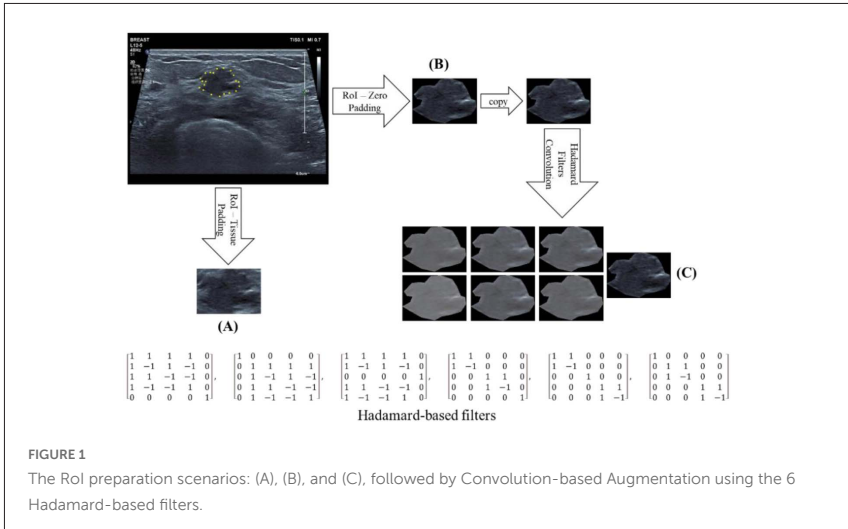
Introduction

Breast cancer is a major global health concern, and early detection is crucial for increasing survival rates [1]. Ultrasound (US) imaging is widely used for breast cancer screening due to its non-invasive nature and availability. However, accurately differentiating between benign and malignant lesions remains challenging. Deep learning (DL) algorithms that utilize convolutional neural networks (CNNs) have shown promise in automatically classifying breast lesions from US images, potentially enhancing diagnosis accuracy and efficiency [2–4].

Effective DL models for BUS image analysis require a large, and well-annotated dataset, which is difficult to obtain [5]. Transfer learning with pre-trained CNN models and image augmentation techniques are frequently used to address this challenge [3, 6–11]. However, using these methods to develop a DL model on a small dataset from a single medical centre still has limitations in sample diversity and potential bias that can impact model performance and generalizability. This study aims to enhance the generalization ability of DL models trained on a relatively small dataset from a single medical centre by convolution-based augmentation with Hadamard kernel filters. The results demonstrate that this effective image augmentation technique significantly enhances the DL model's generalizability on an external testing dataset.

Materials and Methods

We utilize 2 datasets: Renmin and BUSI. The Renmin dataset contains 524 balanced BUS images from Pudong New District Renmin Hospital in Shanghai, China. The lesion boundary points were marked by an experienced radiologist, and the class label of the lesions was confirmed by biopsy tests [2]. The dataset is used to develop/train the DL models. The BUSI dataset also contains 524 BUS images (343 benign and 181 malignant lesions) after removing those without lesions or with severe artefacts. The dataset was obtained from Baheya Hospital for Early Detection and Treatment of Women's Cancer in Cairo, Egypt, and labelled with given masks and lesion status [12]. This dataset is used as an external test (Generalization test).



During training, images are pre-processed using 3 scenarios: **(A)** Conventional Rol cropping, whereby the minimum surrounding rectangular box around the marked Rol tumour area is tissue padded; **(B)** Tight Rol cropping, whereby the region between the surrounding box and the marked Rol tumour area is zero-padded [2]; and **(C)** Each tightly cropped Rol images (as in **(B)**) is augmented by convolving with six 5x5 Hadamard-based filters. These 5x5 filters are created by block diagonals of 1x1, 2x2 and 4x4 Hadamard matrices [13, 14] (See Fig. 1). For each training scenario, the test/ BUSI images are only pre-processed by the adopted scenario.

Hadamard filters are well-conditioned by design (condition numbers ≤ 2) and their choice for convolution-based augmentation is influenced by the fact that their action on image patches is not sensitive to tolerable perturbation of pixel values. Gaussian filters can be used but their random generation may result in highly ill-conditioned filters.

Experimental Results, Discussion and Conclusion

For the experiments, 4 pre-trained CNN architectures, including (AlexNet, VGG16, VGG19, and ResNet18), are fine-tuned with 10 epochs training. The 5-fold cross-validation was used as the training/testing protocol. Table 1, below, presents the classification performance of the 5-fold trained DL models on the BUSI external testing dataset, For the pre-processing (A, B, and C) scenarios, displaying the Ave ± Stdev of accuracy, sensitivity, specificity, F1-score, and AUC rates.

TABLE 1: CNN model performance on BUSI dataset with the scenarios: (A), (B), and (C)

(A)	Accuracy	Sensitivity	Specificity	F1-score	AUC
	Ave ± Stdev	Ave ± Stdev	Ave ± Stdev	Ave ± Stdev	Ave ± Stdev
AlexNet	0.71 ± 0.04	0.99 ± 0.01	0.56 ± 0.06	0.70 ± 0.03	0.77 ± 0.03
VGG16	0.72 ± 0.03	0.99 ± 0.01	0.58 ± 0.05	0.71 ± 0.02	0.78 ± 0.02
VGG19	0.75 ± 0.02	0.99 ± 0.00	0.63 ± 0.03	0.73 ± 0.02	0.81 ± 0.02
ResNet18	0.68 ± 0.05	0.98 ± 0.01	0.52 ± 0.08	0.68 ± 0.03	0.75 ± 0.03
(B)					
AlexNet	0.89 ± 0.03	1.00 ± 0.00	0.84 ± 0.05	0.87 ± 0.03	0.92 ± 0.02
VGG16	0.78 ± 0.09	1.00 ± 0.00	0.66 ± 0.14	0.76 ± 0.07	0.83 ± 0.07
VGG19	0.82 ± 0.04	1.00 ± 0.00	0.72 ± 0.06	0.79 ± 0.04	0.86 ± 0.03
ResNet18	0.66 ± 0.08	1.00 ± 0.00	0.49 ± 0.13	0.68 ± 0.05	0.74 ± 0.06
(C)					
AlexNet	0.90 ± 0.03	0.99 ± 0.00	0.85 ± 0.05	0.88 ± 0.03	0.92 ± 0.02
VGG16	0.92 ± 0.02	1.00 ± 0.00	0.88 ± 0.03	0.90 ± 0.02	0.94 ± 0.01
VGG19	0.88 ± 0.03	1.00 ± 0.00	0.81 ± 0.04	0.85 ± 0.03	0.90 ± 0.02
ResNet18	0.83 ± 0.05	0.99 ± 0.00	0.74 ± 0.08	0.80 ± 0.05	0.86 ± 0.04

The results show that the accuracy of all the CNN models has improved from (A) to (B) and then to (C). Specifically, under (A), the models have an average accuracy of 0.72, whereas under (B), the average accuracy has increased to 0.79, and under (C), it has further improved to 0.88. The above pattern of improved accuracy is almost replicated for F1 score and AUC rates. While all scenarios achieve optimal sensitivity rates, specificity rates are less impressive. All models perform in an unbalanced way with varying gaps between sensitivity and specificity, but the gap is narrowest in the (C) scenario. The VGG16 model benefitted most from the augmentation. The ResNet18 model performance improved significantly as a result of augmentation in comparison to its disappointing performance in scenarios (A) and (B). These results compare favourably with other conventional augmentation schemes, but due to restricted space, results of the other schemes are included in the pending PhD thesis of the first author.

In conclusion, we have shown that convolution-based augmentation with Hadamard filters applied on tightly cropped Rols significantly improves the generalization of DL models for BUS lesion classification. This approach can be further exploited to train CNN models from scratch when the availability of training data is limited (e.g., medical image analysis tasks) without incurring significant overfitting challenges.

References

- [1] Breast Cancer Statistics | How Common Is Breast Cancer?, <https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html>, last accessed 2023/05/05.
- [2] Hassan, T., Alzoubi, A., Du, H., Jassim, S.: Towards optimal cropping: breast and liver tumor classification using ultrasound images. In: Aghaian, S.S., Jassim, S.A., DelMarco, S.P., and Asari, V.K. (eds.) *Multimodal Image Exploitation and Learning 2021*. p. 15. SPIE, Online Only, United States (2021). <https://doi.org/10.1117/12.2589038>.

[3] Hassan, T., AlZoubi, A., Du, H., Jassim, S.: Ultrasound image augmentation by tumor margin appending for robust deep learning based breast lesion classification. In: Aghaian, S.S., Jassim, S.A., DelMarco, S.P., and Asari, V.K. (eds.) *Multimodal Image Exploitation and Learning 2022*. p. 8. SPIE, Orlando, United States (2022). <https://doi.org/10.1117/12.2618656>.

[4] Han, S., Kang, H.-K., Jeong, J.-Y., Park, M.-H., Kim, W., Bang, W.-C., Seong, Y.-K.: A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Phys. Med. Biol.* 62, 7714–7728 (2017). <https://doi.org/10.1088/1361-6560/aa82ec>.

[5] Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T.: Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering*. 5, 261–275 (2019). <https://doi.org/10.1016/j.eng.2018.11.020>.

[6] Daoud, M.I., Abdel-Rahman, S., Bdair, T.M., Al-Najar, M.S., Al-Hawari, F.H., Alazrai, R.: Breast Tumor Classification in Ultrasound Images Using Combined Deep and Handcrafted Features. *Sensors*. 20, 6838 (2020). <https://doi.org/10.3390/s20236838>.

[7] Cao, Z., Duan, L., Yang, G., Yue, T., Chen, Q.: An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC Medical Imaging*. 19, 51 (2019). <https://doi.org/10.1186/s12880-019-0349-x>.

[8] Zeimarani, B., Costa, M.G.F., Nurani, N.Z., Bianco, S.R., De Albuquerque Pereira, W.C., Filho, C.F.F.C.: Breast Lesion Classification in Ultrasound Images Using Deep Convolutional Neural Network. *IEEE Access*. 8, 133349–133359 (2020). <https://doi.org/10.1109/ACCESS.2020.3010863>.

[9] Ahmed, M., AlZoubi, A., Du, H.: Improving Generalization of ENAS-Based CNN Models for Breast Lesion Classification from Ultrasound Images. In: Papiież, B.W., Yaqub, M., Jiao, J., Namburete, A.I.L., and Noble, J.A. (eds.) *Medical Image Understanding and Analysis*. pp. 438–453. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-80432-9_33.

[10] Shorten, C., Khoshgoftaar, T.M.: A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*. 6, 60 (2019). <https://doi.org/10.1186/s40537-019-0197-0>.

[11] Goceri, E.: Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev.* (2023). <https://doi.org/10.1007/s10462-023-10453-z>.

[12] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. *Data in Brief*. 28, 104863 (2020). <https://doi.org/10.1016/j.dib.2019.104863>.

[13] Aгаian, S.S. ed: Hadamard transforms. SPIE, Bellingham, Wash (2011).

[14] Hassan, T.M.: Data-independent vs. data-dependent dimension reduction for pattern recognition in high dimensional spaces, <http://bear.buckingham.ac.uk/199/>, (2017).

Investigation of the structural characteristics of the extracellular matrix

Author

Youssef Arafat – giCentre, Department of Computer Science, City, University of London, UK

Cristina Cuesta Apausa – Tumour–Stroma Signalling Lab, Universidad de Salamanca, Salamanca, Spain

Esther Castellano – Tumour–Stroma Signalling Lab, Universidad de Salamanca, Salamanca, Spain

Constantino Carlos Reyes–Aldasoro – giCentre, Department of Computer Science, City, University of London, UK

Citation

Arafat, Y., Apausa, C.C., Castellano, E., Reyes-Aldasoro, C.C. Investigation of the structural characteristics of the extracellular matrix.

Introduction

Whilst the morphological characteristics of cells have been widely analysed in cancer and many other conditions [1], the microenvironment that surrounds these cells is now recognised to play a major role [2]. Many studies now focus on elements like vasculature, macrophages, angiogenesis or soluble factors [3]. However, less attention has been given to the extracellular matrix (ECM). Perhaps this is due to the more complicated nature of this 3D network, which itself consist of many elements like collagen, enzymes and glycoproteins. One approach to investigate the ECM used 3D printed inclined surfaces to tilt the petri dishes where cells were grown to adjust the

fibril alignment [5]. An increase of the inclination of the surface promoted an increase in the alignment of the collagen as well as an increase in the elastic modulus. Another study investigated whether fibronectin fibres remain aligned after the removal of cells or not [6]. The results showed that the removal of cells using a decellularization procedure did not affect the alignment of the fibrils. This was analysed using Fast Fourier Transform analysis prior to, and after the decellularization procedure. The designing of implant surfaces to enhance the interaction between stem cells and implant material was studied in [7]. The study used Fibronectin as the ligand to promote cell adhesion and migration. The width and the gap sizes of the fibronectin lines were modified to observe how these affected the behaviour of the cells. It was noticed that as the size of the fibronectin fibres decreased, the speed and direction of cell migration increased.

In this work, the structural characteristics of the ECM, as observed by the content of fibronectin are explored. We conjecture that the structural features of these glycoproteins can reveal interesting characteristics related to conditions of health and disease. To this effect, *in vitro* experiments showed that removal of RAS-PI3K interaction in fibroblasts severely impairs their ability to form aligned ECM. To better understand the role of this signalling pathway in the architecture of ECM, fluorescent images were acquired and then analysed. Using image processing algorithms, edges were detected, their orientations and lengths measured and the gaps space between them calculated. Two of these metrics showed statistical difference between two populations (WT $n=5$, RBD $n=5$). Whilst the dataset is small, the results encourage further experiments.

Materials and Methods

Cells and Images

Cells were prepared as described previously [4], but briefly. p110 α Ras Binding Domain-mutant (RBD) and Wild Type (WT) fibroblasts were grown on a gelatine layer crosslinked with glutaraldehyde and used as a substrate to produce the ECM. For each case, 5 sets of fluorescent images for expression of DAPI, Phalloidin and Fibronectin were acquired (Figure. 1)

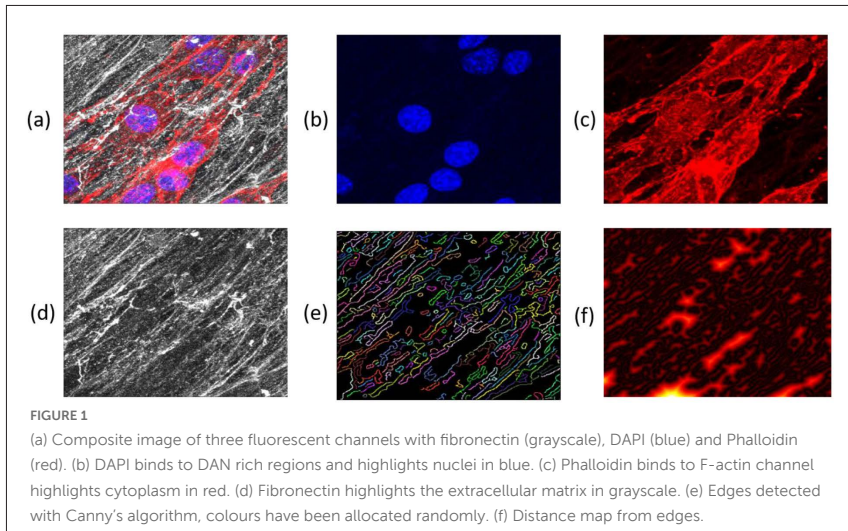


Image Processing of the Fibronectin

Canny edge detection was applied to the grayscale images of fibronectin. Mean and standard deviation of orientation of all edges was calculated (first metric Figure. 2a). Average major and minor axis lengths for the all the edges and the absolute values in pixels were divided by the number of columns of the image to obtain relative values (second metric Figures. 1e, 2b). Finally, a distance transform was obtained calculating the Euclidean distance of every pixel not covered by edges (called gaps) to its closest edge, and the relative area of the gaps was calculated (third metric Figures. 1f, 2c).

Results and Discussion

Two sample *t*-tests provided the following *p*-values: 0.490, 0.035, 0.002 for the orientation, edges and gaps respectively. Whilst the samples analysed in this work are small, the results are encouraging and motivate a further examination of the extracellular matrix in cancer studies.

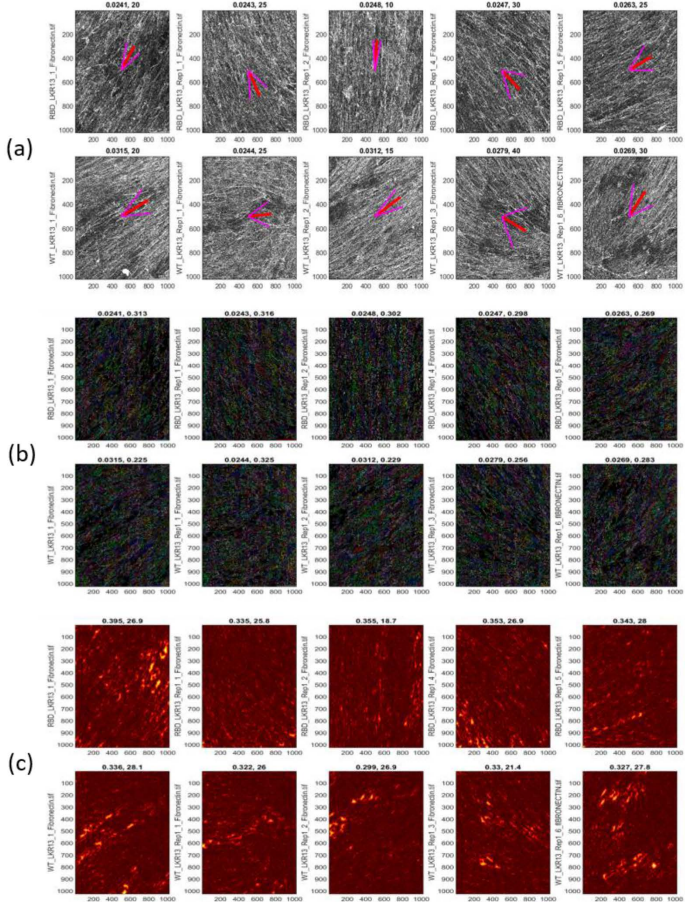


FIGURE 2

Analysis of the extracellular matrix with three different approaches. In all cases top 5 images correspond to RBD mutants and bottom 5 to WT cells. (a) Variance in the fibre orientation is illustrated with one red line and two magenta lines with the angle corresponding to standard deviation. (b) Edges detected with Canny's algorithm, edges have been allocated random colours for discrimination purposes. (c) Distance maps growing from the edges. Brighter regions are further away from edges.

References

- [1] Hu, M., Li, M., Huang, H., Lu, C.: Isolated cancer stem cells from human liver cancer: morphological and functional characteristics in primary culture. *Clin. Transl. Oncol.* 24, 48–56 (2022). <https://doi.org/10.1007/s12094-021-02667-w>.
- [2] Fukumura, D., Duda, D.G., Munn, L.L., Jain, R.K.: Tumor Microvasculature and Microenvironment: Novel Insights Through Intravital Imaging in Pre-Clinical Models. *Microcirculation.* 17, 206–225 (2010). <https://doi.org/10.1111/j.1549-8719.2010.00029.x>.
- [3] Balkwill, F.R., Capasso, M., Hagemann, T.: The tumor microenvironment at a glance. *J. Cell Sci.* 125, 5591–5596 (2012). <https://doi.org/10.1242/jcs.116392>.
- [4] Gupta, S., Ramjaun, A.R., Haiko, P., Wang, Y., Warne, P.H., Nicke, B., Nye, E., Stamp, G., Alitalo, K., Downward, J.: Binding of Ras to Phosphoinositide 3-Kinase p110 α Is Required for Ras- Driven Tumorigenesis in Mice. *Cell.* 129, 957–968 (2007). <https://doi.org/10.1016/j.cell.2007.03.051>.

[5] Sapudom, J., Karaman, S., Quartey, B.C., Mohamed, W.K.E., Mahtani, N., Garcia-Sabaté, A., Teo, J.: Collagen Fibril Orientation Instructs Fibroblast Differentiation Via Cell Contractility. *Adv. Sci.* n/a, 2301353. <https://doi.org/10.1002/advs.202301353>.

[6] Garrison, C.M., Schwarzbauer, J.E.: Fibronectin fibril alignment is established upon initiation of extracellular matrix assembly. *Mol. Biol. Cell.* 32, 739–752 (2021). <https://doi.org/10.1091/mbc.E20-08-0533>.

[7] Kasten, A., Naser, T., Brüllhoff, K., Fiedler, J., Müller, P., Möller, M., Rychly, J., Groll, J., Brenner, R.E.: Guidance of Mesenchymal Stem Cells on Fibronectin Structured Hydrogel Films. *PLoS ONE.* 9, e109411 (2014). <https://doi.org/10.1371/journal.pone.0109411>.

Assessing robustness of network-based correlation analysis with preclinical total-body PET data

Authors

Abigail F. Hellman – School of Physics and Astronomy, University of Edinburgh, Edinburgh, United Kingdom

Paul S. Clegg – School of Physics and Astronomy, University of Edinburgh, Edinburgh, United Kingdom

Adriana A. S. Tavares – University/British Heart Foundation (BHF) Centre for Cardiovascular Science, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, United Kingdom ; Edinburgh Imaging, University of Edinburgh, Edinburgh, United Kingdom

Citation

Hellman, A.F., Clegg, P.S., Tavares, A.A.S. Assessing robustness of network-based correlation analysis with preclinical total-body PET data.

Introduction

While traditional PET is limited to a small axial field-of-view (FOV) and produces images with low resolution that require long scan times and a significant radiation dose delivered to the patient, total-body PET takes advantage of an extended FOV that covers the entire body. Extending the coverage of the PET scanner increases the image resolution, which can allow for either quicker or reduced-dose scanning. An important research benefit is that it provides a scan that shows how the entire body is responding physiologically to a radiotracer at one time. This is advantageous for studying systems biology, a rapidly growing field involving the study of complex interactions between tissues, organs, and organ systems within the body.

Like most traditional PET scanners, total-body PET can be co-registered with a structural imaging modality, such as computed tomography (CT), for the precise overlay of metabolic and anatomic data [2]. Network analysis is a method for image analysis that is particularly helpful in conjunction with total-body PET as a research tool for understanding systems biology and the systemic effect of disease. Networks can be created based on an assessment of the correlations between activity values in different tissues, where highly correlated tissues are taken to be connected in the network. While network analysis has been used extensively in single-organ studies, particularly of the brain, it is largely unexplored for multi-organ and whole-body research. Thus, in order to perform future studies using network analysis with total-body PET, it is necessary first to assess if the networks are robust. Robustness refers to the structural integrity of a network following changes in the data at either a local or global scale [1]. If networks are similar for similar subjects and the structure of the network is maintained when a population average of the data for each organ is used, then the displayed connections may be biologically significant for the study population. The robustness of network analysis with total-body PET was tested using dynamic scans of four healthy mice following intravenous bolus injection of [^{18}F]Fluorodeoxyglucose ([^{18}F]FDG) as a way to assess glucose metabolism in bones. Bone was chosen for analysis because the skeletal system provides a good model for studying complex interactions, as it serves multiple purposes including organ protection, allowing for motion, and the formation of blood cellular components, along with recently discovered major endocrine functions [6]. Additionally, there is existing published data regarding network analysis of bone metabolism that can be used for comparison [6].

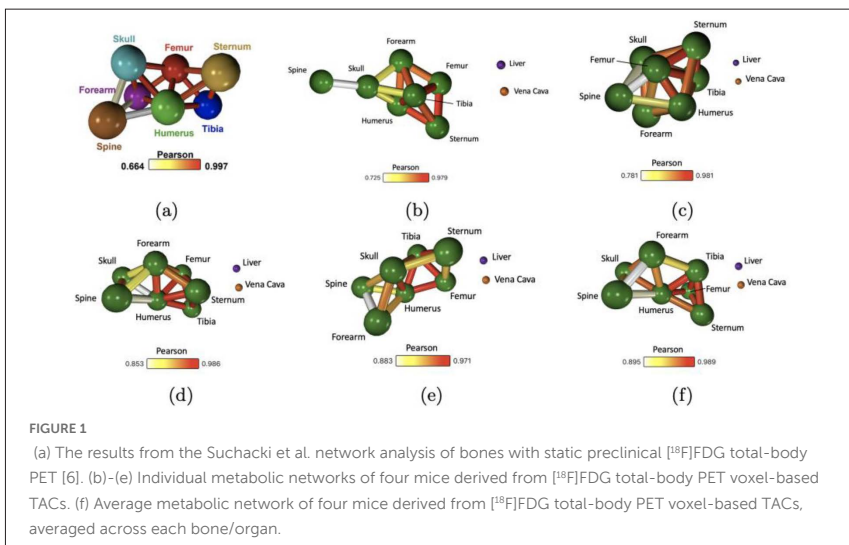
Methods

Total-body PET/CT scans of four healthy mice were acquired through the Edinburgh Preclinical Imaging Facility at the University of Edinburgh. These scans were then processed using PMOD 3.7 software (PMOD Technologies LLC, Switzerland). Two to three cuboid volumes of interest (VOIs) were placed in each bone, 1x1x1 millimetre in size. Each cube then contained twenty-seven voxels, as determined by the scanner resolution. Additionally, one cube each was added

to the vena cava and the liver to act as controls, as both are expected to behave quite differently from each other and from bone, being part of different organ systems. The time-activity curves (TACs), in kBq/cc versus seconds, were then extracted for every voxel in every cube. For each mouse, the voxel data within a cube was averaged, each cube within a bone was averaged, and left and right bones were averaged. This provided one time-activity curve per bone/organ per mouse. This data was then loaded into the network analysis tool Graphia [4], using the Pearson correlation coefficient (PCC) to determine the connections. PCC quantifies linear correlation on a scale from negative one (anti-correlation) to one (perfect correlation) [3]. A network was also created from averaged TACs for each organ across all four mice. In the final networks (see Figure 1), the nodes represent each bone or organ and the colour-coded edges between them represent the PCC between the nodes [4].

Results

The four individual mice networks can be seen in Figures 1b, 1c, 1d and 1e. While they may all look a bit different from each other, there are some



general properties that remain consistent and agree with the previously published results shown in Figure 1a, which were derived with a different image processing method [6]. In each mouse, there is high connectivity between long bones, whereas the spine has some connectivity to the skull but very minimal connectivity to other bones [6]. This provides independent confirmation of the role of long bones in glucose metabolism [7]. The humerus was found to consistently be a central feature of each network. Additionally, the liver and the vena cava were both found to be uncorrelated to each other and to each of the bones, which agrees with expectations. The averaged network (Figure 1f) also displays these properties and agrees with the expected results.

Discussion

So far, the networks created with voxel-based analysis of [^{18}F]FDG total-body PET scans are robust. More work is necessary to further confirm this, though. Future steps will involve adding more mice to the cohort then comparing the individual and averaged networks again. Additionally, a network will be created with all the data from each mouse (no averaging). If these networks continue to be robust and in agreement with existing results, then the method will be applied to more scans, such as those involving the radiotracer [^{18}F]Sodium Fluoride, which is commonly used to image bones as it targets calcification in the body [5]. If this method of image analysis continues to provide robust networks, then it will be used to study systems biology in healthy and diseased mice and humans.

References

- [1] Bullmore, E.T., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* **10**(3), 186–198 (Mar 2009). <https://doi.org/10.1038/nrn2575>
- [2] Cherry, S.R., Jones, T., Karp, J.S., Qi, J., Moses, W.W., Badawi, R.D.: Total-body PET: Maximizing sensitivity to create new opportunities for clinical research and patient care. *Journal of Nuclear Medicine* **59**(1), 3–12 (Jan 2018). <https://doi.org/10.2967/jnumed.116.184028>

[3] Freedman, D., Pisani, R., Purves, R.: Statistics, chap. 8 Correlation. W.W. Norton & Company (Feb 2007)

[4] Freeman, T.C., Horsewell, S., Patir, A., Harling-Lee, J., Regan, T., Shih, B.B., Prendergast, J., Hume, D.A., Angus, T.: Graphia: A platform for the graph-based visualisation and analysis of high dimensional data. PLOS Computational Biology **18**(7) (Jul 2022). <https://doi.org/10.1371/journal.pcbi.1010310>

[5] Gonzalez-Galofre, Z.N., Alcaide-Corral, C.J., Tavares, A.A.S.: Effects of administration route on uptake kinetics of ¹⁸F-sodium fluoride positron emission tomography in mice. Scientific Reports **11**(5512) (Mar 2021). <https://doi.org/10.1038/s41598-021-85073-0>

[6] Suchacki, K.J., Alcaide-Corral, C.J., Nimale, S., Macaskill, M.G., Stimson, R.H., Farquharson, C., Freeman, T.C., Tavares, A.A.S.: A systems-level analysis of total-body PET data reveals complex skeletal metabolism networks in vivo. Frontiers in Medicine **8** (Sep 2021). <https://doi.org/10.3389/fmed.2021.740615>

[7] Zoch, M.L., Abou, D.S., Clemens, T.L., Thorek, D.L., Riddle, R.C.: *In vivo* radiometric analysis of glucose uptake and distribution in mouse bone. Bone Research **4**:16004 (April 2016). <https://doi.org/10.1038/boneres.2016.4>

Towards automatic scoring of spinal X-ray for ankylosing spondylitis

Author

Yuanhan Mo – Big Data Institute, University of Oxford, Oxford, UK

Yao Chen – Novartis Pharmaceutical Company, East Hanover, NJ, USA

Aimee Readie – Novartis Pharmaceutical Company, East Hanover, NJ, USA

Gregory Ligozio – Novartis Pharmaceutical Company, East Hanover, NJ, USA

Thibaud Coroller – Novartis Pharmaceutical Company, East Hanover, NJ, USA

Bart–łomiej W. Papież – Big Data Institute, University of Oxford, Oxford, UK

Citation

Mo, Y., Chen, Y., Readie, A., Ligozio, G., Coroller, T., Papież, B.W. Towards automatic scoring of spinal X-ray for ankylosing spondylitis.

Abstract

Manually grading structural changes with the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS) on spinal X-ray imaging is costly and time-consuming due to bone shape complexity and image quality variations. In this study, we address this challenge by prototyping a 2-step auto-grading pipeline, called VertXGradeNet, to automatically predict mSASSS scores for the cervical and lumbar vertebral units (VUs) in X-ray spinal imaging. The VertXGradeNet utilizes VUs generated by our previously developed VU extraction pipeline (VertXNet) as input and predicts mSASSS based on those VUs. VertXGradeNet was evaluated on an in-house dataset of lateral cervical and lumbar X-ray images for axial spondylarthritis patients. Our results show that VertXGradeNet can predict the mSASSS score for each VU when the

data is limited in quantity and imbalanced. Overall, it can achieve a balanced accuracy of 0.56 and 0.51 for 4 different mSASSS scores (i.e., a score of 0, 1, 2, 3) on two test datasets. The accuracy of the presented method shows the potential to streamline the spinal radiograph readings and therefore reduce the cost of future clinical trials.

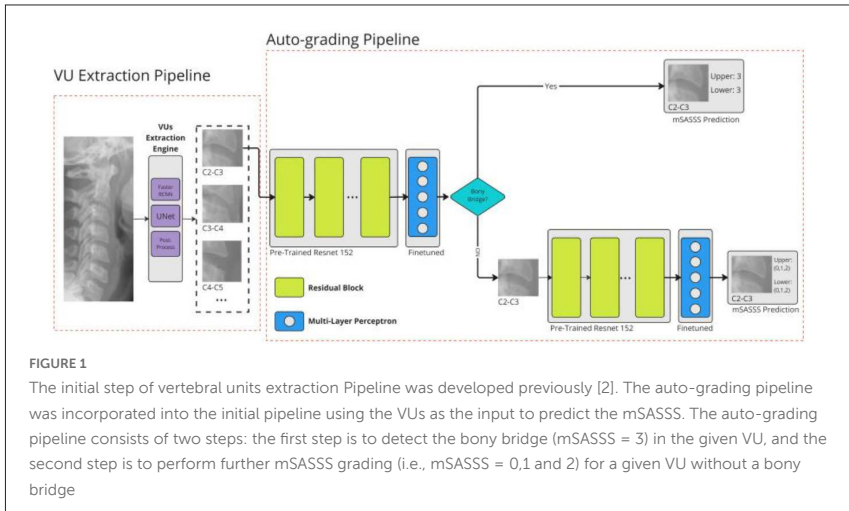
Introduction

X-ray imaging is one of the imaging modalities, utilized to monitor the structural progression of ankylosing spondylitis (AS), and this progression can be quantified by the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS) [6]. Applying the mSASSS to each VU in an X-ray is usually performed manually by expert readers (radiologists) due to its complexity, making it a costly and tedious process. Moreover, such a manual process can introduce inter- and intra-reader variability in the final dataset. Therefore, to address these problems, we aim to propose an automatic pipeline for mSASSS grading.

In this paper, we propose a 2-step auto-grading pipeline, VertXGradeNet, for estimating mSASSS scores from the given spinal X-ray images. The proposed auto-grading pipeline was built on top of the previously developed VUs extraction pipeline [2] and predicts mSASSS scores based on extracted VUs. Then the proposed pipeline was validated utilizing clinical trial data from radiographic and non-radiographic axial spondyloarthritis patients.

Method

Inspired by work conducted by Koo et al. [5], a two-step grading pipeline is proposed (Figure 1) which uses a deep neural network called ResNet [4] as the backbones. In the test phase, given a VU extracted by our in-house VU extraction pipeline, the first step of the proposed pipeline is to detect if a given VU has a bony bridge or not (i.e., mSASSS = 3 versus others). If the given VU does not have a bony bridge, then the VU will proceed to the second step for further mSASSS prediction (i.e., mSASSS = 0,1 and 2). In the second step, the pipeline will produce two predictions on both the upper and lower corners of the given VU (e.g. mSASSS = 0, 1, or 2) that is not classified



as a bony bridge. Finally, each given VU will have the estimated mSASSS given by the auto-grading pipeline.

The ResNet152 [4] is used twice in the two-step auto-grading pipeline. In the first step, the ResNet 152 takes a VU as input and performs a binary classification, namely if the VU does or does not have a bony bridge. In the second step, Another ResNet 152 takes the remaining VUs (mSASSS ≤ 2) as the inputs and predicts the rest of the mSASSS scores for both the upper and lower corners of a VU.

In the training stage, a pre-trained ResNet 152 (on ImageNet) is fine-tuned on the extracted VUs/mSASSS scores from the MEASURE1 dataset. The training process of the two stages of the auto-grading pipeline is not end-to-end. Therefore, each step is trained independently.

Experiments and Results

Data. The method was developed based on the anonymized datasets of secukinumab radiographic and non-radiographic axial spondyloarthritis clinical studies (MEASURE 1 [1] and PREVENT [3]). A total of 7239 and 9313 VUs were successfully extracted from the two studies.

Experiments. Extracted VUs were randomly split into 5 folds at the patient level for MEASURE 1 and performed 5-fold cross-validation. The performance of the model, trained on MEASURE 1 data only, was evaluated on the PREVENT dataset. The ground truth mSASSS scores for the training VUs were provided by the clinical trial. The detailed results are shown in Table 1 and 2.

TABLE 1: Results of 5-fold cross-validation for MEASURE 1 dataset

mSASSS	Precision	Recall	AUC(ROC)	F1-score
0	0.934(0.010)	0.918(0.007)	0.897(0.011)	0.926(0.010)
1	0.200(0.097)	0.240(0.103)	0.809(0.076)	0.218(0.096)
2	0.390(0.069)	0.300(0.020)	0.849(0.021)	0.332(0.026)
3	0.654(0.067)	0.800(0.023)	0.959(0.009)	0.718(0.034)
Micro average	0.544(0.033)	0.564(0.032)	0.898(0.011)	0.548(0.032)
Macro average	0.860(0.014)	0.856(0.014)	0.879(0.020)	0.856(0.014)

TABLE 2: Results for PREVENT dataset

mSASSS	Precision	Recall	AUC(ROC)	F1-score	Support
0	0.99	0.95	0.97	0.825	15201
1	0.01	0.12	0.02	0.759	25
2	0.15	0.23	0.18	0.857	244
3	0.14	0.73	0.23	0.958	64
Micro average	0.32	0.51	0.35	0.826	15534
Macro average	0.97	0.93	0.95	0.850	15534

Conclusion

We have prototyped a 2-step auto-grading pipeline for automatic mSASSS scoring. The current approach, which now can be considered as a benchmark, improves the grading performance compared to the preliminary results. However, limited training samples and class imbalance issues still limit the current performance of the auto-grading pipeline. Thus, further analysis is required to address the aforementioned problems.

References

- [1] Baeten, D., et al.: Secukinumab, an interleukin-17a inhibitor, in ankylosing spondylitis. *New England journal of medicine* **373**(26), 2534–2548 (2015)
- [2] Chen, Y., Mo, Y., Readie, A., Ligozio, G., Coroller, T., Papiez, B.W.: Vertxnet: Automatic segmentation and identification of lumbar and cervical vertebrae from spinal x-ray images (2022). <https://doi.org/10.48550/ARXIV.2207.05476>, <https://arxiv.org/abs/2207.05476>
- [3] Deodhar, A., Blanco, R., Dokoupilová, E., Hall, S., Kameda, H., Kivitz, A.J., Pod- dubnyy, D., van de Sande, M., Wiksten, A.S., Porter, B.O., et al.: Improvement of signs and symptoms of nonradiographic axial spondyloarthritis in patients treated with secukinumab: primary results of a randomized, placebo-controlled phase iii study. *Arthritis & Rheumatology* **73**(1), 110–120 (2021)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), <http://arxiv.org/abs/1512.03385>

[5] Koo, B.S., Lee, J.J., Jung, J.W., Kang, C.H., Joo, K.B., Kim, T.H., Lee, S.: A pilot study on deep learning-based grading of corners of vertebral bodies for assessment of radiographic progression in patients with ankylosing spondylitis. *Therapeutic Advances in Musculoskeletal Disease* **14**, 1759720X221114097 (2022)

[6] Van Der Heijde, D., Braun, J., Deodhar, A., Baraliakos, X., Landewé, R., Richards, H.B., Porter, B., Riedinger, A.: Modified stoke ankylosing spondylitis spinal score as an outcome measure to assess the impact of treatment on structural progression in ankylosing spondylitis. *Rheumatology* **58**(3), 388–400 (2019)

Scottish Medical Imaging (SMI) – providing safe and secure access to research-ready, population- scale health and imaging data

Author

Susan Krueger– School of Medicine, University of Dundee, Dundee DD2 4BF, UK

Jacqueline Caldwell– eDRIS, Public Health Scotland, Edinburgh, EH16 4UX, UK

Rob Wallace

Ruairidh MacLeod– EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Bianca Prodan– EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Andrew Brooks– EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Smarti Reel– School of Medicine, University of Dundee, Dundee DD2 4BF, UK

Laura Moran– EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Kara Moraw– EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Guneet Kaur,

James Sutherland– School of Medicine, University of Dundee, Dundee DD2 4BF,
UK

Emily Jefferson–School of Medicine, University of Dundee, Dundee DD2 4BF, UK

Citation

Krueger, S., Caldwell, J., Wallace, R., MacLeod, R., Prodan, B., Brooks, A., Reel, S.,

Moran, L., Moraw, K., Kaur, G., Sutherland, J., Jefferson, E.

Routinely collected or ‘real-world’ medical images linked with associated text-based health care records are extremely useful for healthcare research. In Scotland, all treatment is recorded under a patient’s Community Health Index (CHI) number [1], making it easy to link images or ‘scans’ with

clinical records spanning decades. This offers a valuable depth of data and longitudinal view of disease progression at a population level. However, using such data is challenging because it contains personally identifiable information (PII) that must be completely removed before it can be used for research. It is by nature large and unwieldy and requires specialist tools and skills to work with. Trusted Research Environments (TREs) such as Scottish Medical Imaging (SMI) provide the capabilities and controls needed to unlock the value of such data whilst ensuring patient confidentiality [2].

Generally, the primary reason for medical image capture is clinical care, not research. The data is organised by 'person' not by 'disease' or other research-relevant category. It's not clean, and the purpose for taking an image is not clear, this means a lot of effort goes into making them research-ready. The SMI platform [3] was developed under the Interdisciplinary Collaboration for efficient and effective Use of clinical images in big data healthcare REsearch programme (PICTURES – a 5-year Medical Research Council funded Grant) [4] in response to three main research questions:

1. How to build research relevant cohorts from messy, unstructured, identifiable data?
2. How to handle big data in a scalable manner?
3. How to protect patient confidentiality?

We share our solutions to some of the major research challenges faced in establishing SMI, namely:

1. Cataloguing – the descriptive metadata that comes with each image, series or study is often inconsistent, incomplete or incorrect, making it difficult to reliably index and catalogue image data for cohort building.
2. Natural Language Processing – useful image descriptors can be found in associated Radiology Reports, specifically free-text notes describing expert observations and diagnosis.

3. Pixel anonymisation – many scanning devices or secondary processing steps result in Personally Identifiable Information (PII) being burnt-in to the images themselves, requiring tools and methods for identifying and removing PII from pixel data at-scale.
4. Image classification from pixel data – machine learning models have been trained to derive labels like MRI sequence type from features extracted from the scans themselves.

A large amount of metadata is generated for each image, thousands of tags depending on the image type. A Metadata Catalogue provides visibility over the SMI data, it reflects the state of data at different stages of the SMI service pipeline, providing statistics of quality, frequency, value distribution and a single source of this information. Searching for images of a particular body part like 'chest' sounds deceptively simple, however the BodyPartExamined field is often missing or unreliable – labelling things correctly is not always first priority in clinical care. Working with a group of clinical experts who generously give their time and expertise to the PICTURES programme, we automated large-scale body-part mapping from a manual activity covering around 52% of CT scans, to a term-dictionary approach delivering highly accurate results with between 81.03% and 99.99% coverage across 11 scan types or 'modalities'.

Where labels cannot be derived from metadata attached to the images, for instance whether a chest scan shows indications of lung nodules, these can often be picked up from the Radiology Report (aka Structured Report). Semantic Search System for Electronic Health Records (SemEHR) [5] is built upon off-the-shelf toolkit Bio-Yodie, and enterprise search system CogStack. This Natural Language Processing (NLP) tool was trained on SMI data to allow searching of free-text for specific terms and biomedical concepts from the Unified Medical Language System (UMLS) Metathesaurus [6]. The beauty of SemEHR is that it doesn't just do pattern matching, although it can. There are different ways of describing lung nodule for instance and they have been mapped to codes and 'concepts' in the UMLS Metathesaurus, allowing us to

search by concept rather than term, and use codes from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) ontology for instance, for more powerful search and discovery of concepts within Radiology Reports.

Once a study cohort is confirmed, relevant linked images can be placed within a secure research environment for access by the researcher. Across the various scan types or 'modalities' of medical imaging, many scanning devices or secondary processing of images results in personally identifiable information (PII) being burnt-in to the images themselves, requiring tools and methods for identifying and removing PII from pixel data at-scale. All routinely collected data must be de-identified before it can be used for research. A huge amount of work has gone into developing and validating the tools for de-identification of pixel data at scale. Modern Optical Character Recognition (OCR) tooling has been a major factor in our success. Just a few years ago this would have been much more difficult if not impossible, due the amount of image pre-processing required for OCR to work reliably. But now we find they handle grey scale very well. One of the main challenges we faced was over-redaction of images due to things like medical devices and teeth being picked up by the machine as text. Through a systematic process of trial and error these challenges have been largely overcome and we continue to apply and refine the pixel de-identification tooling across modalities until all are safe enough for research.

Where labels cannot be derived from scan metadata or Radiology Reports, we seek to derive them from the images themselves. Our model architecture consists of two parts: an autoencoder used to derive numerical features from the pixel data, and a classifier that is trained on those features. Both networks are trained separately. This decoupling means that the extracted features do not depend on the classification and can in fact be reused for other analyses.

The SMI Service is now live and able to support researchers and industry partners to train new models on routinely collected imaging data linked to healthcare records from the whole Scottish Population.

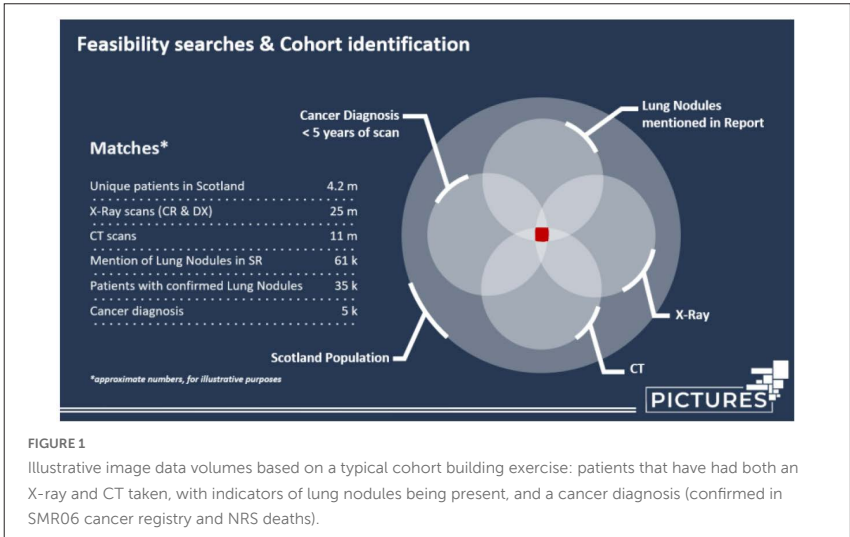


Figure 1 shows illustrative image data volumes based on a typical cohort building exercise: patients that have had both an X-ray and CT taken, with indicators of lung nodules being present, and a cancer diagnosis (confirmed in SMR06 cancer registry and NRS deaths).

References

[1] SMR Datasets | Patient Identification and Demographic Information | Community Health Index (CHI) Number | ISD Scotland | Data Dictionary, <https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/Patient-Identification-and-Demographic-Information/Community-Health-Index-Number/>, last accessed 2023/06/24.

[2] ISD Services | Electronic Data Research and Innovation Service (eDRIS) | ISD Scotland, <https://www.isdscotland.org/Products-and-Services/eDRIS/>, last accessed 2022/11/09.

[3] Nind, T., Sutherland, J., McAllister, G., Hardy, D., Hume, A., MacLeod, R., Caldwell, J., Krueger, S., Tramma, L., Teviotdale, R., Abdelatif, M., Gillen, K., Ward, J., Scobbie, D., Baillie, I., Brooks, A., Prodan, B., Kerr, W., Sloan-Murphy, D., Herrera, J.F.R., McManus, D., Morris, C., Sinclair, C., Baxter, R., Parsons, M., Morris, A., Jefferson, E.: An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population. *GigaScience*. 9, giaa095 (2020). <https://doi.org/10.1093/gigascience/giaa095>.

[4] PICTURES: Supporting the use of data for health care research - Image on a Mission, <https://www.imageonamission.ac.uk/>, last accessed 2022/11/09.

[5] Wu, H., Toti, G., Morley, K.I., Ibrahim, Z.M., Folarin, A., Jackson, R., Kartoglu, I., Agrawal, A., Stringer, C., Gale, D., Gorrell, G., Roberts, A., Broadbent, M., Stewart, R., Dobson, R.J.: SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc*. 25, 530–537 (2018). <https://doi.org/10.1093/jamia/ocx160>.

[6] Metathesaurus, https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html, last accessed 2023/06/24.

Text-based medical image classification by body part

Author

Bianca Prodan – EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Laura Moran – EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK

Susan Krueger – School of Medicine, University of Dundee, Dundee DD2 4BF, UK

Emily Jefferson – School of Medicine, University of Dundee, Dundee DD2 4BF, UK

Citation

Prodan, B., Moran, L., Krueger, S., Jefferson, E. Text-based Medical image classification by body part.

Medical research questions are usually formed around specific conditions, treatments, procedures, or demographics. Building cohorts around these questions is difficult, requiring a level of medical expertise, knowledge of the data and multiple build iterations [1]. To support the cohort building process, we aim to apply an automatic solution to the classification of medical images by body part. The result of this classification is labelled data which enables filtering during cohort building.

Medical images are commonly stored as Digital Imaging and Communications in Medicine (DICOM) objects, containing pixel data and related metadata. DICOM metadata can describe various aspects of the image and the conditions under which it was obtained in attributes referred to as DICOM tags. DICOM tags contain patient information, scan settings, machine information, and medical notes. Most classification solutions look at pixel data due to its high reliability in comparison with attached

metadata, ignoring the metadata due its sparseness and messiness. In comparison, text-based classification would be faster, more scalable, and more computationally efficient. Rather than seeking to replace pixel-based classification, we aim to use both the text and the image as complementary techniques to provide a confidence indicator.

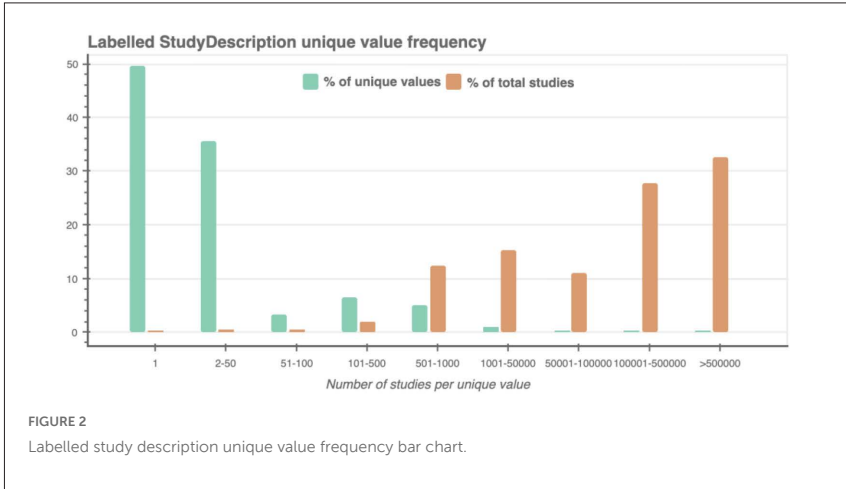
To decide whether there is value in metadata text classification, a radiologist examined a collection of common DICOM tags and measured their reliability for body part classification by comparing them with their respective pixel data. An analysis of the tag frequency and completeness paired with validation against the pixel data revealed that the “StudyDescription” tag is the primary source of body part descriptors, with “SeriesDescription” and “BodyPartExamined” as secondary sources.

The Scottish Medical Imaging (SMI) dataset contains ~25 million studies across 11 modalities. To ensure efficient labelling, the metadata was first reduced to a list of unique values by applying data cleaning and grouping, with a count of studies, series and images containing each particular value. For each of the selected tags, a list of unique values was extracted from across the 11 modalities. These were run through the data cleaning process consisting of removal of errant spaces, special characters, and capitalisation. This resulted in a reduction of ~25 million individual values to ~33 thousand unique values for the main tag “StudyDescription”, a reduction of ~99%.

By analysing the list of unique values with a focus on the “StudyDescription” tag, we created a dictionary of common medical terms mapped to one of 9



FIGURE 1
Metadata labelling process.



body part categories: *head, neck, chest, abdomen, pelvis, upper limb, lower limb, spine, and whole body*. This dictionary tries to encapsulate synonyms, different spellings, incorrect spellings, different languages, abbreviations, irregular plurals, acronyms, and medical codes. By reducing medical terms to their lemma (e.g., thoracic, thorax → thora), we could address multiple of these issues with a smaller number of mappings, resulting in a dictionary of 223 terms. Applying this dictionary to the unique values achieved a coverage of ~93%.

As reflected in Figure 1, manual validation of the labels against the unique values as well as pixel data samples is necessary to determine the accuracy of their application. To make this more feasible, we analysed how many of the unique values represent the majority of studies. As shown in Figure 2, by focusing on the top ~10% of the list of unique values we can cover ~95% of the total studies, with 50% of unique values representing single use cases.

Manual verification of this coverage showed ~92% accuracy. Several issues were identified in ~4% of the unique values, including false positives, harsh abbreviations, incorrect spellings, double meanings, negations, and body part ranges, however these represented a small amount of the data.

Labelling with a focus on the “StudyDescription” tag acted as a proof of concept that we are currently building on by incorporating other tags and expanding the medical term dictionary with DICOM’s anatomic region codes mapped to “BodyPartExamined” defined terms [2]. So far, we have grouped the three selected tags: “StudyDescription”, “SeriesDescription”, and “BodyPartExamined”, labelling them separately as well as together, introducing a “confidence score” based on how many of the tags are in agreement over the applied label(s). This confidence score provides a base for introducing other sources such as labelling from pixel-based classification solutions and would enable cohort builders to specify a minimum level of confidence when selecting data. Once we have a stable and validated medical term dictionary, as well as labelled and unlabelled datasets, we plan on introducing Machine Learning (ML) classification methods and exploring Natural Language Processing (NLP) options for solving identified issues.

While this study is in early stages, it so far shows promising results. If we can show that there is consistent value in DICOM metadata, text-based medical image classification could prove to be a valuable contribution to medical research cohort building. Long-term viability of this solution depends on creating a scalable and automated training process for the text-based classifier. To this end, domain-specific NLP capable of dealing with medical jargon, in limited context and fragmented text sources, will need to be developed in concert with medical experts. This must be supported by a robust verification and validation process, making it a prime candidate for exploring Explainable Artificial Intelligence (XAI). As the format of the data is a combination of pixel, text, and potentially video and audio, this work could also benefit from the use of Multimodal ML models.

References

- [1] E. Jefferson *et al.*, 'An architecture for building cohorts of images from real-world clinical data from the whole Scottish population supporting research and AI development.', *Int J Popul Data Sci*, vol. 7, no. 3, p. 1916, doi: 10.23889/ijpds.v7i3.1916.
- [2] dicom.nema.org. (n.d.). 'L Correspondence of Anatomic Region Codes and Body Part Examined Defined Terms'. [online] Available at: https://dicom.nema.org/medical/dicom/current/output/chtml/part16/chapter_L.html#chapter_L [Accessed 23 Jun. 2023].

Natural language process of radiology reports

Author

Andrew Brooks – EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK
Honghan Wu – Institute of Health Informatics, University College London, London, NW1 2DA, UK

Citation

Brooks, A., Wu, H. Natural language process of radiology reports.

The Scottish Medical Imaging (SMI) archive holds a copy of Scotland's radiology images for use by researchers. It includes radiology reports in text format for CT, MRI, Ultrasound, and other modalities. The PICTURES project has been funded to do research and development of tools and techniques for making the archive more easily accessible and usable.

The text can be useful in several ways:

- As a resource for researchers to consult when interpreting the images with which they are associated,
- As a standalone resource for researchers using their own Natural Language Processing (NLP) tools,
- As a way to find images about a specific topic (disease, drug or medical intervention), i.e. a searchable database for building cohorts,
- As a way to create new metadata

In all use cases, before being allowed to use or release the text, all documents must be de-identified, so they must be examined and have Personally Identifiable Information (PII) removed. This includes names, addresses, postcodes, dates of birth, telephone numbers, GMC registration numbers, and so on. The natural language processing (NLP) work package has been doing research into methods for removing all PII from the free-text reports. The process must be robust enough to give confidence that reports can be delivered to researchers without the risk of re-identifying patients.

A second aim is to catalogue the text to make it easier to find reports about specific medical conditions, treatments or drugs. This labels the text using concepts defined in the UMLS metathesaurus, a comprehensive medical ontology, and it has the ability to spot different phrases with the same meaning. The medical reports are all associated with radiology images so this approach to building cohorts can be used to search the archive for relevant images as well as reports.

This poster describes the work to build an acceptable NLP solution for detecting PII, the validation process, the design of a process for extracting medical concepts from reports and a database for cataloguing them, and a web service with cohort building tool plugin for querying the catalogue.

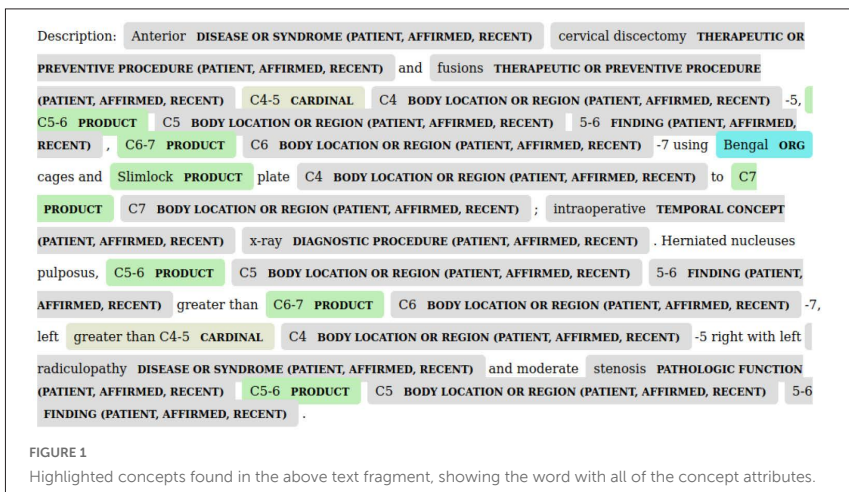
The first problem to solve was the file format. Reports are stored in DICOM which is designed for storing images. The text is stored as an item of metadata in a file which has no image pixels. The metadata includes specific fields for Patient Name, Patient Date of Birth, and others which obviously contain PII, plus a set of fields relating to the study and associated images, which may or may not contain PII. A tool called the Clinical Trial Processor (CTP) is used to remove the fields containing PII, which will include removal of the actual clinical report text. In parallel our own de-identification tool parses the DICOM file to extract the text, de-identifies it, and places this text into the output from CTP.

The second problem to solve was the text structure. The reports use a sub-type of DICOM called Structured Reports, but the text itself is not structured

into useful sections (about the patient, the condition, the treatment, the outcome) so it needs to be decoded before it can be used.

After the text has been extracted it is de-identified using a mixture of techniques including dictionaries, rules-based and contextual information. NLP may be used but care must be taken not to remove names of body parts or diseases such as Monro, Parkinson, etc.

The free-text reports will contain mentions of body parts, diseases, conditions, drugs, treatments, and so on. It is essential to identify these concepts so that we can search for them and also use them for medical analytics. We use NLP to parse the de-identified text and extract concepts from phrases, including context such as who experienced it, whether it is in the past or ongoing, and whether it is negated (e.g. "scan shows patient does not have lesions"). The NLP handles concepts which can be described using different words or phrases, and gives them a unique concept identifier.



The database of concepts has been built using the JSON features of PostgreSQL, with indexes on the array of concepts in each document and on the free text itself. Figure 1 shows all of the concepts found in the above text fragment, showing the word with all of the concept attributes:

Validation has been performed in both aspects of the project:

1. To ensure that the de-identification removes as much PII as possible, and does not remove non-PII, thus reducing the risk to PHS of releasing documents to researchers that could identify any person
2. To ensure that the medical ontology classifies concepts in a way which is useful for searching through the archive and building cohorts.

Two tools have been used for this. One is a very specific tool for checking reports that gives a very easy Yes/No facility to classify words as PII. The other tool is a customised version of eHost which allows multiple annotators to work on a set of reports, checking for PII, marking elements which were missed or were incorrectly flagged. It has also been used on the concept annotations, for example to classify body parts. The system has the ability to learn from researcher corrections to the annotations, to make future search queries return more appropriate results for the study.

In conclusion, we have developed a robust method for de-identifying the free-text clinical reports which accompany radiology images. It achieves accuracy high enough to be used by PHS for supplying reports to researchers. We have developed methods and deployed a pipeline for using NLP to extract meaningful information from free-text reports which can subsequently be used for search queries, improved cohort creation and for research analytics. We have developed and deployed GUI tools streamlining the validation workflow and the annotation workflow.

Medical image anonymisation

Author

Andrew Brooks – EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK
Guneet Kaur – School of Medicine, University of Dundee, Dundee DD2 4BF, UK

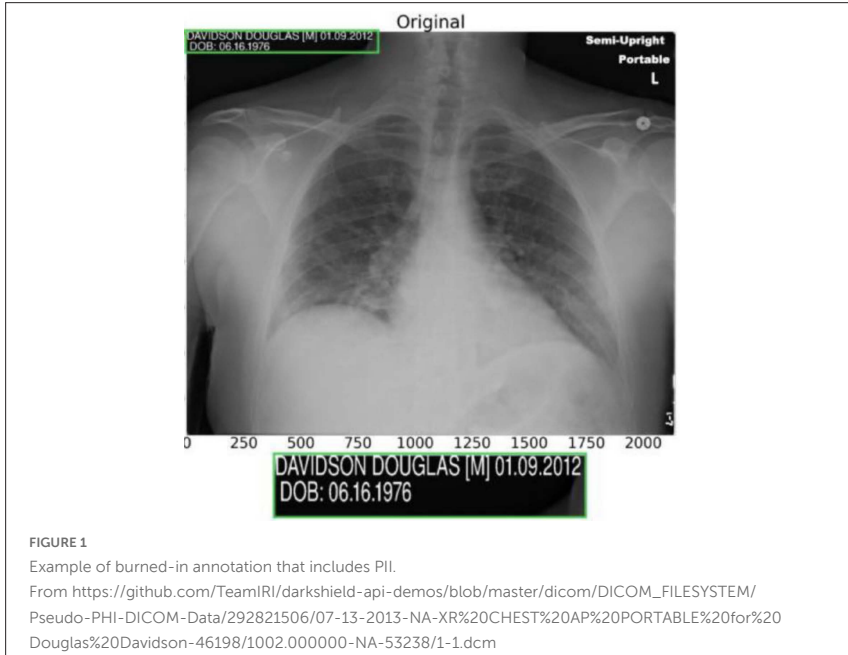
Citation

Brooks, A., Kaur, G. Medical image anonymisation.

The Scottish Medical Imaging (SMI) service, run by Public Health Scotland (PHS) and EPCC (University of Edinburgh), maintains an archive of radiology images for use by researchers. These images are from all NHS Health Boards across Scotland, and include CT, MRI, Ultrasound, and other modalities. The PICTURES project has been funded to do research and development of tools and techniques for making the archive more easily accessible and usable.

Before PHS can release images to researchers all images must be de-identified, so they must be examined and have Personally Identifiable Information (PII) removed. The main problem is that the patient's name, and other PII, may be written on the image itself, burned into the pixels. PII also resides in the image metadata, and removing this is a separate task not described here. How can we safely remove all references to people and other sensitive identifiers from medical images to make them safe for use in research?

The image anonymisation work package has been doing research into methods for removing all Personally Identifiable Information (PII) which has been 'burned in' to the image pixels. In contrast to most anonymisation tools, which only remove PII from the image metadata, the aim is to detect and remove names, dates, and other PII which appears written into the image pixels. An example of such a burned-in annotation is shown in figure 1.



The images in the archive arrive in their raw form, in DICOM format but with no post-processing to clean them up. The first task was to determine the scale of the problem: which images might contain text, where the text is written, and what the text contains. In an archive containing billions of images, this is a Big Data task, so it was started by looking at the metadata, to try and find patterns between machines (by Manufacturer and Model), by Image Type tags, or by metadata such as the BurnedInAnnotation tag. The sheer variety of text in images rendered this impractical so an OCR method was tried instead.

This poster describes the work to find an acceptable OCR solution for detecting text, the tools which were created to assist in the validation process, and the end result being a pipeline for delivering anonymised images to researchers. All of the code has been made open source.

The first questions were:

- How many images do we have?
- How prevalent is this problem?
- Which machines write text onto images?
- Where is the text on the images?
- Is there any metadata indicating the presence of text, or the location on the image?

The archive holds over 2½ billion image files in DICOM format. It is split into several modalities; X-Ray, Ultrasound, CT, MRI, etc. Many files have multiple image frames and can also contain overlay images. An impossible number to manually check.

There is a metadata field called “BurnedInAnnotation” defined in the DICOM standard, taking values YES or NO. A scan for this tag proved that it’s not always present, and it’s often used wrongly, so not helpful. Ultrasound (only) has metadata indicating clinical regions, which is useful, but not a complete guarantee. Some machines only write text in the corners (like the example above), but other machines can place text literally anywhere on the image.

Dividing the archive by ImageType or SOPClass metadata might reduce the scale of the problem. Dividing the archive by specific Manufacturer and Model Number would allow samples from each combination to be checked manually. However, there’s still tens of thousands of combinations.

Overlays are the recommended way to add text, so removing overlays should solve the problem, but a) text is found on main images too, and b) the overlays may contain useful clinical information.

Having analysed representative samples from CT and MRI images it was determined that those of type "ORIGINAL/PRIMARY" never contain burned in text, if overlays are excluded. This releases billions of images to researchers simply by removing overlays.

Unfortunately, just removing overlays is not good enough for other modalities.

Scanning using OCR and rejecting all images which contain text would prevent most images of some modalities from being released.

The only solution is to redact those parts of the images known to contain text. Some work has already been done to classify these images and produce a set of rules. Unfortunately these rules are not comprehensive enough for our archive.

We have developed a unique GUI tool which allows every single image frame and overlay frame to be visually checked for PII. Rectangles can be drawn over regions to be redacted. Common patterns can be identified and redaction rules created for specific models of scanner.

Several OCR algorithms have been tested, adapted and validated. A software pipeline has been developed to run OCR on all frames and redact any text found.

OCR results have been validated and rules created for images from machines which are known to be troublesome.

Natural Language Processing (NLP) has been tested to see if Named Entity Recognition (NER) can spot PII. Unfortunately, the lack of context around

names means it doesn't work well enough, so all text found by OCR would need to be redacted, not just names.

An allow-list has been crafted which can prevent redaction of specific radiological phrases and patient orientations, to preserve information useful to researchers.

Machine-learning algorithms have been developed to spot non-clinical images; these are typically text-heavy forms which are not to be released.

A combination of all of the above techniques is used in the de-identification pipeline.

In conclusion, we have done a comprehensive analysis of a national archive of clinical radiology images for metadata and patterns which indicate how they can be released safely for research. Software has been developed which can find text burned into DICOM images and redact it. The results have been validated and possible problems identified. The residual risk has been determined to be sufficiently low that this mechanism can be used to release the X-Ray and Ultrasound modalities to researchers. The same techniques can be applied to other modalities, once they too have been visually validated to check there are no surprises.

Automated segmentation of cerebral small vessel disease from field-cycling MRI

Author

Nicholas Senn – Aberdeen Biomedical Imaging Centre, University of Aberdeen, AB24 3FX, UK

Vasiliki Mallikourti – Aberdeen Biomedical Imaging Centre, University of Aberdeen, AB24 3FX, UK

P. James Ross – Aberdeen Biomedical Imaging Centre, University of Aberdeen, AB24 3FX, UK

Lionel M. Broche – Aberdeen Biomedical Imaging Centre, University of Aberdeen, AB24 3FX, UK

Gordon D Waiter – Aberdeen Biomedical Imaging Centre, University of Aberdeen, AB24 3FX, UK

Mary-Joan MacLeod – Institute of Medical Sciences, University of Aberdeen, AB24 3FX, UK

Citation

Senn, N., Mallikourti, V., Ross, P.J., Broche, L.M., Waiter, G.D., MacLeod, M. Automated segmentation of cerebral small vessel disease from field-cycling MRI.

Abstract

New clinically viable approaches are needed to realise the potential of non-invasive imaging to monitor changes to the severity of cerebral small vessel disease (SVD) in patients. Field-cycling imaging (FCI) is an emerging whole-body MRI technology that provides unique access to underlying tissue features by varying the magnetic field during acquisition, at strengths up to 10,000 times lower than conventional fixed-field MRI. The low-field nature of FCI further means that it has the potential to be developed towards a variety

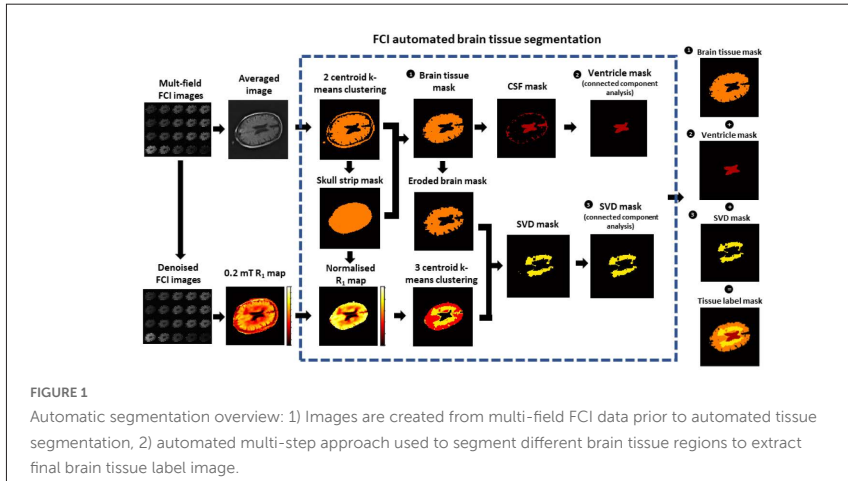
of accessible and impactful clinical applications. The aim of this preliminary work was to investigate the feasibility of FCI to quantify SVD severity when combined with a fully automated segmentation algorithm. In reference to segmented tissue labels obtained from 3T MRI, FCI can accurately segment regions of brain matter (mean Dice coefficient of 0.89), ventricle (0.91), and SVD (0.52). The variation between SVD regions may partly be explained by differences in sensitivity of FCI to underlying pathophysiological processes involved with SVD. This preliminary work is a crucial step towards assessing the clinical feasibility of FCI for SVD clinical applications.

Introduction

Cerebral small vessel disease (SVD) is associated with increased stroke risk and contributes to cognitive decline [1]. New clinically viable non-invasive imaging approaches are needed to monitor changes to SVD severity in patients, unravel the pathophysiological processes involved SVD and inform development of new treatments [2]. Field-cycling magnetic resonance imaging (MRI) is an emerging technology pioneered at the University of Aberdeen [3]. In contrast to conventional fixed field MRI scanners, typically operating at 1.5 and 3T, field-cycling imaging (FCI) provides a unique opportunity to acquire images over multiple magnetic field strengths to provide endogenous contrast R_1 maps at magnetic fields between 0.2 and 200 mT. We hypothesise that FCI combined with an automated segmentation approach provides a clinically feasible approach to assess SVD severity.

Proposed Segmentation Method

An automated segmentation approach has been developed to segment regions of white matter changes associated with SVD from surrounding white matter using R_1 images generated at 0.2 mT from FCI (see Fig. 1). Tissue label masks are created for brain tissue, ventricle, and small vessel disease. The automated approach was written in MATLAB (MathWorks, USA). A constrained k-means clustering based approach was implemented to utilise the inherent contrast between hypointense regions of R_1 corresponding to SVD white matter changes and hyperintense regions of R_1 corresponding to surrounding white matter (see Fig. 1: R_1 map). A multi-step process is used



to generate additional tissue masks which are then used to differentiate SVD from cerebrospinal fluid (CSF) regions by accounting for the overlapping isointense R_1 values.

Experiments and Results

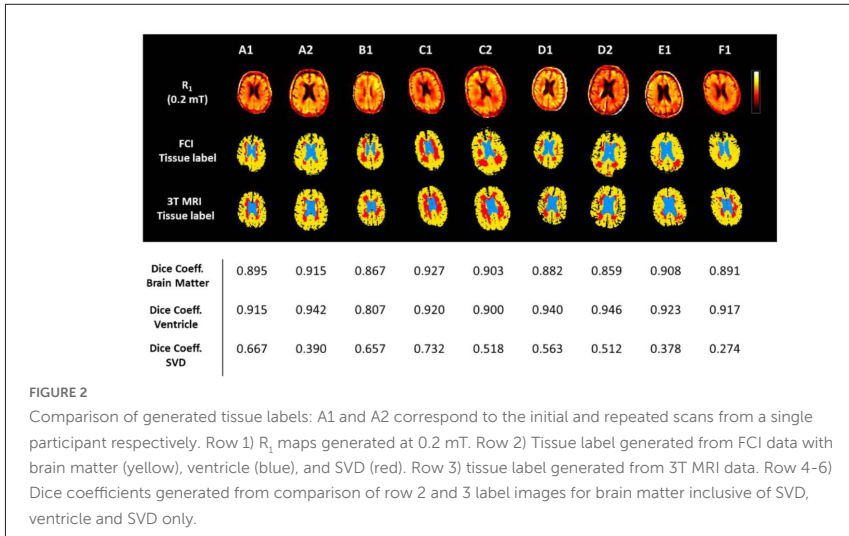
Methods. The study was approved by the North of Scotland Research Ethics Committee (21/NS/0128). A total of 9 data sets were included from the first patients recruited to the study, who attended an initial 3T MRI (Philips 3T dStream) and FCI scan ($N = 6$) and repeated scans after 30 days ($N = 3$). FCI images were acquired across four evolution fields of 0.2, 2, 20 and 200 mT, 5 logarithmically spaced evolution times, TE of 16 ms, matrix size of 90 x 90, in plane resolution of 3.1 mm, and slice thickness of 10 mm. Prior to generation of R_1 maps at each evolution field, FCI images were denoised using a pretrained denoising convolutional neural network contained within MATLAB. A separate brain tissue label was generated from 3T MRI data using an existing automated approach to segment regions of white matter hyperintensity [4], and co-registered to FCI images using a landmark-based approach.

Results

Mean and range of Dice coefficients were obtained for brain matter inclusive of SVD (**0.89**, 0.86 – 0.93), ventricle (**0.91**, 0.81 – 0.95), and SVD only (**0.52**, 0.27 – 0.73), (see Fig. 2). Visual inspection of the SVD tissue label shows regions of both false positive and false negative disagreement. Lower values of Dice coefficient obtained from SVD labels appear to correspond to lower volumes of SVD segmented from FCI compared to that obtained from 3T MRI, (see. Fig. 2: A2 and F1). A significant Pearson correlation was obtained between SVD fractions ($R = 0.861$, $P = 0.003$), calculated as the ratio between number of SVD and brain matter voxels.

Discussion

FCI combined with an automated segmentation approach has the potential to inform radiological assessment of SVD severity and monitor disease progression. The preliminary results obtained from this study demonstrate



the feasibility of FCI to differentiate SVD changes to white matter and inform automated segmentation of these regions. Differences between tissue labels generated from FCI and 3T MRI may partly be underpinned by different sensitivity of imaging approaches to underlying pathophysiological processes involved with SVD changes to white matter [2]. Future work is required to interrogate the sensitivity of FCI to underlying SVD processes and develop further the automated segmentation method presented here.

Conclusion

The preliminary results demonstrate the feasibility of FCI to inform automated segmentation of SVD brain changes and quantification of disease severity.

References

- [1] Østergaard, L., et al.: Cerebral small vessel disease: Capillary pathways to stroke and cognitive decline. *J Cereb Blood Flow Metab* **36**(2), 302-25 (2016)
- [2] Stringer, M.S., et al: A Review of Translational Magnetic Resonance Imaging in Human and Rodent Experimental Models of Small Vessel Disease. *Transl. Stroke Res.* **12**, 15–30 (2021)
- [3] Broche, L.M., et al.: A whole-body Fast Field-Cycling scanner for clinical molecular imaging studies. *Sci Rep* 9,10402 (2019)
- [4] Waymont, J.M.J., et al: Validation and comparison of two automated methods for quantifying brain white matter hyperintensities of presumed vascular origin. *J Int Med Res.* **48**(2), 300060519880053. (2020)

^{68}Ga FAPI imaging in cancer diagnosis: A promising approach for targeted molecular imaging

Author

Sidharth Vinod – Degree of Master of Science in Medical Physics Student, University of Aberdeen

Citation

Vinod, S. ^{68}Ga FAPI imaging in cancer diagnosis: A promising approach for targeted molecular imaging.

The diagnosis and treatment of cancer have been greatly improved by the use of molecular imaging techniques, resulting in substantial progress in the field. The emergence of ^{68}Ga FAPI, a radiotracer designed to specifically target fibroblast activation protein (FAP), has shown great potential in preclinical studies and initial clinical trials. This study aims to thoroughly evaluate the efficacy, feasibility, and accuracy of ^{68}Ga FAPI imaging for diagnosing cancer in individuals.

The process of ^{68}Ga FAPI imaging utilizes a molecular imaging technique that involves the introduction of a radiotracer called ^{68}Ga FAPI. This radiotracer specifically binds to cells expressing fibroblast activation protein (FAP) within tumors or fibrotic tissues. Subsequently, positron emission tomography (PET) imaging is conducted to capture the emitted gamma rays, generating detailed images that provide valuable information about the location, extent, and metabolic activity of the disease. Through the visualization and evaluation of these factors, ^{68}Ga FAPI imaging provides accurate insights into the presence and characteristics of tumors and fibrotic

diseases. This, in turn, aids in the diagnosis, planning of treatment, and monitoring of such conditions.

FAPI 4 and FAPI 2 are variants of FAPI compounds for cancer imaging using Gallium-68. Both FAPI 4 and FAPI 2 bind to and inhibit fibroblast activation protein (FAP) found in tumor stroma. These compounds show promise in molecular imaging, particularly PET/CT. To create imaging tracers, FAPI compounds are combined with radioactive Gallium-68 (^{68}Ga), resulting in [^{68}Ga] FAPI 4 and [^{68}Ga] FAPI 2. Gallium-68 is a short-lived positron-emitting radionuclide obtained from a $^{68}\text{Ge}/^{68}\text{Ga}$ generator system, where ^{68}Ge decays to produce ^{68}Ga .

The preparation of [^{68}Ga] FAPI 4 and [^{68}Ga] FAPI 2 radiotracers involves the labeling of FAPI 4 and FAPI 2 with ^{68}Ga using a chelation process. Chelation refers to the creation of a stable coordination complex between the FAPI molecule and the ^{68}Ga isotope. This ensures that the ^{68}Ga remains securely attached to the FAPI molecule until it is administered to the patient. Upon administration, the [^{68}Ga] FAPI 4 or [^{68}Ga] FAPI 2 radiotracer circulates throughout the patient's body. It selectively binds to FAP present in the tumor stroma, allowing for the visualization of FAP expression during PET imaging.

PET/CT imaging involves the detection of positrons emitted by decaying ^{68}Ga atoms using a PET scanner. These detected signals are reconstructed to generate three-dimensional images that accurately represent the spatial distribution and concentration of the [^{68}Ga] FAPI 4 or [^{68}Ga] FAPI 2 radiotracer. This imaging technique plays a crucial role in providing valuable insights into FAP expression and the precise localization of tumor lesions. It greatly aids in the diagnosis, staging, and assessment of treatment response in cancer patients.

A meticulous and comprehensive retrospective analysis was conducted on a well-defined group of 50 patients affected by various types of cancer. These patients underwent [^{68}Ga] FAPI positron emission tomography/computed tomography (PET/CT) as part of their diagnostic and treatment

process. The primary objective of this analysis was to quantitatively evaluate the uptake of the radiotracer within the tumor lesions using standardized uptake values (SUVs). The results obtained from this quantitative assessment were then carefully compared with histopathological findings, conventional imaging techniques, and subsequent clinical outcomes to gain a more comprehensive understanding of the overall implications.

[⁶⁸Ga] FAPI imaging uses a radiotracer containing a FAPI (fibroblast activation protein inhibitor) compound. FAPIs are small molecules designed to bind and inhibit fibroblast activation protein (FAP) found in excessive amounts in tumor stroma of various cancers. The FAPI compound is labeled with Gallium-68 to create the [⁶⁸Ga] FAPI radiotracer. ⁶⁸Ga is a positron-emitting radionuclide with a 68-minute half-life, obtained from a ⁶⁸Ge/⁶⁸Ga generator system. The preparation of [⁶⁸Ga] FAPI radiotracer involves chelation, forming a stable coordination complex between the FAPI molecule and ⁶⁸Ga. This ensures secure attachment of ⁶⁸Ga to the FAPI molecule until administration to the patient.

After administration, the [⁶⁸Ga] FAPI radiotracer undergoes circulation throughout the patient's body. Within the tumor stroma, which encompasses cancer-associated fibroblasts, there is a high expression of FAP. The FAPI molecule exhibits selective binding to FAP present in the tumor stroma, enabling the visualization of FAP expression during PET imaging.

In the process of positron emission tomography/computed tomography (PET/CT) imaging, the PET scanner detects the positrons emitted by decaying ⁶⁸Ga atoms. These detected signals are reconstructed to create three-dimensional images that accurately depict the spatial distribution and concentration of the [⁶⁸Ga] FAPI radiotracer. This imaging technique plays a crucial role in providing valuable information about the expression and localization of FAP in tumor lesions. Consequently, it greatly assists in cancer diagnosis, staging, and the assessment of treatment response. Consequently, [⁶⁸Ga] FAPI imaging integrates the selective binding characteristics of the

FAPI compound with the radioactive attributes of the ^{68}Ga isotope. This combination allows for the non-invasive visualization and quantification of FAP expression in cancerous tissues.

The application of [^{68}Ga] FAPI PET/CT has demonstrated remarkable effectiveness in the detection of primary tumors and metastatic lesions across various cancer types, exhibiting high levels of sensitivity and specificity. The average maximum standardized uptake value (SUV_{max}) of FAPI-avid lesions showed a significant increase compared to the surrounding normal tissues ($p < 0.001$). Moreover, [^{68}Ga] FAPI imaging revealed previously unseen lesions in a notable proportion of cases (28%), which were not detected by traditional imaging methods. As a result, this prompted adjustments in patient management and treatment strategies. The relationship between the uptake of [^{68}Ga] FAPI and the aggressiveness of tumors was further validated through thorough histopathological analysis.

The results of this study provide significant insights into the immense potential of [^{68}Ga] FAPI imaging as a highly promising and effective tool for cancer diagnosis and staging. The substantial sensitivity and specificity observed within this cohort unequivocally establish its clinical value in facilitating well-informed treatment decisions and accurately assessing treatment response. Furthermore, the remarkable ability of [^{68}Ga] FAPI PET/CT to detect additional lesions missed by conventional imaging techniques serves as compelling evidence of its capacity to improve overall patient outcomes. To further validate these findings and explore the extensive clinical implications of [^{68}Ga] FAPI imaging, additional research efforts and larger-scale studies are undoubtedly warranted.

References

- [1] ^{68}Ga -FAPI PET/CT: Tracer Uptake in 28 Different Kinds of Cancer
- [2] ^{68}Ga -FAPI-PET/CT in patients with various gynecological malignancies

[3] FAPI-PET/CT in Cancer Imaging: A Potential Novel Molecule of the Century

[4] ^{68}Ga -FAPI-PET/CT improves diagnostic staging and radiotherapy planning of adenoid cystic carcinomas - Imaging analysis and histological validation

[5] [^{68}Ga]Ga-FAPI-46 PET for Visualization of Postinfarction Renal Fibrosis

[6] Performance and Prospects of [^{68}Ga]Ga-FAPI PET/CT Scans in Lung Cancer

[7] Clinical Evaluation of ^{68}Ga -FAPI-RGD for Imaging of Fibroblast Activation Protein and Integrin $\alpha\text{v}\beta\text{3}$ in Various Cancer Types

[8] New imaging method superior for diagnosing multiple types of cancer, with potential for targeted treatment

Cubic bézier curve approximation for the estimation of perivascular spaces measurements in MRI brain scans^{1*}

Author

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

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

José Bernal Moyano – German Centre for Neurodegenerative Diseases, Magdeburg

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

Citation

Coello, R.D., Hernández, M.V., Moyano, J.B., Wardlaw, J. Cubic bézier curve approximation for the estimation of perivascular spaces measurements in MRI brain scans.

Abstract

The morphometrics of perivascular spaces (PVS) such as diameter and length have been associated with stroke and hypertension. While the conventional ellipsoid approximation used for measuring diameter and length holds for straight PVS, significant inaccuracies occur for curved PVS. We propose a model based on cubic Bézier curves to cope with this limitation. On curved

^{1*} Supported by the Hilary & Galen Weston Foundation

digital reference objects, we show that our proposal outperforms the conventional method.

Introduction

The computational quantification of MRI-visible perivascular spaces (PVS) in the brain has increasingly gained attention because of its importance in validating PVS as a biomarker of brain health function. PVS morphometrics can provide useful information to complement volume and count. Median PVS diameter and length have been associated with stroke and hypertension [1]. The use of digital reference objects (DRO) has been instrumental in establishing the limits of validity of the PVS segmentation methods [2]. It has shown inaccuracies in the assessment of the diameter and length of curved PVS if the ellipsoid approximation model, implemented by the function `regionprops3` in MATLAB [4], is used. Here we propose the use of a cubic Bézier curve approximation model and compare the two methods using a curved DRO.

Methods

In this section, we describe the construction of our curved DRO and the proposed method for measuring the diameter and length of the simulated PVS.

Curved DRO

We constructed our curved DRO using the following parametric equations:

$$p(t) = [x(t) \quad y(t) \quad z(t)]^T = [0 \quad \cos(t/2) \quad t]^T, \quad (1)$$

where $-t_1 < t < -t_2$, (see Figure 1 (Left)), the PVS-DRO model is constructed from all the points such that the shortest distance to the line described in (1) is smaller than the radius $d/2$. The length of the line is given by:

$$l = d + \int_{t_1}^{t_2} \sqrt{\left(\frac{\partial x}{\partial t}\right)^2 + \left(\frac{\partial y}{\partial t}\right)^2 + \left(\frac{\partial z}{\partial t}\right)^2}. \quad (2)$$

For our simulation, we set the $t_1 = -6$ and $t_2 = 6$ and we vary the diameter of the PVS-DRO from 0.3 to 3.0 millimetres. Please note the choice of the curve for the DRO was arbitrary and any smooth continuous line can be used.

```

Algorithm 1 Cubic Bézier approximation
1 Input: Skeleton of the image  $S(n)$  with  $N$  voxels, initial control points  $\mathbf{b}_0^{(0)}, \mathbf{b}_1^{(0)}, \mathbf{b}_2^{(0)}, \mathbf{b}_3^{(0)}$ 
2 Optimise control points by alternating minimisation
3   for  $i = 0, 1, \dots$  do
4     Estimated cubic Bézier curve
5        $T^{(i)}(t) = [\mathbf{b}_0^{(i)} \ \mathbf{b}_1^{(i)} \ \mathbf{b}_2^{(i)} \ \mathbf{b}_3^{(i)}]Mf(t)$ 
6     Update the skeleton  $t$  values using the value of the curve closest point
7       for  $n = 0, 1, \dots, N$  do
8          $\mathbf{t}^{(i)}(n) = \underset{0 \leq t \leq 1}{\operatorname{argmin}} \left( S_p(n) - T^{(i)}(t) \right)^2 \left( S_p(n) - T^{(i)}(t) \right)$ 
9       end for
10    Update control points using least square error
11     $T = [T^{(i)}(\mathbf{t}^{(i)}(1)) \ T^{(i)}(\mathbf{t}^{(i)}(2)) \ \dots \ T^{(i)}(\mathbf{t}^{(i)}(N))]$ 
12     $C = TM^T = [c_0 \ c_1 \ c_2 \ c_3]^T$ 
13     $S = [S_p^{(1)} \ S_p^{(2)} \ \dots \ S_p^{(N)}]$ 
14     $E = S - \mathbf{b}_0 c_0 - \mathbf{b}_3 c_3$ 
15     $[\mathbf{b}_1 \ \mathbf{b}_2] = ([c_1 \ c_2]^T [c_1 \ c_2])^{-1} [c_1 \ c_2]^T E$ 
16  end for

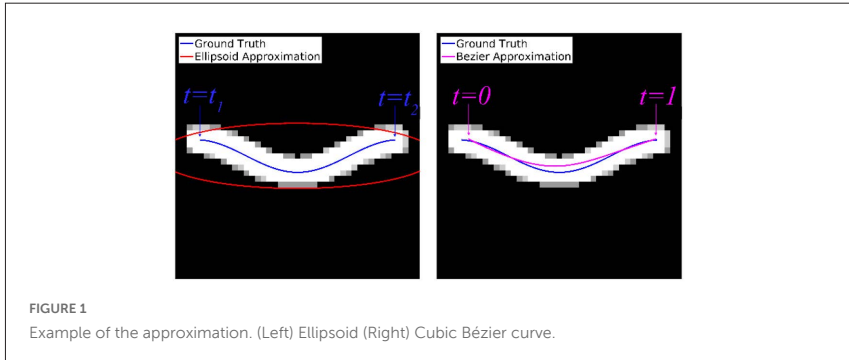
```

Proposed Cubic Bézier Curve Approximation

Cubic Bézier curves can be described with the following equation [3]:

$$p(t) = BMf(t) = [b_0 \ b_1 \ b_2 \ b_3] \begin{bmatrix} 1 & -3 & 3 & -1 \\ 0 & 3 & -6 & 3 \\ 0 & 0 & 3 & -3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \end{bmatrix}, \tag{3}$$

where the vectors conforming \mathbf{B} are control points. This curve is defined in the interval of $0 \leq t \leq 1$ (see Figure 1 (Right)). The \mathbf{b}_0 and \mathbf{b}_3 are the initial ($t = 0$)

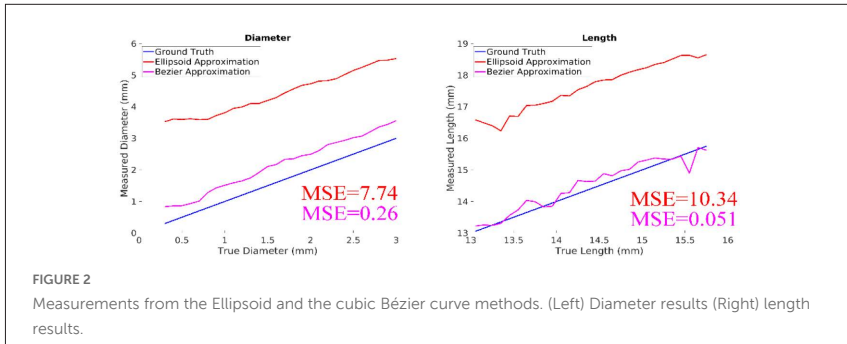


and the end ($t = 1$) points, respectively. From the skeleton ordered voxel list $S(n)$ we approximate the cubic Bézier curve using our proposed algorithm 1.

Results and Conclusions

Our PVS-DRO uses a voxel size of $0.35 \times 0.35 \times 0.35 \text{ mm}^3$, and the binary mask considers voxels containing any proportion of the PVS-DRO. Figure 2 shows the measurements of the length and diameter using the ellipsoid (red) and the cubic Bézier curve (magenta) models. The Bézier curve approximation performs better than the ellipsoid, being closer to the ideal measurements.

Our proposed method creates a DRO for curved tubular objects (e.g., PVS) by giving the parametric equation of the curve. Our optimisation method approximates a Bézier curve allowing us to accurately measure the diameter and length of curved objects representing PVS in brain MRI scans. It outperforms the ellipsoid approximation for measuring the length and diameter of curved objects.



Acknowledgements

This work received funds from the Hilary and Galen Weston Foundation (ref UB190097); Foundation Leducq Network of Excellence for the Study of Perivascular Spaces in Small Vessel Disease (16 CVD 05); the Row Fogo Charitable Trust (BRO-D.FID3668413); and the UK Dementia Research Institute at the University of Edinburgh funded by the Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

References

- [1] Ballerini, L., Booth, T., Hernández, M.d.C.V., Wiseman, S., Lovreglio, R., Maniega, S.M., Morris, Z., Pattie, A., Corley, J., Gow, A., et al.: Computational quantification of brain perivascular space morphologies: Associations with vascular risk factors and white matter hyperintensities. a study in the lothian birth cohort 1936. *NeuroImage: Clinical* 25, 102120 (2020)
- [2] Bernal, J., Valdés-Hernández, M.D., Escudero, J., Duarte, R., Ballerini, L., Bastin, M.E., Deary, I.J., Thrippleton, M.J., Touyz, R.M., Wardlaw, J.M.: Assessment of perivascular space filtering methods using a three-dimensional computational model. *Magnetic Resonance Imaging* 93, 33–51 (2022)

[3] Farin, G.: Curves and surfaces for computer-aided geometric design: a practical guide. Elsevier (2014)

[4] The MathWorks Inc.: MATLAB version: 9.14.0.2239454 (R2023a) (2023), <https://www.mathworks.com>

Independent testing of a publicly available CNN tool for hippocampal image segmentation

Author

FN Sneden, KJ Ferguson, S Muñoz Maniega, M Valdés Hernández, JM Wardlaw –
Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

Citation

Sneden, F.N., Ferguson, K.J., Maniega, S.M., Hernández, M.V., Wardlaw, J.M.
Independent testing of a publicly available CNN tool for hippocampal image segmentation.

Abstract

Hippocampal atrophy is a common measure in ageing and dementia. Automated segmentation methods are often computationally expensive. Hipp-Mapp3r is a CNN-based algorithm trained on patients with significant hippocampal degeneration. We tested HippMapp3r in the Lothian Birth Cohort 1936. We ran HippMapp3r with 2 preprocessing methods: 1) on skull stripped images (recommended) and 2) transformed to standard space on 602 participants with T1-weighted imaging and manually corrected ground truth masks. We found significant differences between using these two preprocessing methods (mean Dice = 0.594 versus 0.676, $p < 0.001$). We found a proportional bias where the greatest variation between our measurements at smaller hippocampal volumes. We suggest automated techniques require multicenter data training and testing on different datasets prior to deployment.

Introduction

The hippocampus plays a crucial role in many cognitive processes (Valdés Hernández et al., 2017). Hippocampal volumes are an important biomarker of ageing and de-mentia but manual segmentation of this structure from brain images is time consuming and often prone to inter-rater variability issues (Goubran et al., 2019). Automated segmentation methods of the brain are often limited by many factors, including training on a locally available population, using a limited number of imaging protocols and some being computationally demanding (Goubran et al., 2019). HippMapp3r is a publicly available CNN-based segmentation tool that attempts to correct these issues by being trained on patients from various samples, including scans with significant atrophy/disease and being computationally inexpensive. HippMapp3r outperforms other automatic segmentation methods, with an average Dice coefficient of 0.89 when compared to manual segmentations (Goubran et al., 2019). The aim of this study is to evaluate the use of HippMapp3r in an ageing population, the Lothian Birth Cohort 1936 (LBC1936) (Deary et al., 2007). We hypothesize:

1. Our use of HippMapp3r will perform differently than described by those who developed it due to differences between the populations studied.
2. Different pre-processing methods will differentially impact in the results. For example skull stripped imagines will have different impact than transforming images to the standard space.

Methods

In this study, we ran HippMapp3r on the first MRI brain scan from participants in LBC1936 that had T1-weighted images (FOV = 256 x 256mm; Matrix = 192 x 192; Slices = 160; Thickness = 1.3mm; voxel size = 1 x 1 x 1.3mm; TR = 10ms; TE = 4ms; TI = 500ms), manually corrected ground truth hippocampal masks (methods by MacLulich et al., 2002) and intracranial volume (ICV) mask. We repeated hippocampal mask generation using HippMapp3r under different circumstances:

1. Preprocessing method 1: T1 images in native image space and skull stripped (recommended by HippMapp3r).
2. Preprocessing method 2: T1 images linear (affine) transformed to MNI152 space using RNiftyReg.

We compared these newly generated hippocampal masks with our ground truth measurements, using Dice similarity scores and positive predictive value, calculated in Matlab 2018. Visual quality control was performed independently in a sample of 40 randomized participants.

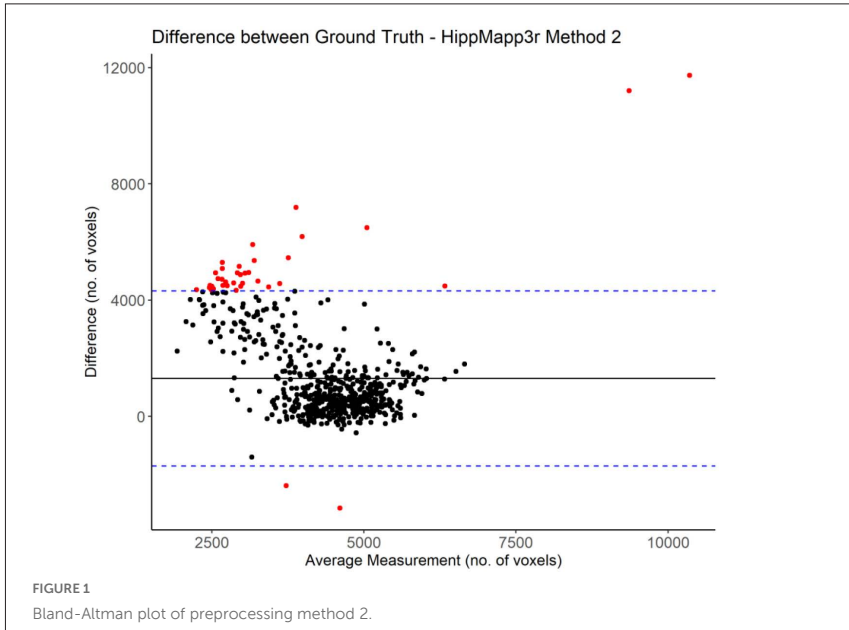
Results

MRI brain scans from 602 participants in the LBC1936 study were included in our analysis. When comparing to ground truth masks, we found HippMapp3r performed significantly better ($p < 0.001$) when images were transformed to MNI152 space before image segmentation; Preprocessing method 1 demonstrated a median Dice score 0.623 versus the preprocessing method 2 with median Dice score 0.760 (Table 1).

We compared the volumes obtained from HippMapp3r masks using Preprocessing method 2 to our ground truth measurements to assess for any bias. We found the greatest variation where hippocampi are smaller (Figure 1).

TABLE 1: Dice similarity scores and other related performance metrics between the two methods. Dice = Dice coefficient; Spec = Specificity; Sens = Sensitivity; PPV = Positive Predictive Value. Assessed using Wilcoxon Signed T

	Preproc. Method 1 (n = 602)			Preproc. Method 2 (n = 602)			
	Mean	Median	SD	Mean	Median	SD	Sig.
Dice	0.594	0.623	0.157	0.676	0.760	0.226	<0.001
Spec	0.999	0.999	0.000	0.997	0.998	0.002	<0.001
Sens	0.494	0.502	0.173	0.807	0.838	0.153	<0.001
PPV	0.787	0.801	0.106	0.619	0.722	0.242	<0.001



Discussion

These data demonstrate differences between the performance of HippMap3r in the original study, and our performance on a different population-based dataset. Our analysis found that normalizing the LBC1936 images to MNI152 space significantly increased performance metrics. This may be due to alterations in image resolution after co-registration or changes in direction of head tilt within the images. Additionally, our data suggests a greater mismatch between ground truth and HippMap3r labels, with a mean Dice of 0.68 achieved in this sample, compared with the previously reported Dice of 0.89. Bland-Altman analysis showed a likely proportional bias where HippMap3r consistently failed to assess smaller hippocampal volumes. Further testing of Hipp-Map3r is required to fully assess reasons for differences between ground truth measurements and

CNN-based masks. HippMapp3r generated masks sometimes excluded anterior portions of the hippocampus and detection of true hippocampal boundaries was affected by the presence of CSF in the hippocampal fissure in some participants. Our results, hence, reinforce the necessity for collaborative training and developing CNN-based algorithms if they are to be widely used, and multi-center testing prior to their deployment in publicly available repositories.

References

- Deary, I.J., Gow, A.J., Taylor, M.D., Corley, J., Brett, C., Wilson, V., Campbell, H., Whalley, L.J., Visscher, P.M., Porteous, D.J., Starr, J.M., 2007. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr* 7, 28. <https://doi.org/10.1186/1471-2318-7-28>
- Goubran, M., Ntiri, E.E., Akhavein, H., Holmes, M., Nestor, S., Ramirez, J., Adamo, S., Ozzoude, M., Scott, C., Gao, F., Martel, A., Swardfager, W., Masellis, M., Swartz, R., MacIntosh, B., Black, S.E., 2019. Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks. *Hum Brain Mapp* 41, 291–308. <https://doi.org/10.1002/hbm.24811>
- MacLulich, A.M.J., Ferguson, K.J., Deary, I.J., Seckl, J.R., Starr, J.M., Wardlaw, J.M., 2002. Intracranial capacity and brain volumes are associated with cognition in healthy elderly men. *Neurology* 59, 169–174. <https://doi.org/10.1212/wnl.59.2.169>
- Valdés Hernández, M. del C., Cox, S.R., Kim, J., Royle, N.A., Muñoz Maniega, S., Gow, A.J., Anblagan, D., Bastin, M.E., Park, J., Starr, J.M., Wardlaw, J.M., Deary, I.J., 2017. Hippocampal morphology and cognitive functions in community-dwelling older people: the Lothian Birth Cohort 1936. *Neurobiology of Aging, Roadmap to Alzheimer's Biomarkers in the Clinic* 52, 1–11. <https://doi.org/10.1016/j.neurobiolaging.2016.12.012>

Identifying MRI sequence type from pixel data to enable cohort building from routinely collected brain scans

Author

Smarti Reel – School of Medicine, University of Dundee, Dundee DD2 4BF, UK
Esma Mansouri–Benssassi – School of Medicine, University of Dundee, Dundee DD2 4BF, UK

Kara Moraw – EPCC, The University of Edinburgh, Edinburgh EH8 9BT, UK
Susan Krueger – School of Medicine, University of Dundee, Dundee DD2 4BF, UK
Emily Jefferson – School of Medicine, University of Dundee, Dundee DD2 4BF, UK;
Health Data Research (HDR), London NW1 2BE, UK

Citation

Reel, S., Mansouri–Benssassi, E., Moraw, K., Krueger, S., Jefferson, E. Identifying MRI sequence type from pixel data to enable cohort building from routinely collected brain scans.

Abstract

To develop robust machine learning (ML) models which are fit for use in real-world clinical scenarios, it is vital to train and evaluate these algorithms on routinely collected medical data. The Scottish Medical Imaging (SMI) service provides researchers with safe and secure access to a large dataset of routinely collected medical images. This dataset contains around 25 million studies across 11 modalities. A key step in granting access to a research-ready dataset is cohort building, i.e. identifying the data requirements of the

research project and extracting a subset of relevant medical images that meet these requirements. However, building cohorts of medical images from routinely collected scans is a challenging process [1]. The Digital Imaging and Communications in Medicine (DICOM) standard provides a wide range of metadata consisting of tags providing additional information about an image. Missing information, inconsistent tags and variability across different scanners make it highly challenging to query DICOM images solely using metadata [2–4] While the DICOM standard provides tags that can indicate the sequence type, this information is often not complete enough in routinely collected images to build meaningful cohorts. The information from pixel data in addition to metadata can enhance the ability to build accurate and reliable cohorts.

As part of the PICTURES [5] project, this work aims to develop a flexible and reusable pipeline applying ML techniques to extract information from the pixel data in routinely collected images. A robust method is proposed that uses pixel images from routinely collected brain magnetic resonance imaging (MRI) scans for identifying their sequence type: T1 or T2 (see Figure 1) for cohort building. Usually, a clinician uses image data to identify T1- and T2-weighted images by visually analysing the Cerebrospinal fluid (CSF), a dark CSF corresponds to T1-weighted imaging and bright to T2-weighted imaging

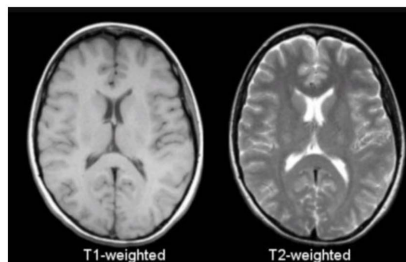


FIGURE 1
Shows T1-weighted and T2-weighted scans [6].

[6]. A more technical solution is required in predicting the sequence type from pixel data for cohort building to facilitate and automate the process. Previous works [7, 8] leveraged ML techniques to classify MRI sequence types and proposed deep Convolutional Neural Network (CNN) architectures that were directly applied to the images and predict the sequence type. In contrast, this work opted for a modular approach that can be easily reused and expanded for further use cases, resulting in a scalable and efficient pipeline for cohort building in the context of the SMI dataset.

The model architecture consisted of two parts: an autoencoder [9] which is used to derive numerical features from the pixel data, and a classifier which was then trained on those extracted features. Both networks were trained separately. This decoupling meant that the extracted features were independent of classification and can in fact be reused for other analyses apart from sequence-type classification. The autoencoder has a simple architecture based on a convolutional neural network. It was first trained for image reconstruction and then the encoder part was used to extract features from pixel images. It was trained to reproduce the input with minimum error. The low-dimensional embeddings calculated by the encoder network were then stored as numerical features of the input images. For the sequence type classification task, a Support Vector Machine (SVM) [10] was trained on those extracted features. Using the features as input instead of an entire image improved the performance and memory efficiency.

The models were trained on the open-source IXI dataset [5]. IXI dataset contains nearly 600 MR images from normal, healthy subjects with data from three different hospitals in London. This work focused on the most common MRI sequences: T1-weighted and T2-weighted scans. MRI brain scans are composed of a varied number of slices, where each slice is a 2D image of a brain slice. Combining these images yields a 3D brain volume, but for sequence-type classification, it seemed unnecessary to examine the entire brain. Instead, initially, a single mid-slice was used to classify the MRI sequence type. Later, to build a robust model for countering challenges associated with routinely collected data such as partial brain scans and varied

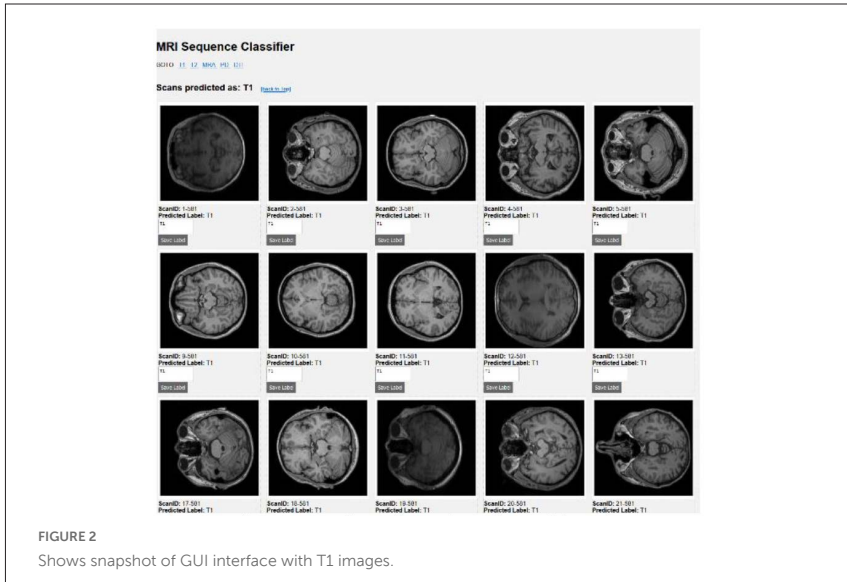


image orientations, 10 slices were sampled from each of the three axes per sample in the IXI dataset. In a preprocessing step, these 2D images were resized and then fed into the autoencoder. The experiments show an overall accuracy of 97% for the MRI sequence type classification when evaluated on the IXI dataset.

This pre-trained model can be applied to routinely collected brain MRI images. An intuitive graphical user interface (GUI) (see Figure 2) was also developed which allows an expert user to correct any misclassifications. These corrected labels can then be fed back to retrain the classifier and refine its prediction performance. The method uses a human-in-the-loop approach to enable different users such as research coordinators, researchers, or data managers to share, annotate and extract relevant cohorts when DICOM tags metadata is not available or is insufficient in identifying datasets of interest. Automated MRI sequence type classification

combined with a human-in-the-loop procedure for correcting predictions yields a powerful tool for deriving information from medical imaging pixel data.

References

- [1] Jefferson, E., Krueger, S., Macleod, R., Sutherland, J., Nind, T., Mudie, R., Prodan, B., Brooks, A., Wallace, R., Morris, C., Caldwell, J., Baxter, R., Parsons, M.: An architecture for building cohorts of images from real-world clinical data from the whole Scottish population supporting research and AI development. *Int J Popul Data Sci.* 7, 1916. <https://doi.org/10.23889/ijpds.v7i3.1916>.
- [2] Park, C., You, S.C., Jeon, H., Jeong, C.W., Choi, J.W., Park, R.W.: Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. *Yonsei Med J.* 63, S74–S83 (2022). <https://doi.org/10.3349/ymj.2022.63.S74>.
- [3] Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing Medical Imaging Data for Machine Learning. *Radiology.* 295, 4–15 (2020). <https://doi.org/10.1148/radiol.2020192224>.
- [4] Basu, A., Warzel, D., Eftekhari, A., Kirby, J.S., Freymann, J., Knable, J., Sharma, A., Jacobs, P.: Call for Data Standardization: Lessons Learned and Recommendations in an Imaging Study. *JCO Clin Cancer Inform.* 3, 1–11 (2019). <https://doi.org/10.1200/CCI.19.00056>.

[5] Pictures: Supporting the use of data for health care research - Image on a Mission, <https://www.imageonamission.ac.uk/>, last accessed 2023/06/24.

[6] David C Preston: MRI Basics, <https://case.edu/med/neurology/NR/MRI%20Basics.htm>, last accessed 2023/06/25.

[7] Vieira de Mello, J.P., Paixão, T.M., Berriel, R., Reyes, M., Badue, C., De Souza, A.F., Oliveira-Santos, T.: Deep Learning-based Type Identification of Volumetric MRI Sequences. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1–8 (2021). <https://doi.org/10.1109/ICPR48806.2021.9413120>.

[8] Ranjbar, S., Singleton, K.W., Jackson, P.R., Rickertsen, C.R., Whitmire, S.A., Clark-Swanson, K.R., Mitchell, J.R., Swanson, K.R., Hu, L.S.: A Deep Convolutional Neural Network for Annotation of Magnetic Resonance Imaging Sequence Type. *J Digit Imaging*. 33, 439–446 (2020). <https://doi.org/10.1007/s10278-019-00282-4>.

[9] Baldi, P.: Autoencoders, unsupervised learning and deep architectures. In: *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop - Volume 27*. pp. 37–50. JMLR.org, Washington, USA (2011).

[10] Zhang, Y.: Support Vector Machine Classification Algorithm and Its Application. In: Liu, C., Wang, L., and Yang, A. (eds.) *Information Computing and Applications*. pp. 179–186. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34041-3_27.

Improving venous tumour thrombus segmentation in clear cell renal cell cancer MRI scans with a two-stage 3D nnU-Net

Author

Robin Haljak – Department of Physics, University of Cambridge, Cambridge, UK

Hanna Wyciszczok – Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

Ines P. Machado – Department of Oncology, University of Cambridge, Cambridge, UK; Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK

Grant D. Stewart – Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; Department of Surgery, University of Cambridge, Cambridge, UK

James O. Jones – Department of Oncology, University of Cambridge, Cambridge, UK; Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK

Stephan Ursprung – Department of Radiology, University of Cambridge, Cambridge, UK; Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

Ferdia A. Gallagher – Department of Radiology, University of Cambridge, Cambridge, UK; Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK;

Department of Surgery, University of Cambridge, Cambridge, UK

Mireia Crispin-Ortuzar – Department of Oncology, University of Cambridge, Cambridge, UK; Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK

Citation

Haljak, R., Wyciszczok, H., Machado, I.P., Stewart, G.D., Jones, J.O., Ursprung, S., Gallagher, F.A., Crispin-Ortuzar, M. Improving venous tumour thrombus segmentation in clear cell renal cell cancer MRI scans with a two-stage 3D nnU-Net.

Abstract

An unusual hallmark of kidney cancer is the biological pre-disposition for vascular invasion, with the extension of the venous tumour thrombus (VTT) into the inferior vena cava occurring in 4-15% of cases. Automated segmentation of the VTT would be beneficial for the diagnostic evaluation of kidney cancer. However, the location, size and shape of the VTT are highly variable, and the training data is limited, making the automatic segmentation task extremely challenging. A two-stage localization-refinement-based 3D nnU-Net model is proposed to significantly increase the segmentation accuracy of the VTT in kidney cancer MRI scans. The proposed model involves two main steps. In the first step, the VTT is localised and an initial segmentation is created. In the second step, the segmentation is expanded to more accurately segment the VTT. Comparative experiments were conducted on the NAXIVA clinical trial data set. The proposed method found 7/12 cases of VTT with a Dice similarity coefficient of 0.402 in the test set. The effect of MRI image quality on automatic segmentation accuracy was also quantified. The results demonstrate the potential of the proposed model for VTT segmentation.

Introduction

Kidney cancer is the 7th most common newly diagnosed cancer. It claims the lives of 4700 individuals annually in the UK and accounts for 3% of all cancers [1]. A form of kidney cancer is the venous tumour thrombus (VTT), which is an extension of a tumour from the kidney into the renal vein and inferior vena cava (Figure 1a). NAXIVA [9] was the first prospective study to evaluate drug treatment in reducing the extent of VTT which required manual segmentation of MRI images. This task is challenging and time-consuming and there is demand for developing machine learning methods for semantic segmentation of VTT MRI scans. Furthermore, improved segmentation might offer features predictive of response. Recently the nnUNet framework has shown promising results for semantic image segmentation in a variety of medical image analysis applications [2, 5]. Robustness to imaging error sources, including imaging noise and the partial volume effect, is a significant requirement for automated segmentation models [8]. Encoder-decoder

architectures (like nnU-Net) have been shown to suffer in segmentation accuracy for structures with low contrast and blurred boundaries [11]. In this work the first automatic segmentation model for VTT, a two-stage localization-refinement-based 3D nnU-Net model, is proposed to segment the VTT in kidney cancer MRI scans.

Materials and Methods

The proposed two-stage model for VTT segmentation consisting of 2 nnU-Net stages is shown in Figure 1a. The 3D nnU-Net cascade [5] serves as the basis for the model. However, a different approach was adopted for the network tasks to accommodate the small data set and foreground-background imbalance in the VTT data set [3]. The first stage is a regular nnU-Net network that detects, and segments the VTT. The image is cropped around the initial segmentation and input into the second stage with the initial segmentation softmax added as an additional layer [10]. The second stage produces a refined segmentation, using a nnU-Net network, that has

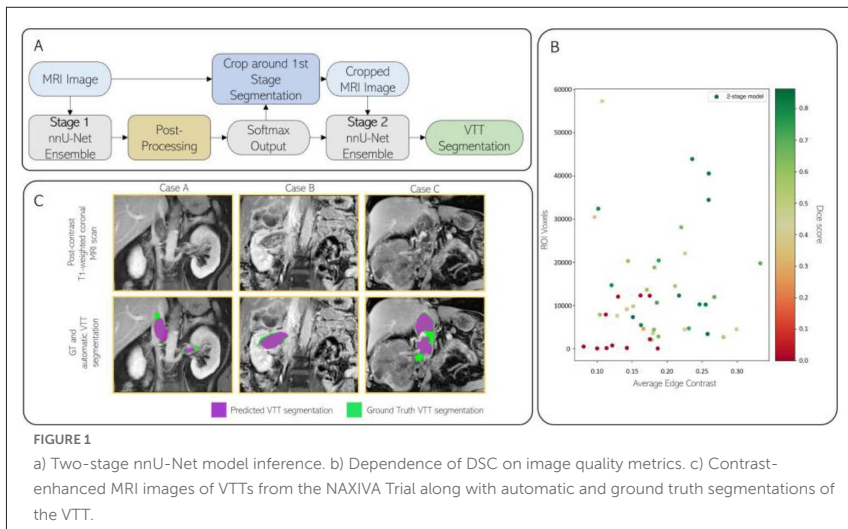


FIGURE 1

a) Two-stage nnU-Net model inference. b) Dependence of DSC on image quality metrics. c) Contrast-enhanced MRI images of VTTs from the NAXIVA Trial along with automatic and ground truth segmentations of the VTT.

been trained on the task of completing a partial segmentation of the VTT, using synthetic partial segmentations. The NAXIVA dataset consists of 47 post-contrast T1-weighted coronal MRI scans of 19 cancer patients. A total of 35 images (split by patient) were used as training set and the remaining 12 were held out as a test set. 5-fold cross-validation was used on the training set. Performance was evaluated using the Dice Similarity Coefficient (DSC) [7] and the Area Under Precision-Recall Curve (AUPRC) [4, 6].

Results and Discussion

The evaluation of the baseline and two-stage nnU-Net models is given in Table

1. The two-stage nnU-Net outperformed the baseline in both DSC and AUPRC. The correlation between image quality metrics and segmentation DSC is given in Table 2. VTTs containing a greater number of voxels exhibited higher automatic segmentation DSC, as they are comparatively easier to detect and localize. It was shown that images with larger voxel sizes, and therefore a larger partial volume effect, have worse automatic segmentation DSC. Higher contrast at the edges of the VTT was also found to increase the automatic segmentation DSC. Metrics of image noise were found not to correlate with segmentation DSC: image noise is not the limiting factor for automatic segmentation quality.

TABLE 1: Comparison of quantitative segmentation evaluation metrics for 2-stage nnU-Net and baseline nnU-Net

Model	VTTs found (Cross-Validation)	VTTs found (Test set)	DSC (Cross-Validation)	DSC (Test set)	AUPRC
nnU-Net baseline	26/35	6/12	0.376 ± 0.330	0.184 ± 0.291	0.454
2-stage nnU-Net	29/35	7/12	0.511 ± 0.287	0.402 ± 0.373	0.475

TABLE 2: Correlations between automatic segmentation Dice Similarity Coefficients and image quality metric

Data set	Voxel Volume	VTT Voxels	SNR	Edge Contrast
Cross-validation sets ($n = 35$)	-0.356 ($p = 0.036$)	0.307 ($p = 0.073$)	-0.107 ($p = 0.543$)	0.331 ($p = 0.052$)
Test set ($n = 12$)	-0.734 ($p = 0.007$)	0.627 ($p = 0.029$)	-0.530 ($p = 0.076$)	0.634 ($p = 0.027$)

Conclusion

A two-stage nnU-Net model based on the localisation-refinement principle was trained for VTT segmentation, achieving a DSC of 0.402 and finding 7/12 VTT cases in the independent test set. This improved upon the baseline nnU-Net DSC of 0.184 with 6/12 cases found. Statistically significant correlations between segmentation DSC and image resolution, VTT voxel number, and edge contrast were found, reinforcing the fact that images with high spatial resolution and contrast are beneficial to automatic segmentation quality.

Acknowledgements

GDS/FAF/MCO are supported by The Mark Foundation for Cancer Research, the Cancer Research UK Cambridge Centre [C9685/A25177 and CTRQQR-2021\100012] and NIHR Cambridge Biomedical Research Centre (NIHR203312). FAF is also supported by Cancer Research UK (C19212/A27150). MCO also received support from the Academy of Medical Sciences (G117526). SU was supported by the Cambridge Trust. Additional support was also provided by the EPSRC Tier-2 capital grant (EP/P020259/1). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

References

[1] Kidney cancer statistics (May 2019), <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/kidney-cancer>

[2] Anaya-Isaza, A., Mera-Jiménez, L., Zequera-Diaz, M.: An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked* **26**, 100723 (Jan 2021). <https://doi.org/10.1016/j.imu.2021.100723>, <https://www.sciencedirect.com/science/article/pii/S2352914821002033>

[3] Bazgir, O., Barck, K., Carano, R.A.D., Weimer, R.M., Xie, L.: Kidney segmentation using 3D U-Net localized with Expectation Maximization. In: 2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). pp. 22–25 (Mar 2020). <https://doi.org/10.1109/SSIAI49293.2020.9094601>, <http://arxiv.org/abs/2003.09075>, arXiv:2003.09075 [cs, eess]

[4] Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240. ICML '06, Association for Computing Machinery, New York, NY, USA (Jun 2006). <https://doi.org/10.1145/1143844.1143874>, <https://dl.acm.org/doi/10.1145/1143844.1143874>

[5] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**(2), 203–211 (Feb 2021). <https://doi.org/10.1038/s41592-020-01008-z>, <https://www.nature.com/articles/s41592-020-01008-z>, number: 2 Publisher: Nature Publishing Group

[6] Ozenne, B., Subtil, F., Maucort-Boulch, D.: The precision–recall curve over-came the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology* **68**(8), 855–859 (Aug 2015). <https://doi.org/10.1016/j.jclinepi.2015.02.010>, <https://www.sciencedirect.com/science/article/pii/S0895435615001067>

[7] Popovic, A., De la Fuente, M., Engelhardt, M., Radermacher, K.: Statistical validation metric for accuracy assessment in medical image segmentation. *International Journal of Computer Assisted Radiology and Surgery* **2**, 169–181 (2007). <https://doi.org/https://doi.org/10.1007/s11548-007-0125-1>

[8] Sharma, N., Aggarwal, L.M.: Automated medical image segmentation techniques. *Journal of Medical Physics* **35**(1), 3 (Mar 2010). <https://doi.org/10.4103/0971-6203.58777>, https://journals.lww.com/jomp/Fulltext/2010/35010/Automated_medical_image_segmentation_techniques.2.aspx

[9] Stewart, G.D., Welsh, S.J., Ursprung, S., Gallagher, F.A., Jones, J.O., Shields, J., Smith, C.G., Mitchell, T.J., Warren, A.Y., Bex, A., Boleti, E., Carruthers, J., Eisen, T., Fife, K., Hamid, A., Laird, A., Leung, S., Malik, J., Mendichovszky, I.A., Mumtaz, F., Oades, G., Priest, A.N., Riddick, A.C.P., Venugopal, B., Welsh, M., Riddle, K., Hopcroft, L.E.M., NAXIVA Trial Group, Jones, R.J.: A Phase II study of neoadjuvant axitinib for reducing the extent of venous tumour thrombus in clear cell renal cell cancer with venous invasion (NAXIVA). *British Journal of Cancer* **127**(6), 1051–1060 (Oct 2022). <https://doi.org/10.1038/s41416-022-01883-7>

[10] Zhang, G., Yang, Z., Huo, B., Chai, S., Jiang, S.: Multiorgan segmentation from partially labeled datasets with conditional nnU-Net. *Computers in Biology and Medicine* **136**, 104658 (Sep 2021). <https://doi.org/10.1016/j.compbimed.2021.104658>, <https://www.sciencedirect.com/science/article/pii/S001048252100452>

[11] Zhou, S., Nie, D., Adeli, E., Yin, J., Lian, J., Shen, D.: High-Resolution Encoder–Decoder Networks for Low-Contrast Medical Image Segmentation. *IEEE Transactions on Image Processing* **29**, 461–475 (2020). <https://doi.org/10.1109/TIP.2019.2919937>, conference Name: IEEE Transactions on Image Processing.

XGBoost classifier-based survival prediction in head and neck cancer patients using pre-treatment PET images

Author

Mahima Philip – Institute of Applied Health Sciences, University of Aberdeen, AB25 2ZD, UK

Jessica Watts – National Health Service Grampian, Aberdeen AB15 6RE, UK

Andy Welch – Institute of Education in Healthcare and Medical Sciences, University of Aberdeen, AB25 2ZD, UK

Fergus McKiddie – National Health Service Grampian, Aberdeen AB15 6RE, UK

Mintu Nath – Institute of Applied Health Sciences, University of Aberdeen, AB25 2ZD, UK

Citation

Philip, M., Watts, J., Welch, A., McKiddie, F., Nath, M. XGBoost classifier-based survival prediction in head and neck cancer patients using pre-treatment PET images,

Abstract

High-dimensional radiomics features may play an important role in the prognosis of Head and Neck Squamous Cell Carcinoma (HNSCC) patients. We investigated the prediction of the survival outcomes in HNSCC patients using features obtained from pre-treatment positron emission tomography (PET) images of primary tumours. The study included 265 HNSCC patients. A total of 241 PET radiomic features were extracted from PET images using LIFEx software. We used 85% of the training data with radiomics features and patient age to develop the XGBoost classifier to predict the survival

outcome and validated the model on the remaining dataset. Before fitting the model, we conducted feature selection by removing features with a higher percentage of missing values, higher correlation and low variance, and pre-processed the data using minmax scaling. We implemented the synthetic minority oversampling technique (SMOTE) to account for target imbalance. The model demonstrated a modest estimate (0.73) of the area under the receiver operating curve (AUC) in the test dataset with reasonable performance metrics. Results suggested PET radiomic features may improve the prediction of overall survival probability in HNSCC patients.

Introduction

Head and Neck Squamous Cell Carcinoma (HNSCC) is a heterogeneous group of tumours developing from the squamous epithelium of the proximal aerodigestive tract. With a high mortality rate of 40 to 50%[1], HNSCC patients often present with locally advanced disease and the treatment is effective in only 50% of the patients [2]. Accurate and early diagnosis and prediction of survival outcomes are vital for patient-specific treatment planning to improve the survival rate and to enhance patient care[3].

Machine learning (ML), a subset of Artificial Intelligence (AI), is a valuable tool to construct prediction models by learning patterns from massive data such as PET radiomics. ML methods have reasonable predictive accuracy for outcomes like cancer susceptibility, recurrence and survival using high-dimensional radiomics data, thus demonstrating prospects of effective disease management in patients[4]. Traditional ML algorithms such as support vector machine (SVM), back propagation neural network (BPNN), multi-layer perception (MLP) and Bayesian network constituted most prediction models for cancer [5]. Extreme gradient boosting (XGBoost), an improvement over gradient-boosted decision tree (GBDT), is a complex and novel machine learning algorithm. XGBoost can achieve remarkable accuracy in multiple regression tasks and has shown satisfactory results in several machine learning competitions, outperforming other algorithms commonly used in predictive model construction[6]. XGBoost algorithm has been widely used in industries, but rarely used in medical research [5].

In this study, we implemented the XGBoost machine learning framework to develop a predictive model of survival outcome in HNSCC patients using high-dimensional PET radiomics features and patient age as predictors and assessed the model performance.

Methods

The study included 265 HNSCC patients (216 survived) with PET images obtained from The Cancer Imaging Archive (TCIA) database. We used LIFEx software (v 7.2.3) to segment the primary tumour on the pre-treatment PET images and extract radiomic features. A total of 241 PET-based features across seven feature classes were obtained. Additionally, we considered the patient's age as a predictor variable. We defined 'not survived' as an outcome if the patient did not survive beyond seven years of follow-up.

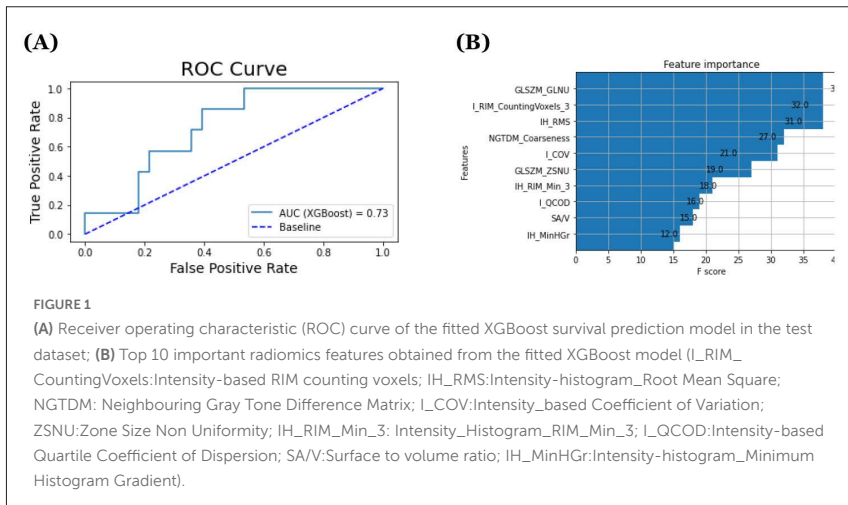
We partitioned the data into training ($n=197$) and test ($n=35$) sets with an 85:15 split. Before the model fitting, we considered different feature selection approaches. We removed features with more than 85% missing values and low variance and retained features that had minimal correlation with each other ($r < 0.95$). Feature selection approaches resulted in 79 features for the final model development. We pre-processed all selected features using minmax scaling. To account for the target imbalance, we implemented the synthetic minority oversampling technique (SMOTE). We developed the XGBoost classifier (Python scikit-learn library) on the training set. We considered a randomised search of hyperparameters and performed five-fold cross-validation to identify the most optimal hyperparameters of the XGBoost classifier. We internally validated the best fitted XGBoost classifier on the test dataset to construct the receiver operating characteristic (ROC) curve, and estimate the area under the ROC (AUC) and other performance metrics [5].

Results

PET images of HNSCC patients were obtained from TCIA. These patients have their pre-treatment PET images available and none of the patients had distant metastasis at the time of presentation. The cohort consisted of 206

male and 59 female patients, with a mean age of 63.88 years. Oropharynx was the primary site of the tumour in 71.7% of the cases and hypopharynx was the least reported site (3.77%).

We calibrated the XGBoost classifier using five-fold cross-validation on the training dataset and identified the best model. The XGBoost classifier exhibited good performance in the training set (AUC=0.82) and test set (AUC=0.73) (Fig. 1A) with reasonable estimates of sensitivity (0.86) and specificity (0.61), suggesting the classifier could reasonably discriminate between the two classes (survived and not survived). The model revealed that GLSZM_GLNU (Gray Level Size Zone Matrix Gray Level Non-Uniformity) was the most important feature and several intensity-based features had similar importance ranking (Fig. 1B).



Discussion

We developed a predictive model using the XGBoost classifier to predict survival in HNSCC patients. We observed that the XGBoost classifier had a good performance in the test cohort (AUC=0.73). To the best of our knowledge, this is the first study to apply an XGBoost classifier with PET radiomics features to predict the survival of HNSCC patients, suggesting the potential of such a predictive model in clinical practice.

Several prediction models other than XGBoost were evaluated for survival in HNSCC from PET radiomics. The AUC values of these models ranged between 0.60-0.87. XGBoost classifier demonstrated the best performance (AUC=0.84) compared to other classifiers in cervical cancer survival prediction[8]. XGBoost algorithm has been reported to effectively predict patient prognosis in other cancer types[5,9–11]. XGBoost has therefore exhibited robust performance for processing large-scale and high-dimensional data.

Our study has some limitations. The sample size used for fitting the classifier was relatively small. The model did not include other clinical features and the model was only internally validated. Future studies may focus on the development of the predictive model with PET radiomics and other demographic and clinical features and validation with external datasets to provide high-level evidence for clinical applications.

References

- [1] Moskowitz J, Moy J, Ferris RL. Immunotherapy for Head and Neck Squamous Cell Carcinoma. *Curr Oncol Rep.* 2018;20:22. <https://doi.org/10.1007/s11912-018-0654-5>
- [2] Solomon B, Young RJ, Rischin D. Head and neck squamous cell carcinoma: Genomics and emerging biomarkers for immunomodulatory cancer treatments. *Semin Cancer Biol.* 2018;52:228–40. <https://www.sciencedirect.com/science/article/pii/S1044579X17302560>

[3] Alabi RO, Youssef O, Pirinen M, Elmusrati M, Mäkitie AA, Leivo I, et al. Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for future—A systematic review. *Artif Intell Med.* 2021;115:102060.

[4] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.* 2020;471:61–71.

[5] Jiang J, Pan H, Li M, Qian B, Lin X, Fan S. Predictive model for the 5-year survival status of osteosarcoma patients based on the SEER database and XGBoost algorithm. *Sci Rep. England;* 2021;11:5542.

[6] Li B, Zhang F, Niu Q, Liu J, Yu Y, Wang P, et al. A molecular classification of gastric cancer associated with distinct clinical outcomes and validated by an XGBoost-based prediction model. *Mol Ther Nucleic Acids.* 2023;31:224–40.

[7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min [Internet].* New York, NY, USA: Association for Computing Machinery; 2016. p. 785–794. <https://doi.org/10.1145/2939672.2939785>

[8] Yu W, Lu Y, Shou H, Xu H, Shi L, Geng X, et al. A 5-year survival status prognosis of nonmetastatic cervical cancer patients through machine learning algorithms. *Cancer Med. United States;* 2023;12:6867–76.

[9] Gong X, Zheng B, Xu G, Chen H, Chen C. Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer. *J Thorac Dis. China;* 2021;13:6240–51.

[10] Huang Z, Hu C, Chi C, Jiang Z, Tong Y, Zhao C. An Artificial Intelligence Model for Predicting 1-Year Survival of Bone Metastases in Non-Small-Cell Lung Cancer Patients Based on XGBoost Algorithm. *Biomed Res Int.* 2020;2020:3462363.

[11] Wei L, Huang Y, Chen Z, Lei H, Qin X, Cui L, et al. Artificial Intelligence Combined With Big Data to Predict Lymph Node Involvement in Prostate Cancer: A Population-Based Study. *Front Oncol.* Switzerland; 2021;11:763381.

Prediction of cystic evolution of tumours in ovarian cancer with CT-derived features

Author

Maria Delgado–Ortet – Department of Radiology, University of Cambridge, Cambridge, United Kingdom; Cancer Research UK Cambridge Centre, Cambridge, United Kingdom

Leonardo Rundo – Department of Radiology, University of Cambridge, Cambridge, United Kingdom; Cancer Research UK Cambridge Centre, Cambridge, United Kingdom; Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno (SA), Fisciano, Italy

Ramona Woitek – Department of Radiology, University of Cambridge, Cambridge, United Kingdom; Cancer Research UK Cambridge Centre, Cambridge, United Kingdom; Research Center for Medical Image Analysis & Artificial Intelligence (MIAAI), Danube Private University, Krems, Austria

Evis Sala – Department of Radiology, University of Cambridge, Cambridge, United Kingdom; Cancer Research UK Cambridge Centre, Cambridge, United Kingdom; Dipartimento Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Policlinico Universitario A. Gemelli IRCCS, Rome, Italy; Dipartimento di Scienze Radiologiche ed Ematologiche, Università Cattolica del Sacro Cuore, Rome, Italy

Lorena Escudero Sánchez – Department of Radiology, University of Cambridge, Cambridge, United Kingdom; Cancer Research UK Cambridge Centre, Cambridge, United Kingdom

Citation

Delgado–Ortet, M., Rundo, L., Woitek, R., Sala, E., Sánchez, L.E. Prediction of cystic evolution of tumours in ovarian cancer with CT-derived features.

Abstract

Background

High Grade Serous Ovarian Carcinoma (HGSOC) presents high levels of macroscopic heterogeneity with cystic/necrotic, solid and calcified tumour regions. This work aims to predict changes in the cystic tumour proportion as a potential indicator of treatment response.

Materials and Methods

CT-based models were developed using shape and first order features computed for pelvic and ovarian lesions in HGSOC for a cohort of 53 neoadjuvant-treated patients.

Results

After feature selection, models rendered high ROC AUCs of 0.81 ± 0.01 for shape features and 0.88 ± 0.08 for the combination with first order features.

Conclusions

Interpretable CT-based radiomics allows for predicting changes in the cystic proportion of ovarian cancer in response to NACT.

Introduction

High Grade Serous Ovarian Carcinoma (HGSOC) is the most prevalent and lethal type of ovarian cancer [1] and presents with high levels of macroscopic heterogeneity with cystic/necrotic, solid, and calcified tumour regions [2]. Solid tumour regions have been reported to generally have higher cellular density and therefore contribute more to adverse prognostic [3], implying that a decrease in solid tumour with a concomitant increase in cystic/necrotic components likely indicates treatment response despite the overall tumour diameter or volume undergoing little change. However, response to treatment is generally measured using RECIST 1.1 criteria [4], which relies upon tumour diameter measurements.

Through the use of computed tomography (CT)-based interpretable radiomic features, this study aims to predict the change in tumour cystic/necrotic proportion under neoadjuvant chemotherapy (NACT) in HGSOC.

Materials and Methods

Patient Cohort

This is a retrospective analysis of a cohort of patients diagnosed with HGSOE who were treated with NACT before delayed primary surgery (DPS). Clinical and imaging data were prospectively collected at Cambridge University Hospitals NHS Foundation trust (Cambridge, UK) between 2009 and 2020 under the CTCR-OV04 observational study. Patients included in this study underwent contrast-enhanced CT scans of the abdomen and pelvis both at baseline (pre-NACT) and follow-up (post-NACT and prior to DPS) and were found to have pelvic and/or ovarian deposits on CT.

Imaging Processing

Lesions in the pelvis and ovaries were manually segmented by a board-certified gynaecological radiologist (RW: 8 years' experience as Consultant Radiologist). Whole tumour segmentations were sub-segmented using a previously developed automated method [5] to extract the cystic/necrotic, solid and calcified regions. The volumetric proportion of each region to the total tumour volume was calculated for each time-point and patients were split into two groups based on their change in cystic proportion between the two time-points (increase versus decrease) (Fig. 1).

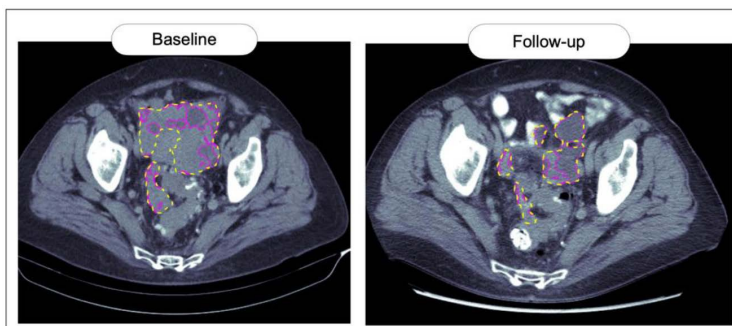


FIGURE 1

Baseline and follow-up CT scans of a patient showing an increase in the cystic proportion (+29%). Whole tumour segmentation is shown in yellow and the cystic region in magenta.

Fourteen shape and 18 first order intensity histogram statistics radiomic features were extracted for each patient at baseline for the whole volume segmentations using the open-source package PyRadiomics version 3.0.1 [6]. Higher order radiomic features were excluded for this study given their limited interpretability and to reduce the curse of dimensionality due to the small sample size. Pre-processing involved re-sampling of the images to the average voxel size ($0.68 \times 0.68 \times 4.71$) mm over the range of values of the cohort using Welch sinc interpolation method to minimise effects of different acquisition and reconstruction parameters [7].

Machine Learning Predictive Modelling

Two Linear Support Vector Classifier (LSVC) models were built to predict an increase or decrease in cystic/necrotic proportion: (A) based on shape features only (shape set) and (B) using a combination of shape and first order features (combined set).

Each feature was linearly scaled to a range (0, 1) to ensure the equal contribution of all features to the model fitting and model learning function even if measured at different scales.

Firstly, a logistic regression model was built for every feature and removed if the prediction accuracy was below 0.5. Secondly, from highly correlated feature pairs (Spearman coefficient $\rho > 0.95$) the feature with the highest predictive power was kept. The final feature selection step was sequential backward feature selection (SBS) [8] based on a stratified 3-fold cross-validation score. SBS was run independently on the shape set and the combined set.

To avoid overfitting, the optimal number of features was selected to ensure a good accuracy of the model for both testing and validation datasets as well as a large area under the receiver operating characteristic (ROC) curve (AUC). The selected features were then used to construct the final LSVC models, using stratified 3-fold cross-validation.

Results

A total of 53 patients were included in this study, 30 (57%) of whom presented an increase in the necrotic/cystic proportion under NACT and a decrease for the remaining 23 (43%).

Two shape and six first order features were removed based on their pair-wise Spearman rank correlation coefficients. Fig. 2 illustrates the cross-validation training and validation accuracy for both feature sets (12 shape features, 12 shape + 12 first order features) by SBS rank. The optimal number of features was set to 3 for the shape set (training accuracy: 0.68 ± 0.07 , validation

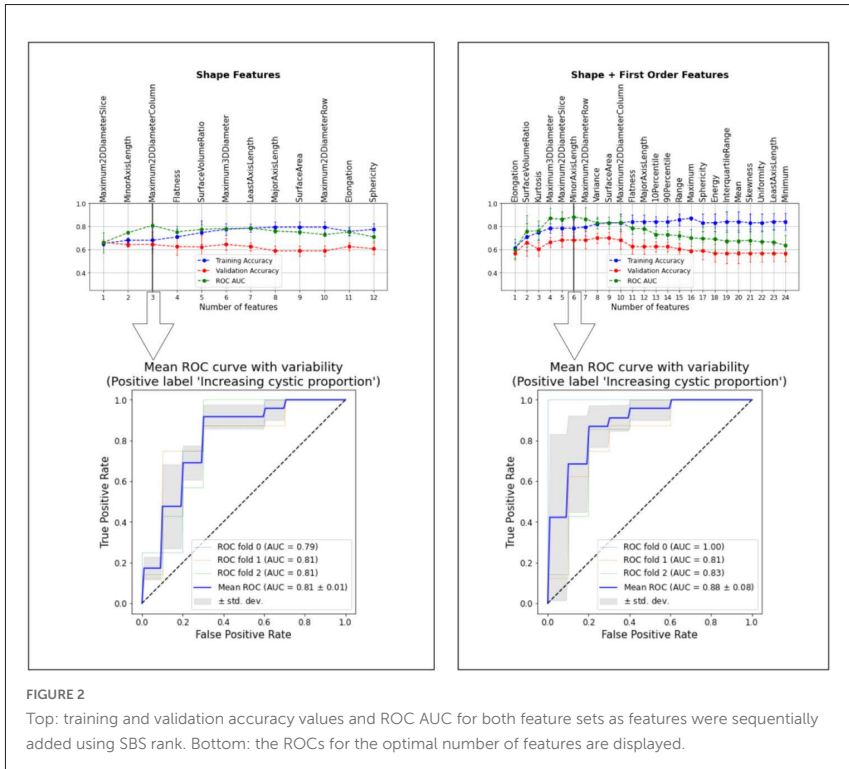


FIGURE 2

Top: training and validation accuracy values and ROC AUC for both feature sets as features were sequentially added using SBS rank. Bottom: the ROCs for the optimal number of features are displayed.

accuracy: 0.64 ± 0.05 , ROC AUC: 0.81 ± 0.01) and to 6 (5 shape, 1 first order) for the combined set (training accuracy: 0.78 ± 0.07 , validation accuracy: 0.68 ± 0.08 , ROC AUC: 0.88 ± 0.08).

Conclusion

This feasibility study demonstrates that interpretable machine learning using shape and first order radiomic features can predict changes in cystic volume proportion for pelvic and ovarian lesions in HGSOc patients under NACT. This is relevant as an increase in the cystic proportion likely represents response to treatment. Further work involves the use of an external test set as well as the evaluation of less conservative, more complex non-linear models with larger datasets.

References

- [1] Jayson, G. C., Kohn, E. C., Kitchener, H.C et al.: Ovarian Cancer. *The Lancet* 384(9951), 1376–1388 (2014).
- [2] Lisio, M. A., Fu, L., Goyeneche, A. et al.: High-grade serous ovarian cancer: basic sciences, clinical and therapeutic standpoints. *International journal of molecular sciences* 20(4), 952 (2019).

[3] Burkill, G. J. C., Allen, S. D., A'hern, R. P. et al.: Significance of tumour calcification in ovarian carcinoma. *The British journal of radiology* 82(980), 640–644 (2009).

[4] Eisenhauer, E. A., Therasse, P., Bogaerts, J. et al.: New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1. 1). *European Journal of Cancer* 45(2), 228–247 (2009)

[5] Rundo, L., Beer, L., Ursprung, S. et al.: Tissue-specific and interpretable sub-segmentation of whole tumour burden on CT images by unsupervised fuzzy clustering. *Computers in Biology and Medicine* 120, 103751 (2016).

[6] van Griethuysen, J. J. M., Fedorov, A., Parmar, C. et al.: Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research* 77(21), e104–107 (2017).

[7] Escudero Sánchez, L., Rundo, L., Gill, A. B. et al.: Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle. *Scientific reports* 11(1), 1–15 (2021).

[8] Ferri, F. J., Pudil, P., Hatef, M., & Kittler, J.: Comparative study of techniques for large-scale feature selection. *Machine intelligence and pattern recognition* 16, 403–413 (1994).

Cross-attention multiple instance learning for interpretable whole slide image classification

Author

Thomas Allcock – School of Computing, University of Leeds, Sir William Henry Bragg Building, Woodhouse Lane, Leeds, UK

Andy Bulpitt – School of Computing, University of Leeds, Sir William Henry Bragg Building, Woodhouse Lane, Leeds, UK

Andrew Hanby – Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK

Rebecca Millican-Slater – Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK

Citation

Allcock, T., Bulpitt, A., Hanby, A., Millican-Slater, R. Cross-attention multiple instance learning for interpretable whole slide image classification.

Abstract

Computerized analysis of histopathology slides using Deep Learning methods has seen success in the diagnosis of cancers. However, the supervised approaches used require large, annotated datasets which are time consuming to produce. To overcome this, weakly-supervised methods were proposed largely based on Multiple Instance Learning. Although these approaches are competitive, they can lack interpretability. Therefore, we propose a prototype-based MIL approach based on a cross-attention transformer framework which allows for end-to-end learning of prototypical patterns and then constructs Whole Slide Image features based on the similarity between the prototypes and the slide.

Introduction

Deep learning based approaches have improved computerized cancer diagnosis [1]. However, Whole Slide Images (WSI) are giga-pixel images and therefore annotated datasets are time-consuming and expensive to produce. Weakly-supervised approaches remedy this by making use of only global slide labels. These approaches are often based on Multiple Instance Learning (MIL) [2]. Attention-based approaches have been proposed to solve the MIL problem [3, 4] and provide insight into the model's decision making through a heatmap of important instances. Prototype-based MIL approaches provide further interpretability [5, 6] by comparing a WSI against a set of prototypes, where each prototype captures a different morphology.

Here we propose a prototype-based MIL approach where prototypes are learned and compared against WSI patches, representing a slide as regions with different tissue structures. An attention mechanism then determines the saliency of each morphological region towards the final prediction. This work aims to improve upon other MIL approaches by understanding the importance of different tissue types towards a classification. Additionally, it differs from other prototype-based MIL approaches by learning prototypes end-to-end instead of through unsupervised clustering (e.g. k -means). This removes the requirement for a prototype discovery step and produces prototypes specific to the classification task. Additionally, the use of cross-attention introduces a learnable similarity function instead of a pre-defined similarity measure, such as euclidean or cosine distance. We validate the proposed approach on lung cancer sub-type prediction and find competitive performance with state-of-the-art MIL approaches.

Method

A WSI x_i is represented as a set of L patches $x_i = \{x_1, \dots, x_L\}$ extracted from non-background regions. Each patch is processed using a ResNet-50 [7] pretrained on ImageNet, reducing each image patch to a feature embedding $h \in R^d$. Each patch feature vector h_{ij} , where J is the index of the patch,

is compared against a set of m prototype vectors $P = \{p_n\}_{n=1}^m$ using cross-attention. Here patch embeddings are mapped to a key K and value V vector using linear layers $W_k \in \mathbb{R}^{d \times d'}$ and $W_v \in \mathbb{R}^{d \times d'}$. The output of the slide when queried against a set of prototypes is,

$$\bar{C} = \text{concat}(C_0, \dots, C_m) \quad (1)$$

where:

$$C_n = \text{softmax}\left(\frac{p_n k^T}{\sqrt{d'}}\right) V \quad (2)$$

The embedding C_n represents the slide information most similar to the n^{th} prototype. Finally, the attention mechanism from [3] aggregates the m embeddings into a single slide embedding by computing an attention coefficient a_n for each feature embedding C_n . The final slide embedding is given by,

$$C_{out} = \sum_{n=1}^m a_n C_n \quad (3)$$

which is passed to a linear classifier producing a score for each class.

Prototype Learning For end-to-end prototype learning two terms are added to the loss function. Alongside cross-entropy for the classification loss, a cluster loss [8] and orthogonality loss [9] are included. The cluster loss encourages each prototype to be the most similar to at least one patch within a WSI. For each patch, the similarity scores between it and each prototype are softmaxed returning m similarity maps $\mathbf{s}_m \in \mathbf{S}$ where $\mathbf{S} = \text{softmax}\left(\frac{PK^T}{\sqrt{d'}}\right)$. Then the maximum similarity score for each prototype

across the slide is taken. This gives m maximum scores which are averaged to return the loss given by,

$$L_{clst} = -\frac{1}{m} \sum_{n=1}^m \max_{s_n \in S} (s_n) \quad (4)$$

The orthogonal loss encourages each prototype to be orthogonal to one another and thus promotes diversity between prototypes,

$$L_{orth} = \|PP^T - I\|^2 \quad (5)$$

where I is the identity matrix and P are the prototypes which are normalized to unit length. The final loss function is therefore,

$$L = L_{CE} + L_{clst} + L_{orth} \quad (6)$$

Experiments and Results

We used data from The Cancer Genome Atlas (TCGA) open-access portal which contains a total of 1053 pathology slides. 541 of the slides contain lung adenocarcinoma (LUAD) and 512 contain lung squamous cell carcinoma (LUSC). 5-fold cross validation was implemented with 80% of data used for training and validation and 20% used for testing. Non-overlapping patches of size 256×256 pixels are extracted at $10 \times$ magnification from all non-background slide regions. The models were trained for a maximum of 30 epochs and the model with the lowest validation f1-score in each fold was selected for testing. 6 prototypes were used and initialized as random vectors. Adam optimizer was used with a fixed learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} . Cross Entropy Loss was used for the slide label loss. Training was implemented with a single NVIDIA V100 GPU on the Advanced Research Computing 4 (ARC4) system at the University of Leeds. Results were compared against several state of the art MIL methods [3, 4, 10].

Results show comparable performance to other state of the art methods as shown in Table 1. From Fig 1 prototypes are discovered for regions such as dense tumor, alveolar, stroma and some tissue edge patches with the highest attention given to dense tumor cells. We found that dense tumor cell regions were given a greater attention on average for LUSC compared to

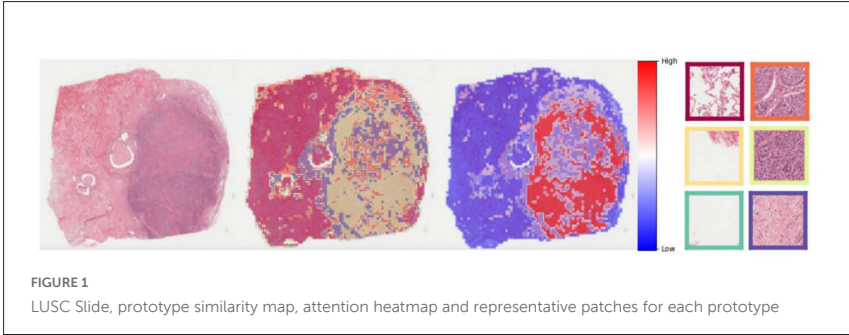


TABLE 1: Mean Test AUC, Accuracy and F1-score across 5-folds \pm standard deviation

Method	AUC	ACC	F1-score
ABMIL	0.922 \pm 0.011	0.869 \pm 0.010	0.869 \pm 0.010
CLAM	0.912 \pm 0.005	0.871 \pm 0.014	0.871 \pm 0.014
TransMIL	0.931 \pm 0.013	0.859 \pm 0.016	0.858 \pm 0.016
Ours	0.942\pm0.006	0.882\pm0.008	0.881\pm0.008

LUAD cases, although still important for LUAD. Alveolar regions had a higher average attention in LUAD cases compared to LUSC where they were found to be unimportant. Stromal regions and regions on the tissue edge had a low attention for both LUAD and LUSC cases.

There are some limitations to the proposed approach. First, each component of the loss function contributes equally to the total loss. Tuning the weighting of each component might produce greater performance. Second, the optimal number of prototypes still requires further investigation. Future work will look to address these limitations and understand the clinical relevance of the discovered prototypes. Further evaluation is also required across other cancer types.

Acknowledgements

This work was supported by The Engineering and Physical Sciences Research Council (EP-SRC) [EP/S024336/1] and Leica Biosystems. The results shown here are based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

References

[1] Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. *Med. Image Anal.* 67, 101813 (2021). <https://doi.org/10.1016/j.media.2020.101813>.

[2] Ghaffari Laleh, N., Muti, H.S., Loeffler, C.M.L., Eche, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., Buelow, R.D., Grabsch, H.I., Brenner, H., Chang-Claude, J., Alwers, E., Brinker, T.J., Khader, F., Truhn, D., Gaisa, N.T., Boor, P., Hoffmeister, M., Schulz, V., Kather, J.N.: Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med. Image Anal.* 79, 102474 (2022). <https://doi.org/10.1016/J.MEDIA.2022.102474>.

- [3] Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: 35th International Conference on Machine Learning, ICML 2018. pp. 3376–3391 (2018).
- [4] Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* 2021 56. 5, 555–570 (2021). <https://doi.org/10.1038/s41551-020-00682-w>.
- [5] Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N.: Handcrafted Histological Transformer (H2T): Unsupervised representation of whole slide images. *Med. Image Anal.* 85, 102743 (2023). <https://doi.org/10.1016/j.media.2023.102743>.
- [6] Yu, J.-G., Wu, Z., Ming, Y., Deng, S., Li, Y., Ou, C., He, C., Wang, B., Zhang, P., Wang, Y.: Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. *Med. Image Anal.* 85, 102748 (2023). <https://doi.org/10.1016/j.media.2023.102748>.
- [7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>.
- [8] Foroughi pour, A., White, B.S., Park, J., Sheridan, T.B., Chuang, J.H.: Deep learning features encode interpretable morphologies within histological images. *Sci. Rep.* 12, (2022). <https://doi.org/10.1038/s41598-022-13541-2>.
- [9] Wang, J., Liu, H., Wang, X., Jing, L.: Interpretable Image Recognition by Constructing Transparent Embedding Space. (2021).
- [10] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. (2021).

Investigating the effect of self-supervised contrastive learning on mitosis classification

Author

Trinh Thi Le Vuong – School of Electrical Engineering, Korea University, Seoul, Korea

Mostafa Jahanifar – Tissue Image Analytics Centre, University of Warwick, Coventry, UK

Neda Zamanitajeddin – Tissue Image Analytics Centre, University of Warwick, Coventry, UK

Jin Tae Kwak – School of Electrical Engineering, Korea University, Seoul, Korea

Nasir Rajpoot – Tissue Image Analytics Centre, University of Warwick, Coventry, UK

Citation

Le Vuong, T.T., Jahanifar, M., Zamanitajeddin, N., Kwak, J.T., Nasir, R. Investigating the effect of self-supervised contrastive learning on mitosis classification.

Abstract

The quality of histology images depends on the tissue type, staining, and digitization procedure, which vary from source to source, causing a domain-shift problem. Despite the great success of deep learning models in computational pathology, sub-optimal generalization on a specific domain still hampers the performance of those models on unseen data due to the domain shift. To overcome this challenge, transfer learning and self-supervised learning have emerged as promising techniques for histology image analysis, especially in scenarios where training data is limited. Self-

supervised learning is usually studied using a large set of training data without requiring them to be labelled or relevant to the downstream task. In this work, we evaluate several self-supervised models under different settings. We train ResNet50 using various transfer and self-supervised learning methods and evaluate their performance on a cross-domain classification of mitosis figure images. The results suggest that the computation pathology image analysis task has its own characteristics; simply applying self-supervised methods requires a large amount of data to succeed. Otherwise, domain knowledge to adapt and design self-supervised methods is needed.

Introduction

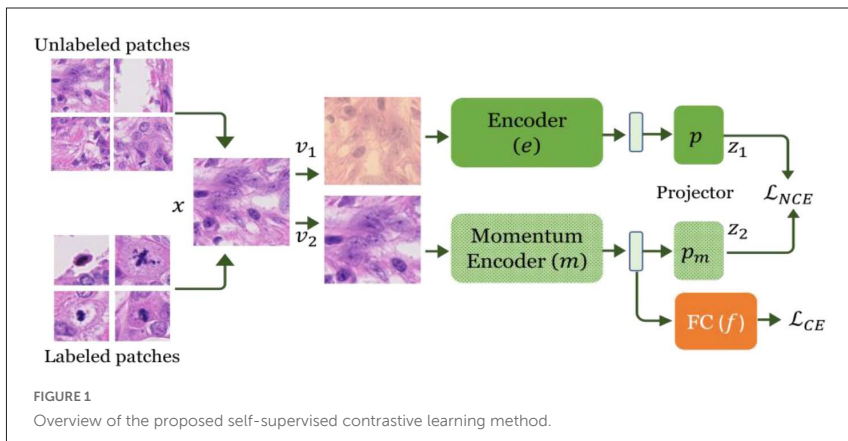
Although deep learning (DL) models are shown to be very powerful in solving various computational pathology (CPath) problems [10], they exhibit vulnerability against the variations in histology images [4]. One of the main challenges in CPath is the domain-shift problem, where the joint distribution of the data and label used for training and testing datasets varies significantly. There are many sources of domain shift in histology images, such as variations in sample preparation and staining protocol, the colour palette of different scanners, and the tissue type itself.

Two of the most studied methods for domain generalisation in CPath are normalisation (SN) [8], stain-augmentation (SA), and domain adaptation [6,7]. Domain adversarial learning has also attracted much attention for domain generalisation in CPath [13]. Other methods use general self-supervised contrastive learning (SSCL) algorithms, such as MoCo-v2 [3], or self-supervised learning (SSL) methods that try to solve modality-specific pretext tasks, like Self-Path [7], to leverage a huge number of unlabelled histology images and pre-train a generalisable model. However, most SSCL methods require a proper design of data augmentation techniques and adjustment of their extent to be trained effectively. This research evaluates various transfer learning and SSCL methods on the mitosis classification task in a series of cross-domain validation experiments.

Method

We evaluate our method using the mitosis classification dataset from MIDOG22 [1,2]. From the training set of the MIDOG22 dataset and provided annotations, we extract a total of 8,337 mitosis patches and 11,051 mimickers (non-mitotic cells which resemble mitosis) patches in four domains: canine lung cancer, canine lymphoma, canine cutaneous mast cell tumour, and human neuroendocrine tumour. Similarly, 2,714 mitoses and 1,721 mimicker patches from human breast cancer samples were extracted to serve as our test set. We trained all SSCL models using two versions of the MIDOG22 dataset: 1) **Labeled patches**: 19,388 patches of size 128×128 pixels and 2) **Unlabeled patches**: 172,434 patches of size 224×224 pixels from the training set.

We evaluate three SSCL methods, including MoCo-v2 [3], InfoMin [11], and IMPaSh [12], in three different settings: 1) Labeled patches: the SSCL encoder is trained using label patches, 2) ImageNet + Labelled Patches: the SSCL encoder is trained using labelled patches but initialised with ImageNet pre-trained weights, and 3) Unlabeled patches: the SSCL encoder is trained using unlabeled patches.



The three SSCL methods have some variations but are still based on the momentum contrastive learning framework, which is shown in Fig. 1. In this framework, an encoder e with learnable parameters θ_e is utilised to extract feature representations from augmented views of the input image. To enhance the optimisation of the contrastive loss by leveraging a larger number of negative samples in each batch, the framework incorporates the *momentum encoder* m whose parameters are momentum updates of the θ_e i.e., $\theta_m \leftarrow \alpha\theta_m + (1 - \alpha)\theta_e$.

Given an image x , we generate two different augmented views v_1 and v_2 . The loss function is a variation of InfoNCE loss [9], which maximises the mutual information between positive representation pairs (v_1, v_2) while minimising the similarity of those views with other K negative representations from the momentum encoder:

$$\mathcal{L}_{NCE} = -\mathbb{E} \left[\log \frac{\exp(z_{1,i} \cdot z_{2,i}/\tau)}{\sum_{j=1}^K \exp(z_{1,i} \cdot z_{2,i}/\tau)} \right] \#(1)$$

where $z_1 = p(e(v_1))$, $z_2 = p_m(m(v_2))$, and $\tau = 0.07$ is the temperature hyper-parameter. We use ResNet50 [5] as the feature extractor e . After pretraining e , we add a classifier f on top of its output feature representations and fine-tune the network using cross-entropy loss (\mathcal{L}_{CE}) and our desired labelled dataset.

Results and Discussion

Tab. 1 reports the performance of ResNet50 trained with different methods on mitosis classification. Fine-tuning the ResNet50 pre-trained on ImageNet achieves good results on the validation set by improving 14.0% Acc from scratch. Interestingly, stain normalisation (ImageNet-SN) degrades the performance to 73.0%. Using the proposed SSCL approach, superior results are achieved with MoCo-v2, InfoMin, or IMPaSh trained on a large amount of unlabeled data. It is natural to think that supervised fine-tuning benefits more if the SSCL training data appeared similar to the supervised data. However, performance improvement is not observed with a small SSCL training set

TABLE 1: Results of the mitosis classification using various transfer learning and SSCL techniques

Method	SSCL pre-trained	Acc (%)	F1	Recall	Spec	κ
Scratch	-	68.2	0.626	0.627	0.674	0.275
ImageNet-SN	-	73.0	0.707	0.703	0.717	0.417
ImageNet	-	82.2	0.806	0.797	0.824	0.613
MoCo-v2	Labelled patches	67.3	0.586	0.601	0.685	0.226
InfoMin	Labelled patches	69.3	0.637	0.637	0.689	0.297
IMPash	Labelled patches	67.0	0.588	0.600	0.674	0.224
MoCo-v2	ImageNet + Labelled Patches	85.9	0.849	0.844	0.856	0.699
InfoMin	ImageNet + Labelled Patches	85.6	0.848	0.846	0.849	0.695
IMPash	ImageNet + Labelled Patches	85.6	0.848	0.847	0.85	0.697
MoCo-v2	Unlabeled Patches	85.3	0.843	0.839	0.85	0.687
InfoMin	Unlabeled Patches	85.2	0.842	0.839	0.847	0.685
IMPash	Unlabeled Patches	85.1	0.843	0.841	0.845	0.686

(11,051 patches) compared to training from scratch. Nonetheless, the SSCL using the labelled set but pre-trained on ImageNet still achieves the highest results.

In summary, we present an empirical evaluation of self-supervised learning methods using labelled and unlabeled data to improve the performance of mitosis figure classification. In future work, we would evaluate the SSCL on mitosis semantic segmentation to construct an end-to-end mitosis detection framework.

Acknowledgement

This work was funded by the Medical Research Council (MRC) UK and South Korea biomedical and health researcher exchange scheme grant No. MC/PC/210-14, and KIAT grant funded by the Korea Government (MOTIE) (P0022543) and National Research Foundation of Korea (NRF) grant funded by the Korea Government (No. 2021K1A3A1A88100920).

References

- [1] Aubreville, M., Bertram, C., Breininger, K., Jabari, S., Stathonikos, N., Veta, M.: Mitosis domain generalisation challenge 2022. In: 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022) (2022)
- [2] Aubreville, M., Stathonikos, N., Bertram, C.A., Klopffleisch, R., Ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., et al.: Mitosis domain generalisation in histopathology images-the midog challenge. *Medical Image Analysis* 84, 102699 (2023)
- [3] Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
- [4] Foote, A., Asif, A., Rajpoot, N., Minhas, F.: Reet: Robustness evaluation and enhancement toolbox for computational pathology. *Bioinformatics* (Oxford, England) p. btac315

- [5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
- [6] Jahanifar, M., Shephard, A., Zamanitajeddin, N., Bashir, S., Bilal, M., Khurram, S., Mihas, F., Rajpoot, N.: Stain-robust mitotic figure detection for the mitosis domain generalisation challenge. In: Aubreville, M., Zimmerer, D., Heinrich, M. (eds.) Mitosis domain generalisation in histopathology images–The MIDOG challenge, chap. MIDOG, pp. 48–52. pringer (2022)
- [7] Koohbanani, N.A., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N.: Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging* 40(10), 2845–2856 (2021)
- [8] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalising histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. pp. 1107–1110. IEEE (2009)
- [9] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018) 9. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. *Medical Image Analysis* 67, 101813 (2021)
- [10] Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. *Medical Image Analysis* 67, 101813 (2021)
- [11] Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems* 33, 6827–6839 (2020)
- [12] Vuong, T.T.L., Vu, Q.D., Jahanifar, M., Graham, S., Kwak, J.T., Rajpoot, N.: Impash: A novel domain-shift resistant representation for colorectal cancer tissue classification. In: *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. pp. 543–555. Springer (2023)

[13] Wilm, F., Marzahl, C., Breininger, K., Aubreville, M.: Domain adversarial retinanet as a reference algorithm for the mitosis domain generalisation challenge. In: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis: MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27–October 1, 2021, Proceedings, pp. 5–13. Springer (2022)

Investigation of topological features of the 3D cellular nuclear envelope as observed with electron microscopy

Author

Kokeb Dese, Cefa Karabağ, Panos Giannopoulos, Constantino Carlos Reyes–Aldasoro – giCentre, Department of Computer Science, City, University of London, London EC1V 0HB, UK

Citation

Dese, K., Karabağ, C., Giannopoulos, P., Reyes–Aldasoro, C.C. Investigation of topological features of the 3D cellular nuclear envelope as observed with electron microscopy.

Introduction

The use of topological data analysis (TDA) in medical and biomedical applications has been growing significantly [1], [2] and has provided interesting results in oncology among other areas [3], [4]. TDA aims to capture the essential features of complex shapes by analysing their topological structure [5]. In the context of TDA, scales can be understood as different levels of detail, i.e., distance thresholds at which initial data points (in a metric space) are filtered to create a topological object. This is different from the more traditional concept of scale, e.g., the zoom of a microscope. As such, TDA can provide complementary tools to analyse complex shapes and datasets that have been previously segmented in three dimensions. These topological features, along with persistent homology, have been

proven to be invariant features of data and are utilized in the segmentation and classification of various diseases, as mentioned in [1], [2], [6]. In this work, we explore a series of topological features extracted from volumetric synthetic datasets and nuclear envelopes of HeLa cells that were observed with electron microscopy. We conjecture that the application of TDA can help characterise and analyse very complex shapes such as those presented here.

Materials and Methods

Volumetric Datasets

Two types of volumetric datasets were analysed in this work. First, a series of simple synthetic shapes: *sphere*, *torus*, *double torus*, *triple torus*, which were used to illustrate the metrics extracted and ease the comparison with the HeLa cells. Second, the *nuclear envelope* of HeLa cells as observed with serial block face scanning electron microscopy. The synthetic shapes were generated in Matlab. The HeLa cells are publicly available in Empiar [7] as a 8,000x8,000x517 voxel dataset. The nuclear envelope of individual cells was segmented as previously described [8] and the 3D surface of the nuclear envelope processed as described below.

TDA Analysis

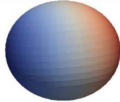



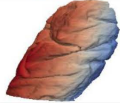
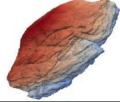
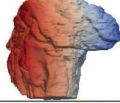
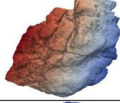
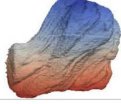
The 3D surface of the nuclear envelopes was exported from Matlab as in Standard Tessellation Language (STL) format. The stl file was imported and processed using the Topology ToolKit (<https://topology-tool-kit.github.io/>), embedded within Paraview (<https://www.paraview.org>) from which the following geometric and topological features of each cell were extracted: number of vertices, edges, faces and Euler characteristics as a geometric features and *Betti* numbers β_0 , and β_1 as a topological features (Table 1). Formally, β_0 , β_1 are defined as the ranks of the 0th, and 1st homology groups of a surface. Informally, β_0 measures the number of connected components, whilst β_1 , measures the maximum number of simple (i.e., not self-intersecting), closed curves (sometimes called loops) on the surface, along which one can cut without separating the surface into more than one components. One can think of such curves as ones that are not the boundary

of any part of the surface. For example, the torus has two such curves while the double torus has four. Note also that for a connected, orientable and closed (i.e, without boundary) surface, we have that $\beta_1 = 2g$, where g is the so called *genus* of the surface. Informally, the genus is the number of 'handles' attached to the surface, e.g., a torus is a sphere with one handles and a double torus is a sphere with two handles; each of these handles correspond to a 'doughnut' hole of the torus.

Results and Discussion

Table 1 illustrates the geometric and topological features of four synthetic shapes and five nuclear envelopes of HeLa cells. The geometric features are directly related with the complexity of the shapes. Lower values of vertices, edges and faces indicate a simpler and smoother the surface. A significant increase can be noticed from the synthetic shapes to the nuclear envelopes. Similarly, the Euler characteristic, which is related to the number of "holes" of a shape, presents a major difference between the two groups. Whilst it is not possible to observe at the resolution presented, we conjecture that the low values of the Euler characteristic are due to very small discontinuities more related to the roughness of the surface rather than holes that would connect two opposite side of the surface of the nuclei. A similar trend, but with positive values follows β_1 . An interesting case is presented by the values of β_0 , which do not follow the complexity as indicated by the geometric features or the Euler characteristic. All surfaces with the exception of rows 7 and 9 have $\beta_0 = 1$. Further work could try to elucidate why these surfaces have the same values, when clearly they are very different. Whilst this is just a preliminary investigation of TDA in biomedical imaging, the geometric and topological features may prove useful in the analysis of biomedical imaging. Future research could focus on the use of machine learning approaches that incorporate topological features to segment and classify three dimensional datasets with different characteristics, i.e., different cell stages, treatments or diseases.

TABLE 1: Geometric (Number of vertices, edges and Faces) and Topological Features (Euler characteristic, Betti numbers 0, 1) of Synthetic 3D volumes and Segmented HeLa Nuclear envelopes

	Object	Vertices	Edges	Faces	Euler	β_0	β_1
1		1,250	3,744	2,496	2	1	0
2		5,872	17,942	12,070	0	1	2
3		7,726	23,184	15,456	-2	1	4
4		11,892	35,688	23,792	-4	1	6
5		142,814	428,550	285,700	-36	1	38
6		114,126	342,498	228,332	-40	1	42
7		155,864	467,786	311,860	-62	8	80
8		122,200	366,905	244,604	-101	1	103
9		189,270	568,294	378,868	-156	10	183

References

- [1] Y. Singh *et al.*, "Topological data analysis in medical imaging: current state of the art," *Insights Imaging*, vol. 14, no. 1, p. 58, Apr. 2023, doi: 10.1186/s13244-023-01413-w.
- [2] Y. Skaf and R. Laubenbacher, "Topological data analysis in biomedicine: A review," *J. Biomed. Inform.*, vol. 130, p. 104082, Jun. 2022, doi: 10.1016/j.jbi.2022.104082.
- [3] A. Bukkuri, N. Andor, and I. K. Darcy, "Applications of Topological Data Analysis in Oncology," *Front. Artif. Intell.*, vol. 4, p. 659037, Apr. 2021, doi: 10.3389/frai.2021.659037.
- [4] J. Yu and X. Chang, "Topological Data Analysis: A New Method to Identify Genetic Alterations in Cancer," *Asia-Pac. J. Oncol. Nurs.*, vol. 8, no. 2, pp. 112–114, Mar. 2021, doi: 10.4103/2347-5625.308301.

[5] C.-S. Hu, A. Lawson, J.-S. Chen, Y.-M. Chung, C. Smyth, and S.-M. Yang, "TopoResNet: A Hybrid Deep Learning Architecture and Its Application to Skin Lesion Classification," *Mathematics*, vol. 9, no. 22, p. 2924, Nov. 2021, doi: 10.3390/math9222924.

[6] N. Byrne, J. R. Clough, I. Valverde, G. Montana, and A. P. King, "A Persistent Homology-Based Topological Loss for CNN-Based Multiclass Segmentation of CMR," *IEEE Trans. Med. Imaging*, vol. 42, no. 1, pp. 3–14, Jan. 2023, doi: 10.1109/TMI.2022.3203309.

[7] A. Iudin, P. K. Korir, J. Salavert-Torres, G. J. Kleywegt, and A. Patwardhan, "EMPIAR: a public archive for raw electron microscopy image data," *Nat. Methods*, vol. 13, no. 5, pp. 387–388, May 2016, doi: 10.1038/nmeth.3806.

[8] C. Karabağ, M. L. Jones, and C. C. Reyes-Aldasoro, "Volumetric Semantic Instance Segmentation of the Plasma Membrane of HeLa Cells," *J. Imaging*, vol. 7, no. 6, p. 93, Jun. 2021, doi: 10.3390/jimaging7060093.

On generalisability of segment anything model for nuclear instance segmentation in histology images

Author

Kesi Xu – Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK

Lea Goetz – Artificial Intelligence and Machine Learning, GSK, London, UK

Nasir Rajpoot – Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK

Citation

Xu, K., Goetz, L., Rajpoot, N. On generalisability of segment anything model for nuclear instance segmentation in histology images.

Abstract

Pre-trained on a large and diverse dataset, the segment anything model (SAM) is the first promptable foundation model in computer vision aiming at object segmentation tasks. In this work, we evaluate SAM for the task of nuclear instance segmentation performance with zero-shot learning and finetuning. We compare SAM with other representative methods in nuclear instance segmentation, especially in the context of model generalisability. To achieve automatic nuclear instance segmentation, we propose using a nuclei detection model to provide bounding boxes or central points of nuclei as visual prompts for SAM in generating nuclear instance masks from histology images.

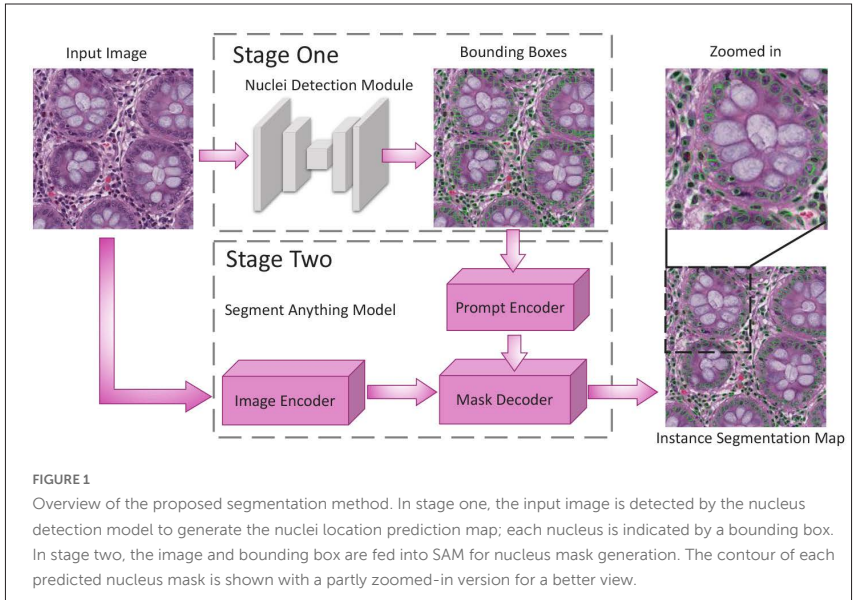
Introduction

In Computational Pathology (CPath), generating a nuclear segmentation mask from digital histology images is vital as it can be used in downstream analysis, such as cancer grading, tumour microenvironment analysis, survival analysis, etc [1–4]... The challenge lies in accurate nuclear segmentation, which is essential for understanding each tissue component's contribution to disease. Current works focused on accurately segmenting overlapping and cluttered nuclei [1, 2]. However, the segmentation performance of machine learning (ML) models often does not generalise across different datasets of domains. Yet, the model's robustness and generalisability are essential requirements for clinical applications. The recently released Segment Anything Model (SAM)[5] is trained on the SA-1B dataset, which contains an unprecedented number of images and annotations. This allows the model to exhibit strong zero-shot generalisation for segmentation tasks. SAM uses an image encoder and prompt encoder, both based on a vision transformer framework, to incorporate user interactions and embed prompts. The extracted features from two encoders are merged in a lightweight mask decoder to generate segmentation results.

In this paper, we evaluate the generalisability of SAM on a nuclear instance segmentation task. As SAM relies on a visual prompt for segmentation, to make a fair comparison, we choose to compare SAM with another state-of-the-art (SOTA) semi-automatic nuclear instance segmentation method – NuClick [2]. NuClick has a similar interactive mechanism as SAM and requires a click inside the designated nuclear object as a visual prompt for nuclear instance mask generation. We also compare the proposed method with a SOTA-supervised learning method [6] in nuclear instance segmentation on the Lizard dataset [7].

Method

We proposed a two-stage method by adding a nucleus detection stage with SAM for nuclear instance segmentation, as shown in Fig. 1. For an input image, we use a nucleus detection model, which is a fine-tuned YOLOv8 [8], to provide bounding boxes of nuclei. The second stage is



nuclear segmentation with SAM. The centre points of the detected nuclei bounding boxes serve as the visual prompts for the SAM prompt encoder. By aggregating the outputs of both the image encoder and prompt encoder, the mask decoder generates the final instance map.

Experiment and Result

Dataset and Experiment Setting

We assessed nuclear segmentation under domain shift using the Lizard dataset [7], the largest publicly available colon tissue nuclei dataset. It includes images from six centres: GlaS [9], CRAG [10], CoNSEP [1], DigestPath, PanNuke [11], and TCGA [12]. We used the first five datasets as training data for the models in Table 2, while the TCGA dataset was the

unseen test data to evaluate the model's domain generalisation. We use the following metrics: Dice score evaluates the semantic segmentation of the nucleus versus background class and considers all instances as a single object. Binary-class panoptic quality (PQ), equal to the detection quality score (DQ) multiplied by the segmentation quality score (SQ), is used to evaluate the performance of nuclei instance segmentation. To finetune SAM model, we freeze the image encoder and prompt encoder. We only finetuned the mask decoder of SAM on the Lizard training dataset, with the nuclear central point prompt as input.

Result

Providing the finetuned SAM with ground truth central points as prompt inputs gives a 2% and 2.2% improvement in averaged PQ score and Dice score, respectively, compared with NuClick (Table 1). While providing the default SAM with ground truth bounding boxes as prompt achieves the best results, drawing bounding boxes as opposed to a single point would be prohibitively time-consuming in clinical practice and impractical requires.

Generalisability

For a fair comparison, we used YOLOv8 for nuclear detection, then used the central point of nuclear as the point prompt for finetuned SAM for nuclear instance segmentation. Table 2 shows that finetuned SAM has better domain generalisability than HoVer-Net, outperforming HoVer-Net by 3.3% in the PQ score. See an example visualised segmentation result in Fig. 2.

TABLE 1: Interactive nuclear segmentation method domain generalizability evaluation

Ground truth prompt type	Segmentation method	Dice	PQ	DQ	SQ
Points	NuClick	0.796	0.663	0.858	0.744
	SAM	0.572	0.339	0.450	0.775
	Finetuned SAM	0.812	0.678	0.872	0.768
Bounding boxes	SAM	0.835	0.703	0.913	0.768

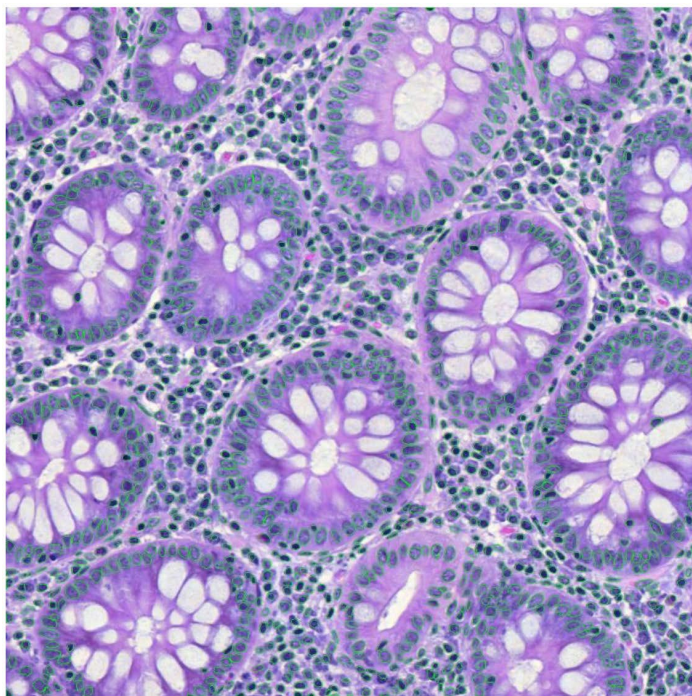


FIGURE 2

The visualisation examples of the nuclear segmentation result of the proposed method.

TABLE 2: Cross-validation external test on TCGA coherent in Lizard dataset

Segmentation Method	Dice	PQ	DQ	SQ
U-Net [13]	0.612	0.390	0.588	0.664
Micro-Net [3]	0.735	0.484	0.654	0.741
HoVer-Net [1]	0.801	0.514	0.656	0.780
YOLOv8+Finetuned SAM	0.745	0.569	0.729	0.778

Conclusion

We have evaluated the domain generalisability of the SAM with and without finetuning the mask decoder. The SAM demonstrates good generalisability in the nuclear segmentation when provided with a ground truth bounding box prompt in zero-shot learning. On a more clinically relevant task, the finetuned SAM using the nuclear central point as prompt, shows better generalisability than HoVer-Net on an external test dataset. We conclude that SAM has the potential to become a foundation model in CPath due to its good generalisability.

References

- [1] Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* 58, 101563 (2019). <https://doi.org/10.1016/j.media.2019.101563>.
- [2] Alemi Koohbanani, N., Jahanifar, M., Zamani Tajadin, N., Rajpoot, N.: NuClick: A deep learning framework for interactive segmentation of microscopic images. *Med. Image Anal.* 65, 101771 (2020). <https://doi.org/10.1016/j.media.2020.101771>.
- [3] 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. *Med. Image Anal.* 52, 160–173 (2019). <https://doi.org/10.1016/j.media.2018.12.003>.

- [4] Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M.: Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans. Med. Imaging.* 35, 1196–1206 (2016). <https://doi.org/10.1109/TMI.2016.2525803>.
- [5] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment Anything, <http://arxiv.org/abs/2304.02643>, (2023).
- [6] Xu, K., Jahanifar, M., Graham, S., Rajpoot, N.: Accurate segmentation of nuclear instances using a double-stage neural network. In: *Medical Imaging 2023: Digital and Computational Pathology*. pp. 506–515. SPIE (2023). <https://doi.org/10.1117/12.2654173>.
- [7] Graham, S., Jahanifar, M., Azam, A., Nimir, M., Tsang, Y.-W., Dodd, K., Hero, E., Sahota, H., Tank, A., Benes, K., Wahab, N., Minhas, F., Raza, S.E.A., El Daly, H., Gopalakrishnan, K., Snead, D., Rajpoot, N.: Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification. In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. pp. 684–693. IEEE, Montreal, BC, Canada (2021). <https://doi.org/10.1109/ICCVW54120.2021.00082>.
- [8] Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics, <https://github.com/ultralytics/ultralytics>, (2023).
- [9] Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R.J., Rajpoot, N.M.: Gland segmentation in colon histology images: The glas challenge contest. *Med. Image Anal.* 35, 489–502 (2017). <https://doi.org/10.1016/j.media.2016.08.008>.
- [10] Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.-A., Snead, D., Tsang, Y.W., Rajpoot, N.: MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. Image Anal.* 52, 199–211 (2019). <https://doi.org/10.1016/j.media.2018.12.001>.

[11] Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N.: PanNuke Dataset Extension, Insights and Baselines, <http://arxiv.org/abs/2003.10778>, (2020). <https://doi.org/10.48550/arXiv.2003.10778>.

[12] Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., Staudt, L.M.: Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* 375, 1109–1112 (2016). <https://doi.org/10.1056/NEJMp1607591>.

[13] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., and Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. pp. 234–241. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28.

On enhancing the robustness of vision transformers in medical imaging: Defensive diffusion

Author

Raza Imam, Muhammad Huzaifa, Mohammed El–Amine Azz – Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Citation

Imam, R., Huzaifa, M., Azz, M.E. On enhancing the robustness of vision transformers in medical imaging: Defensive diffusion.

Abstract

Privacy and confidentiality of medical data are of utmost importance in healthcare settings. ViTs, the SOTA vision model, rely on large amounts of patient data for training, which raises concerns about data security and the potential for unauthorized access. Adversaries may exploit vulnerabilities in ViTs to extract sensitive patient information and compromising patient privacy. This work address these vulnerabilities to ensure the trustworthiness and reliability of ViTs in medical applications. In this work, we introduced a defensive diffusion technique as an adversarial purifier to eliminate adversarial noise introduced by attackers in the original image. By utilizing the denoising capabilities of the diffusion model, we employ a reverse diffusion process to effectively eliminate the adversarial noise from the attack sample, resulting in a cleaner image that is then fed into the ViT blocks. Our findings demonstrate the effectiveness of the diffusion model in eliminating attack-agnostic adversarial noise from images. Additionally, we propose combining knowledge distillation with our framework to obtain a lightweight

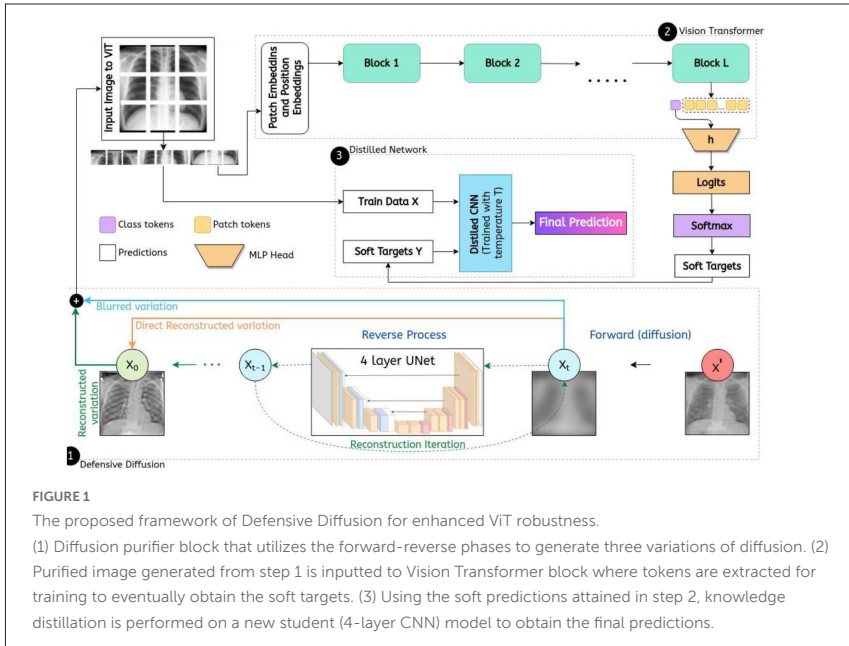
student model that is both computationally efficient and robust against gray box attacks. Comparison of our method with a SOTA baseline method, SEViT, shows that our work is able to outperform the baseline. Extensive experiments conducted on a publicly available Tuberculosis X-ray dataset validate the computational efficiency and improved robustness achieved by our proposed architecture. Our code is publicly available at https://github.com/Muhammad-Huzaifaa/Defensive_Diffusion.

Introduction

The privacy and confidentiality of medical data are crucial in healthcare settings. Vision Transformers (ViTs), being state-of-the-art vision models, raise concerns about data security and unauthorized access due to their reliance on large patient datasets for training. The objective of this work is to improve both computational efficiency and the overall robustness of medical imaging systems using vision transformers. We aim to demonstrate the efficacy of the diffusion model in eliminating attack-agnostic adversarial noise from images. We hypothesize that (1) reverse diffusion process is effective in removing adversarial noise, (2) adversarial images generated specifically for the vision transformer would not be transferable to the student model (CNN). Faris et al. [1] conducted a study revealing that Vision Transformers (ViTs) lack robustness against white box attacks. In our research, we address this issue from a different perspective by leveraging the capabilities of a diffusion model to mitigate attack-agnostic noise. By exploiting the denoising properties of the diffusion model, we employ a reverse diffusion process to remove adversarial noise and obtain a clean image. Combining knowledge distillation with our framework to obtain a lightweight student model that is both computationally efficient and robust against gray box attacks [2].

Methodology

Defensive Diffusion In our approach, we begin by training the deblurring diffusion model in an end-to-end manner [3], using the training set from the TB-dataset [4]. We then apply this trained model to process adversarial images generated using a vision transformer. The forward diffusion



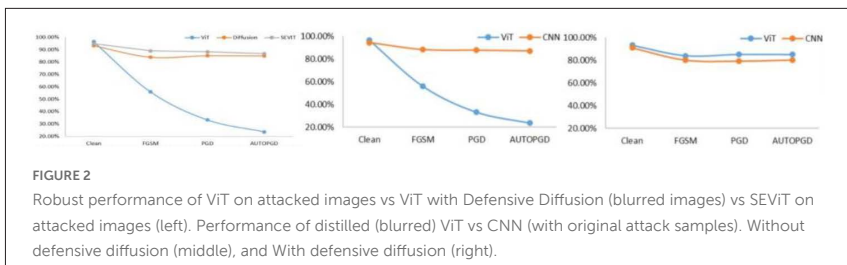
process is employed to introduce blur to the images, effectively mixing the adversarial noise with the blurred noise [5]. The resulting blurred images are subsequently fed into the deblurring diffusion model to obtain denoised images as shown in Figure 1(1). To evaluate the effectiveness of our method, we compare the quality of the blurred and reconstructed images. Unlike the approach proposed in [6], which focuses on adversarial purification, our method leverages the deblurring diffusion model. Here we consider three variation of diffusion model phases- blurred, reconstruction, and direct-reconstruction phases. In the 'blurred phase' we evaluate the final output obtained from forward diffusion process [5]. In the 'algorithmic reconstruction' phase, we employ the method described in [3]. This approach involves reconstructing the image by combining it with the image from the

previous time step, followed by the addition of variable noise to the image until time t . As the time steps progress, the amount of added noise gradually decreases. To facilitate 'direct reconstruction', we leverage the trained U-net model, which allows us to reconstruct the clean samples in a single step [7].

Knowledge Distillation Knowledge distillation trains a teacher model to generate soft targets, used to train a smaller student model [8]. The student model improves robustness by introducing uncertainties in its output, making it harder for attackers to generate deceiving adversarial examples. Soft targets provide valuable information that can't be conveyed through a single hard target, enabling better generalization. CNNs are used due to the transferability of adversarial attacks within the model family [9]. Training with soft probabilities avoids severe overfitting observed with hard probabilities [8], introducing smoothness to the optimization landscape. The lightweight CNN is trained on soft probabilities, taking input from the last ViT layer (Figure 1(3)).

Results and Discussions

Results on Diffusion with Raw ViT vs Diffusion ViT vs SEViT We evaluated our proposed diffusion method on ViT and SEViT models (Figure 2). Our objective as defenders was to maintain high robust accuracy while defending against attacks. With the original ViT, clean accuracy dropped to around 55% when using attack samples. However, when applying our diffusion purifier, all three diffusion variations showed higher robust accuracy. The



blurred variation performed the best, achieving approximately 85% robust accuracy against PGD attacks. SEViT exhibited higher robust accuracy without diffusion, around 89% against FGSM attacks (Table 1). However, applying diffusion decreased robust accuracy due to over-smoothing. The blurred variation of diffusion showed similar robustness to SEViT, with a 1-2% difference, making it an efficient and practical solution compared to the complex ensemble of MLPs in SEViT.

Results on Teacher ViT vs Student CNN We compared the performance of the 5 convolution-layer student CNN model and the teacher ViT model on

TABLE 1: Defensive diffusion when performed with ViT (a) and SEViT (b) across Clean & Attack (FGSM [10], PGD [11], AutoPGD [12]) samples. Here, the column 'Original' are attack samples without any diffusion phase, whereas the other three variations (blurred, reconstructed and direct reconstructed) are samples generated when input attack samples are passed through the diffusion phase. In SEViT, we are ensembling 3 MLPs. Clean & Robust performance on distilled CNN across the same combination (c)

	Attack Type	Original	Reconst.	Blurred	DirRec.
(a) Diffusion with ViT	Clean	96.37%	94.34%	93.18%	95.21%
	FGSM	55.85%	65.63%	83.85%	66.22%
	PGD	33.18%	60.89%	85.03%	60.89%
	AUTOPGD	23.56%	58.81%	84.88%	57.63%
(b) Diffusion with SEViT	Clean	94.78%	93.04%	91.30%	92.46%
	FGSM	89.03%	87.70%	83.55%	87.55%
	PGD	88.29%	87.55%	82.96%	87.11%
	AUTOPGD	86.67%	87.11%	83.25%	86.96%
(c) Distilled CNN	Clean	94.27%	91.40%	90.83%	93.27%
	FGSM	88.23%	82.56%	79.80%	85.90%
	PGD	87.65%	81.98%	78.92%	86.05%
	AUTOPGD	87.06%	82.70%	79.94%	86.05%

adversarial images and images obtained through defensive diffusion (Figure 2). The ViT model experienced a significant decrease in accuracy, dropping to around 20% under white box attacks. In contrast, the CNN model maintained

a high accuracy of over 80% even under adversarial attack (Table 1). The CNN model demonstrated superior robustness and resilience compared to the ViT model. Surprisingly, the accuracy of the ViT model was effectively restored when evaluated on blurred images generated by our method. The CNN model maintained its accuracy levels, indicating its inherent robustness. These results validate the effectiveness of our proposed defensive diffusion technique and the resilience of the CNN model against adversarial attacks.

Conclusion and Future Work

In conclusion, our defensive diffusion method effectively eliminates adversarial noise and outperforms the state-of-the-art SEViT in most attack scenarios.

Future work includes experimenting with different diffusion models, exploring partial reconstruction, employing an ensemble-based approach, and conducting tests on diverse natural images. These efforts aim to enhance the robustness and generalizability of our approach, further improving its performance and applicability in real-world scenarios.

References

- [1] Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar. Self-ensembling vision transformer (sevit) for robust medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 376–386. Springer, 2022.
- [2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. in *2017 IEEE Symp. on Security and Privacy (SP)*. *IEEE*. (10.1109/SP.2017.49), pages 39–57, 2017.
- [3] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022.

- [4] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.
- [5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [6] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.
- [7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
- [9] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021.
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020.

Stain-invariant representation for tissue classification in histology images

Author

Manahil Raza, Saad Bashir, Talha Qaiser, Nasir Rajpoot – Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K.

Citation

Raza, M., Bashir, S., Qaiser, T., Rajpoot, N., Stain-invariant representation for tissue classification in histology images.

Abstract

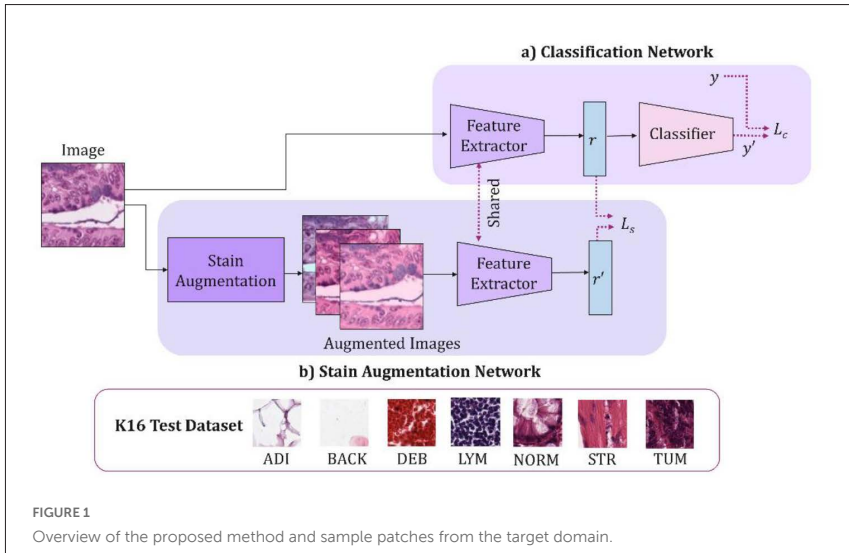
The process of digitizing histology slides involves multiple factors that can affect a whole slide image's (WSI) final appearance, including the staining protocol, scanner, and tissue type. This variability constitutes a domain shift and results in significant problems when training and testing deep learning (DL) algorithms in multi-cohort settings. As such, developing robust and generalizable DL models in computational pathology (CPath) remains an open challenge. In this regard, we propose a framework that generates stain-augmented versions of the training images using stain matrix perturbation. Thereafter, we employed a stain regularization loss to enforce consistency between the feature representations of the source and augmented images. Doing so encourages the model to learn stain-invariant and consequently, domain-invariant feature representations. We evaluated the performance of the proposed model on cross-domain multi-class tissue type classification of colorectal cancer images and have achieved improved performance compared to other state-of-the-art methods.

Introduction

The advent of Deep Learning (DL) has revolutionised the field of Computational Pathology (CPath) and has enabled the automated and quantitative analysis of histology images [1,2]. Despite the success of DL methods, they are vulnerable to domain-specific variations [3]. Some major sources of variations include staining and scanning processes, where distinct institutions may employ different staining protocols and use various scanners, resulting in differences in the visual appearance of whole slide images (WSIs). This variability poses a significant challenge known as domain shift for training robust DL models and encumbers their ability to generalise well across diverse histology datasets. Addressing the domain shift problem has been a focal point for CPath researchers, leading to several efforts in stain normalisation [4-6], augmentation and adaptation. Stain augmentation (SA) aims to mitigate the effects of domain shift by generating augmented variations of the source images to mimic the stain variations present in the target domain for improving the model's generalisability on unseen data [7-9]. Tellez *et al.* [24] has stressed upon the importance of using stain augmentations for histopathology images for a more robust classification performance. Abbet *et al.* [11] proposed a novel domain adaptation (DA) method, Self-Rule to Multi-Adapt (SRMA), for single-source and multi-source tissue classification with multiple datasets by using in-domain and cross-domain losses. Unlike in DA, domain generalisation (DG) methods cannot leverage unlabelled data from the target domain [10]. To this effect, Vuong *et al.* [12] adopted a self-supervised contrastive learning approach using a combination of encoders and momentum encoders for colorectal cancer classification using patch shuffling augmentations. Our proposed method, inspired by [14] uses stain augmentations to help extract domain-invariant feature representations for colorectal cancer tissue images, thus ensuring that the class labels assigned to an image remain consistent in the face of staining variability.

Methodology

The proposed framework comprises of two modules, one for classification and the other for stain augmentation, as shown in Fig. 1. Each image x



with label y is passed through ResNet-18 based feature extractor f_e for extracting the feature representation as $f_e(x) = r$. This feature representation is then passed onto an MLP-based classifier f_c for a classification decision, \hat{y} as $f_c(f_e(x)) = \hat{y}$. During the training process, we also employ the stain augmentation network, which generates stain-altered version(s) of the source image as $x' = \{x'_1..x'_N\}$. For this purpose, we use the Vahadane [18] method for extracting the stain matrix. The stain concentrations are perturbed to create the stain-altered images using [15]. These images are then passed onto the same feature extractor $f_e(x') = r' = \{r'_1..r'_N\}$ to extract feature representations of the stain-altered images. Two loss functions are employed in the proposed workflow. We use the cross-entropy loss as the primary classification loss L_c between the predicted label \hat{y} and the ground-truth label y . Additionally, we employed a mean squared error (MSE) loss as a stain regularisation loss, $L_s = ||r - r'||_2^2$, which measures the distance between the extracted feature representations of the source and augmented images. The MSE loss acts as a strong penalisation factor, enforcing

consistency in the face of stain augmentations. The overall loss function combines the two loss functions, $L = L_c + L_s$.

Results and Discussions

We have employed two datasets to validate the proposed framework 1). Kather-19 (K19) [16], which contains 100,000 images (224x224 pixels) from 9 tissue classes and 2). Kather-16 (K16) [17], which consists of 5,000 images (150x150 pixels) from 8 classes. Since there are discrepancies between the class labels of the two datasets, we followed the strategy for relabelling [11] and grouped the data into seven classes, namely adipose, background, debris, lymphocytes, normal colon mucosa, stroma and colorectal adenocarcinoma epithelium. Sample images from K16 are shown in Fig.1. The results of the experiments are reported in Table 1, where ImageNet Upper Bound denotes an experiment where both the training and testing are performed with the same dataset (K16). The degradation in performance in ImageNet Lower bound is due to the presence of a domain shift when the model is trained and tested on different source (K19) and target domains (K16). We observe that the proposed method which generated 6 augmented images for each input image, outperforms the ImageNet baseline by 22% in terms of accuracy. Whereas it performed 20% and 12% better as compared to MocoV2 [22] and InfoMin [23]. IMPaSh [12] is a combination of InfoMin [23] and PatchShuffling augmentations, and our methods still outperform it by 1% while being less computationally expensive. To summarise, the proposed workflow leveraged stain augmentations to encourage the DL model to learn stain and domain-invariant feature representations and outperformed other state-of-the-art methods which begs the question "Is Stain Augmentation really all you need for Domain Generalisation?" Our future work will include automating the selection of the optimal number of augmentations.

TABLE 1: Experimental Results between source (K19) and target (K16) domains

Method	Training Dataset	Accuracy	Recall	Precision	F1-Score
ImageNet - Upper Bound	K16 (Target)	0.942	0.942	0.941	0.941
ImageNet - Lower Bound	K19	0.654	0.654	0.741	0.626
SN Macenko [19]	K19	0.660	0.660	0.683	0.645
SN Vahadane [18]	K19	0.683	0.683	0.696	0.656
InsDis [20]	K19	0.694	0.694	0.766	0.659
PIRL [21]	K19	0.818	0.818	0.853	0.812
MocoV2 [22]	K19	0.675	0.675	0.816	0.642
InfoMin [23]	K19	0.750	0.750	0.824	0.752
IMPaSh [12]	K19	0.868	0.868	0.887	0.865
Proposed Method	K19	0.878	0.878	0.887	0.877

References

- [1] Wu, Y., Cheng, M., Huang, S., Pei, Z., Zuo, Y., Liu, J., Yang, K., Zhu, Q., Zhang, J., Hong, H. and Zhang, D., 2022. Recent advances of deep learning for computational histopathology: Principles and applications. *Cancers*, 14(5), p.1199.
- [2] Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T. and Kather, J.N., 2021. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4), pp.686-696.
- [3] Stacke, K., Eilertsen, G., Unger, J. and Lundström, C., 2020. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2), pp.325-336.

[4] Shaban, M.T., Baur, C., Navab, N. and Albarqouni, S., 2019, April. Staingan: Stain style transfer for digital histological images. In 2019 IEEE 16th international symposium on biomedical imaging (Isbi 2019) (pp. 953-956). IEEE.

[5] Jia, Q., Guo, J., Du, F., Yang, P. and Yang, Y., 2022, December. A Fast Texture-to-Stain Adversarial Stain Normalization Network for Histopathological Images. In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2294-2301). IEEE.

[6] Zhao, B., Han, C., Pan, X., Lin, J., Yi, Z., Liang, C., Chen, X., Li, B., Qiu, W., Li, D. and Liang, L., 2022. RestainNet: a self-supervised digital re-stainer for stain normalization. *Computers and Electrical Engineering*, 103, p.108304.

[7] Jahanifar, M., Shepard, A., Zamanitajeddin, N., Bashir, R.S., Bilal, M., Khurram, S.A., Minhas, F. and Rajpoot, N., 2022. Stain-robust mitotic figure detection for the mitosis domain generalization challenge. In *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis: MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27–October 1, 2021, Proceedings* (pp. 48-52). Cham: Springer International Publishing.

[8] Yamashita, R., Long, J., Banda, S., Shen, J. and Rubin, D.L., 2021. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40(12), pp.3945-3954.

[9] Chang, J.R., Wu, M.S., Yu, W.H., Chen, C.C., Yang, C.K., Lin, Y.Y. and Yeh, C.Y., 2021. Stain mix-up: Unsupervised domain generalization for histopathology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24* (pp. 117-126). Springer International Publishing.

[10] Ghifary, M., Balduzzi, D., Kleijn, W.B. and Zhang, M., 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7), pp.1414- 1430.

[11] Abbet, C., Studer, L., Fischer, A., Dawson, H., Zlobec, I., Bozorgtabar, B. and Thiran, J.P., 2022. Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection. *Medical image analysis*, 79, p.102473.

[12] Vuong, T.T.L., Vu, Q.D., Jahanifar, M., Graham, S., Kwak, J.T. and Rajpoot, N., 2023, February. IMPaSh: A Novel Domain-Shift Resistant Representation for Colorectal Cancer Tissue Classification. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* (pp. 543-555). Cham: Springer Nature Switzerland.

[13] Raipuria, G., Shrivastava, A. and Singhal, N., 2022, September. Stain-AgLR: Stain Agnostic Learning for Computational Histopathology Using Domain Consistency and Stain Regeneration Loss. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings* (pp. 33-44). Cham: Springer Nature Switzerland

[14] Pakzad, A., Abhishek, K. and Hamarneh, G., 2023, February. CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV* (pp. 203-219). Cham: Springer Nature Switzerland.

[15] Pocock, J., Graham, S., Vu, Q.D., Jahanifar, M., Deshpande, S., Hadjigeorgiou, G., Shephard, A., Saad Bashir, R.M., Bilal, M., Lu, W. and Epstein, D., 2021. TIAToolbox: an end-to-end toolbox for advanced tissue image analytics. *bioRxiv*, pp.2021-12

- [16] Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D. and Jansen, L., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1), p.e1002730.
- [17] Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A. and Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1), pp.1-11.
- [18] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I. and Navab, N., 2016. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8), pp.1962-1971.
- [19] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C. and Thomas, N.E., 2009, June. A method for normalizing histology slides for quantitative analysis. In 2009 IEEE international symposium on biomedical imaging: from nano to macro (pp. 1107-1110). IEEE.
- [20] Wu, Z., Xiong, Y., Yu, S.X. and Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3733-3742).
- [21] Misra, I. and Maaten, L.V.D., 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6707-6717).
- [22] Chen, X., Fan, H., Girshick, R. and He, K., 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- [23] Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C. and Isola, P., 2020. What makes for good views for contrastive learning?. *Advances in neural information processing systems*, 33, pp.6827-6839.

[24] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F. and Van Der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58, p.101544.

Radiology report generation using multi-layer visual representation

Author

Chenyu Wang, Stephen McKenna, Vladimir Janjic – School of Science & Engineering University of Dundee, Dundee, DD1 4HN, UK

Citation

Wang, C., McKenna, S., Janjic, V. Radiology report generation using multi-layer visual representation.

Abstract

Chest X-ray images are crucial for diagnostics and treatment of various diseases, but analysing them requires highly skilled and experienced medical professionals. This leads to a high burden on radiologists. This paper investigates a method for automatic generation of radiology reports and incorporates visual features extracted from multiple layers of a convolutional visual encoder network. The use of multi-level features can provide more information to the text generator, potentially improving generated reports. Experimental results are reported on the MIMIC-CXR dataset using natural language generation metrics.

Introduction

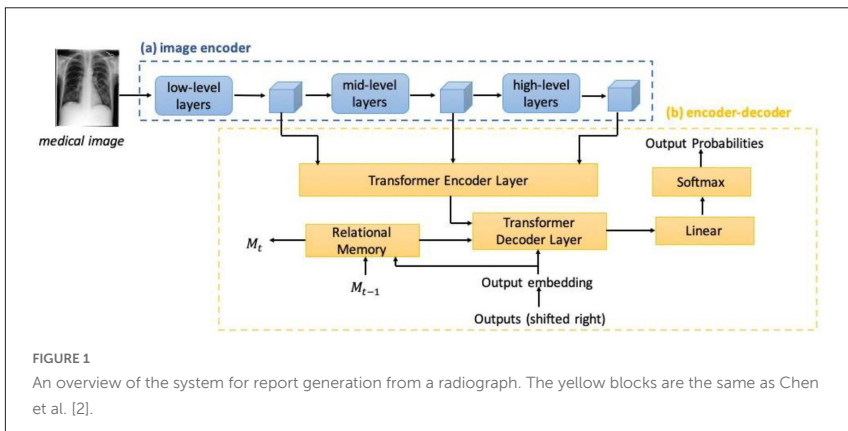
In radiology practice, radiologists are required to undertake a time-consuming process of reading chest X-ray images and writing reports to narrate their findings. Most work on automatic generation of radiology reports (e.g. [2,6–8,10]) applies conventional encoder-decoder frameworks,

starting with a standard convolutional neural network (CNN) as the encoder to extract visual features from radiology images, and subsequently feeding the extracted features to a recurrent neural network for radiology report generation.

We propose a modification to the method of Chen et al. [2]. They proposed a memory-driven transformer which incorporates a relational memory and a memory-driven conditional layer normalization. To provide visual input to this report generator they used a ResNet101 (pretrained on ImageNet) to extract 2,048-dimensional feature vectors from each patch in a grid of image patches. Instead, we apply a single ResNet101 network to the entire radiograph, and form the visual representation by concatenating vectors obtained at multiple layers of the network. We evaluate the resulting report generation method using the MIMIC-CXR dataset, adopting the commonly used natural language generation (NLG) metrics BLEU1-4 [9], METEOR [3] and ROUGE-L [4]) to evaluate report quality.

Method

Figure 1 presents a schematic overview of the system. The blue blocks illustrate the multi-level visual encoder; the yellow blocks are the same



as in the system proposed by Chen et al. [2]. Training aims to produce text sequences that align to the target text representation, given training radiographs.

Convolutional visual encoders learn representations that are hierarchical and compositional. Features extracted from a single convolutional layer may not capture the relevant visual information across various granularities. Therefore, we extract visual features from multiple convolutional layers to construct a built-in feature pyramid [5]. Specifically, we concatenate the feature vectors obtained from the last three residual blocks of a ResNet101 encoder. These are denoted as x_s^L , x_s^M and x_s^H for low, mid, and high-level representations, respectively. To encode the extracted visual features, and then decode the encoded content, we follow the transformer designed by Chen et al. [2] for radiology report generation.

Experiment and Results

Following the experimental setting from Chen et al. [2] we excluded samples without reports and adopted the official dataset split for MIMIC-CXR (222,758 for training, 1,808 for validation, and 3,269 for testing). Training took approximately 8 hours per epoch running on a computer with an Intel(R) Core(TM) i9-9820X CPU and one GeForce RTX 2080 GPU. The system was trained for 20 epochs.

Table 1 compares the proposed method with recent state-of-the-art work on the same dataset. The method performed competitively, obtaining the best BLEU metrics. It was slightly better on all metrics than Chen et al. [2], suggesting that the change of visual encoder was helpful.

Figure 2 shows an example of a chest X-ray image and associated report from MIMIC-CXR, as well as a report generated automatically using the proposed system. Aligned medical observations are highlighted in different colours. In this instance, the generated report covers most of the observations present in the ground-truth report.

TABLE 1: Comparison of the proposed method with previous studies on MIMIC-CXR test data. The symbol † indicates results that are directly cited from the referenced paper, while the symbol ‡ represents results obtained by running released code

Model	NLG Metrics					
	BLEU1	BLEU2	BLEU3	BLEU4	MEREOR	ROUGE-L
R2GEN [2] ‡	0.355	0.217	0.145	0.103	0.140	0.273
R2GEN [2] †	0.353	0.218	0.145	0.103	0.142	0.277
TOPDOWN [1] †	0.317	0.195	0.130	0.092	0.128	0.267
PPKED [7] †	0.360	0.224	0.149	0.106	0.149	0.284
CMCL[6] †	0.344	0.217	0.140	0.097	0.133	0.281
OURS	0.362	0.225	0.150	0.108	0.147	0.278

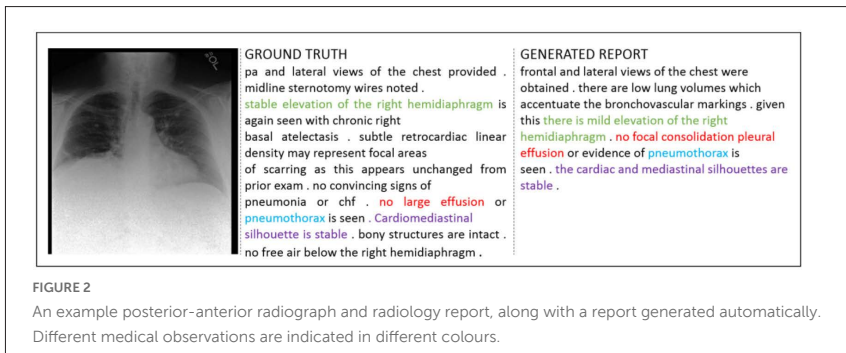


FIGURE 2
 An example posterior-anterior radiograph and radiology report, along with a report generated automatically. Different medical observations are indicated in different colours.

Discussion and Conclusion

We modified a previously proposed system for radiology report generation [2] by replacing the patch-wise visual encoders with a single multilevel image encoder. Evaluation on the MIMIC-CXR dataset using NLG metrics suggests that this modification is beneficial. However, further experiments are needed to assess significance.

There remains a sizeable gap between the performance of state-of-the-art radiology report generation systems and the performance needed for reliable deployment. Generating radiology reports is challenging for reasons which include the need to generate multi-sentence text using specialist language, and the need to ensure that abnormalities are highlighted in reports given datasets in which normal regions dominate. In future work, we will use clinically-focused metrics in addition to NLG metrics for evaluation.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proc. IEEE CVRP 2018, pp 6077–6086, 2018.
- [2] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Proc. EMNLP 2020, pages 1439–1449, 2020.
- [3] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proc. Sixth Workshop on Statistical Machine Translation, pages 85–91, 2011.

[4] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proc. ACL-02 Workshop on Automatic Summarization*, volume 4, pages 45–51, 2002.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. IEEE CVRP 2017*, pp. 2117–2125, 2017.

[6] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In *Proc. ACL IJCNLP 2021 (Volume 1: Long Papers)*, pp. 3001–3012, Online, 2021.

[7] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proc. IEEE CVRP 2021*, pp. 13753–13762, 2021.

[8] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 269–280, 2021.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[10] Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. Crossmodal contrastive attention model for medical report generation. In *Proc. COLING 2022*, pp. 2388–2397, 2022.

Multi-lesion segmentation for diabetic retinopathy

Author

Mohammed Ali Athar, Kashif Rajpoot – University of Birmingham

Citation

Athar, M.A., Rajpoot, K. Multi-lesion segmentation for diabetic retinopathy.

Abstract

Medical image segmentation plays a crucial role in the diagnosis and treatment of many diseases, including diabetic retinopathy (DR). DR is a leading cause of blindness and involves abnormalities in the blood vessels of the retina. Our contributions include identifying binary class map weighting as a solution for more accurate segmentation of the microaneurysm class. This study provides insights into improving the accuracy of DR lesion segmentation, which will ultimately aid in the diagnosis and treatment of DR.

Introduction

Globally, approximately 95 million diabetic patients have diabetic retinopathy (DR), of which one-third have vision-threatening DR¹. DR is a complication of diabetes that affects the retina; if left undiagnosed and untreated, it can cause total blindness.

Identification of DR in a patient's eyes requires the assessment of microvascular abnormalities (called lesions) present in the patient's eye. This manual detection by experts is time-consuming and subjective. Furthermore,

¹Diabetic Retinopathy

the global shortage of ophthalmologists results in millions of undiagnosed patients and permanent blindness.

The prevalence of eye fundus image datasets, lesion segmentation annotation maps, and DR disease grades have led to automated solutions through computer vision – primarily via deep learning methods. Among these methods, the U-Net model has emerged as a famous architecture due to its effectiveness in medical imaging problems where the data is limited and there is a high degree of class imbalance.

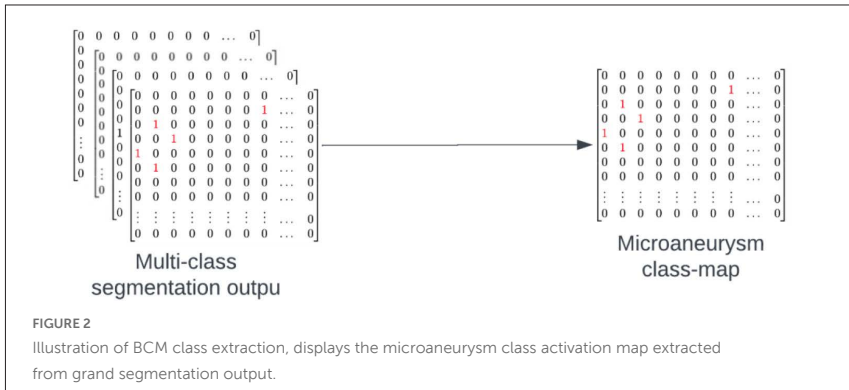
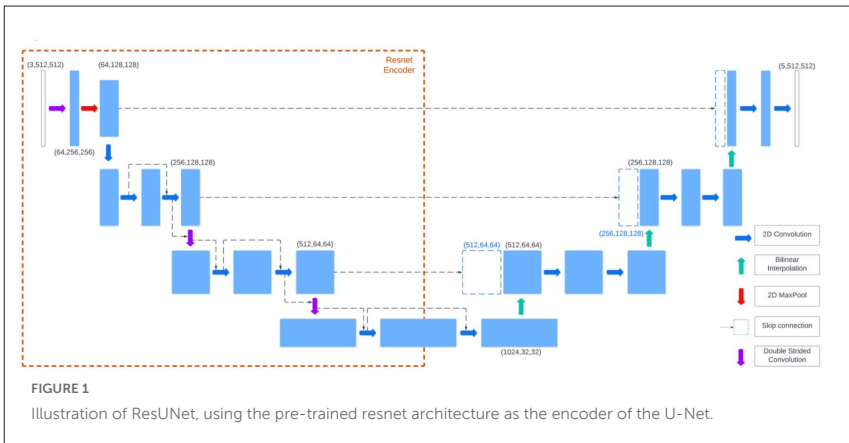
In recent years there have been many advancements in medical image segmentation using UNets due to various modifications proposed in network architecture, such as: using pre-trained encoders, attention mechanisms, and multi-task training. The evolution of loss functions for segmentation has also contributed to U-Nets' success in image segmentation.

Despite advances in various fields, segmentation of lesions for DR remains a difficult problem due to the difficult characteristics of lesions present in the fundus, which is exacerbated by limited dataset sizes. In particular, the microaneurysm class has been known to be more difficult to segment than other classes. Through this work we aim to introduce a novel yet simple training approach which can improve the performance of the model on difficult classes – Binary Classmap Weighting.

Motivation for this work stems from the urgent need for accurate and efficient automated segmentation of DR lesions in fundus images. The prevalence of DR and its potential for blindness in millions of patients worldwide highlights the critical importance of developing effective computer vision solutions for its diagnosis and treatment. Through this research, we aim to introduce a training mechanism which will improve the performance of the model on weakly performing classes. The potential success of this work can have significant implications for the diagnosis and treatment of DR, potentially improving the lives of diabetic patients. Additionally, the identified approaches and techniques may also benefit the wider image segmentation community by providing insights into effective segmentation methods in limited data scenarios.

Methods

The U-Net model we train consists of a pretrained encoder (VGG or ResNet-50) and a simple decoder consisting of bilinear interpolation blocks. In particular, the ResUNet is a variant of the U-Net architecture that uses the pretrained ResNet network as the encoder as shown in Figure 1.



This allows for the use of strided convolutions instead of max-pooling for down-sampling, reducing the loss of local features and improving performance. Other pretrained networks, such as VGG and EfficientNet, can also be used as the encoder of the U-Net, leveraging their powerful feature extraction capabilities to enhance segmentation performance. By incorporating pretrained networks into the U-Net architecture, we can improve the performance of medical image segmentation tasks.

Binary Class Map Weighting

Microaneurysms are small and subtle lesions that are challenging to detect in DR images. In order to address this issue, we have proposed the use of binary BCM² weighting. This technique allows the model to focus more on the difficult class that it is performing poorly on. BCM weighting computes a binary segmentation loss on a selected class from the multi-class segmentation output, which helps the model pay more attention to the specific class of lesions that are harder to detect, such as microaneurysms. The process is depicted in Figure 2.

$$loss_{total} = loss_{seg} + \gamma * loss_{bin}$$

The loss defined above combines the multi-label segmentation loss computed on the full segmentation output with the binary segmentation loss computed on the selected binary class map extracted from the full mask, where γ is a parameter that controls the contribution of the binary segmentation loss.

Results

In this section we present results of our experiments which demonstrate the effectiveness of our proposed training method. We train both a VGG-U-Net and the ResUNet from Figure 1 in two methods, the first using weighted crossentropy, and the second using our proposed BCM training approach. Our method has shown to greatly improve the performance on the microaneurysm class.

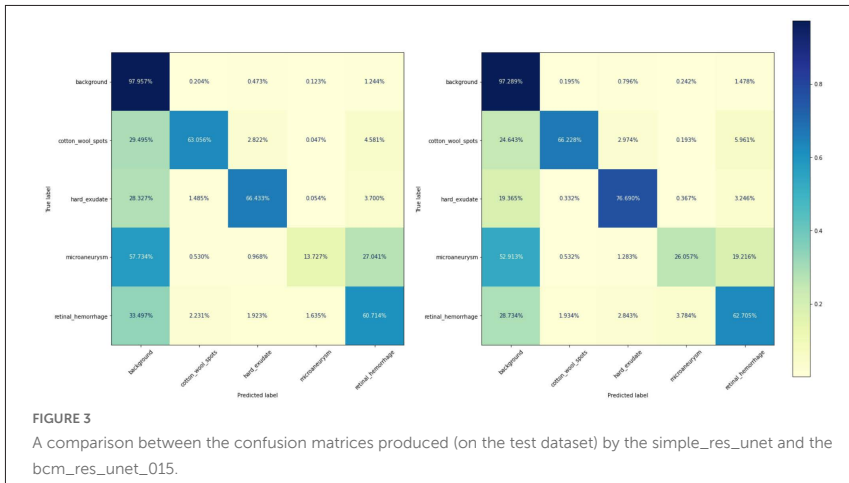
² Binary Class Map

TABLE 1: This table lists the comparison of the F1 score for each lesion class between U-Net models trained with different approaches

Model	cotton_wool_spots	hard_exudate	microaneurysm	retinal_hemorrhage
simple_vgg_unet	0.223	0.606	0.235	0.447
bcm_vgg_unet_05	0.208	0.586	0.308	0.461
bcm_vgg_unet_015	0.349	0.573	0.327	0.451
simple_res_unet	0.358	0.625	0.208	0.476
bcm_res_unet_05	0.356	0.614	0.306	0.457
bcm_res_unet_015	0.394	0.606	0.319	0.469

Furthermore, the models trained with $\gamma = 0.5$ (bcm_vgg_unet_05 and bcm_res_unet_05) showed an increase in performance for microaneurysm, but a decline in performance for other classes such as cotton_wool_spots and hard_exudates as shown in Table 1. This suggests that a higher gamma value causes the binary segmentation loss to conflict with multi-class segmentation loss. On the other hand, the models trained with $\gamma = 0.15$ showed improved performance for microaneurysm without significant deterioration in performance for other classes.

Figure 3 further illustrates the superior performance of the BCM trained U-Net on the microaneurysm class compared to the same model trained using the standard approach. The simple_res_unet classifies the microaneurysm as the retinal hemorrhage more often than the model trained using BCM. Additionally, an increase in performance over all other lesions can be observed for the bcm_re.



The prevalence and association of coronary artery calcification in patients with chronic obstructive pulmonary disease: A systematic review and meta-analysis

Author

Khalid Hakami – Division of Systems Medicine, School of Medicine, University of Dundee

Mohammad Alghamdi – Division of Systems Medicine, School of Medicine, University of Dundee

Abdulmalik Arab – Division of Systems Medicine, School of Medicine, University of Dundee

James Chalmers – Molecular and Clinical Medicine, School of Medicine, University of Dundee

Faisal Khan – Division of Systems Medicine, School of Medicine, University of Dundee

Citation

Hakami, K., Alghamdi, M., Arab, A., Chalmers, J., Khan, F. The prevalence and association of coronary artery calcification in patients with chronic obstructive pulmonary disease: A systematic review and meta-analysis.

Background

The prevalence of coronary artery calcification (CAC) is higher in chronic obstructive pulmonary disease (COPD) patients. The prevalence has not been assessed in a systematic review and meta-analysis, so we aimed to conduct

a systematic review and meta-analysis to quantitatively synthesise data from studies regarding coronary calcification in COPD patients and its risk factors.

Methods

Articles containing keywords such as "COPD," "CAC," and "prevalence" were searched in PubMed, CINAHL, and Web of Science databases. Eligibility screening, data extraction, and quality assessment of retrieved articles were conducted by two reviewers independently. To determine CAC prevalence as well as risk factors among COPD patients, meta-analyses were conducted. To examine the sources of heterogeneity, meta-regressions were conducted.

Results

Across 24 studies, the pooled prevalence was 52% (95% CI 0.36-0.77). It was found that Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages were positively associated with CAC prevalence ($r = 14.93$, $p = 0.01$), while scoring method ($r = 1.9288$, $p = 0.8589$), study design ($r = 3.1946$, $p = 0.0739$), imaging protocol ($r = 1.9831$, $p = 0.9917$), sample size ($r = 1.4777$, $p = 0.4777$) were positively correlated but not significantly. A higher mortality risk was found in patients with COPD compared to those without (HR 1.24; 95%CI 1.11–1.37).

Conclusions

The pooled prevalence of CAC is high, thus patients with COPD should be screened for CAC and risk factors that contribute to it. A larger clinical trial is still needed due to the high degree of heterogeneity.

Where scientists empower society

Creating solutions for healthy lives on a healthy planet

frontiersin.org

Why publish with us?

Open access

All Frontiers journals are fully open access, meaning every research article we publish is immediately and permanently free to read.

Peer review

Our collaborative peer review is handled by experts in the field who objectively certify the quality, validity, and rigor of research.

Technology

With the latest custom-built technology and artificial intelligence, we're revolutionizing the way research is published, evaluated, and communicated.

Impact

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, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact