CLINICCAI DETAILED PROGRAM

CLINICCAI2024 ORAL PRESENTATIONS

Time	Paper Title	Primary Contact Author Name
08:30-09:30	**Oral session 1: Neuroscience** **Chairs: Sandrine De Ribaupierre & Saad Slimani**	
08:30-08:42	Augmented Reality-guided External Ventricular Drain Placement: A Technical Note on a Clinically Fully Integrable System	Jesse van Doormaal
08:42-08:54	Introducing ClinIQ: A Non-Invasive Clinical Tool for Molecular Profiling of Brain Gliomas	Joseph Maldjian
08:54-09:06	Determining cerebral white matter integrity and volume of brain structures in patients with primary Sjögren's syndrome in DTI and FLAIR with AI tools.	Michał Sobański
09:06-09:18	Generating Computed Tomography Angiography from Time-of-Flight Magnetic Resonance Angiography Using Diffusion-Based Models	Dietmar Frey
09:06-09:18	AI-based Outcome Prediction in Ischemic Stroke: A Study on Robust External Validation Principles	Dietmar Frey
09:18-09:30	Deep Learning-Enabled Meningitis Screening In Young Infants Based on a Novel Non-Invasive Transfontanellar Device: Initial Performance Results	Sara Ajanovic Andelic
11:15-12:30	**Oral session 2: Radiology & Pathology** **Chairs: Joe Yeong & Saad Slimani**	
11:15-11:27	Preoperative Rotator Cuff Tear Prediction from Shoulder Radiographs using a Convolutional Block Attention Module-Integrated Neural Network	Ikbeom Jang
11:27-11:39	Fully Automated PACS-Integrated Volumetric Assessment of Abdominal Aortic Aneurysms	David Weiss
11:39-11:51	Artificial Intelligence for Prediction of Abdominal Aortic Aneurysm Status Using Multimodal Patient Data in the VASCULAID-RETRO-AAA Study	Lotte Rijken
11:51-12:03	Machine learning-based prediction of spinal cord ischemia after complex endovascular aneurysm repair using clinical and imaging-derived features	Kaj Kappe
12:03-12:15	H&E-based cell prediction multi-classification models to capture morphologically distinct subpopulations of CD8+ T cells	Muzammil Arif Din Abdul Jabbar
12:15-12:27	From Concept to Clinic: Improving Colorectal Cancer Diagnosis through Deep Learning-Driven Lymph Node Screening in Clinical Routine	Amjad Khan
14:30-15:30	**Oral session 3: Surgery** **Chairs: Abdourahmane Ndong & Idriss Ahmedou**	
14:30-14:42	Deep learning for intraoperative navigation in minimally invasive liver surgery: a multicenter external validation	Namkee Oh
14:42-14:54	Validation of instance segmentation of instruments during laparoscopic cholecystectomy using MedSAM	António S Soares
14:54-15:06	Advancing Laparoscopic Surgery: Addressing Adoption Challenges with User-Centric Augmented Reality Solutions	Pooja P Jain
15:06-15:18	Bridging innovation and practice: the journey of FAROS from technical design to in-vivo animal validation	Nicola A Cavalcanti
15:18-15:30	Deep Learning Artificial Intelligence Tool for Automated Radiographic Determination of Posterior Tibial Slope in Patients with Anterior Cruciate Ligament Injuries	Yining Lu
16:00-17:00	**Oral session 4: Dermatology, fetal Ultrasound, Opthalmology** **Chairs: Joël Lavanchy & Joe Yeong**	
16:00-16:12	Facial Wrinkle Segmentation with Weakly Supervised Pretraining and Supervised Finetuning with Multi-Annotator Labels	Ik Jun Moon
16:12-16:24	Preeclampsia screening by minimally trained operators on ultrasound using Artificial Intelligence, a paired diagnostic pilot study	Ikbeom Jang
16:24-16:36	A Randomized Controlled Trial of Artificial Intelligence to Assist in Screening Fetal Ultrasound Scans	Saad Slimani
16:36-16:48	EyeLiner: Longitudinal retinal fundus image registration through clinically guided keypoint detection	Thomas Day
16:48-17:00	Automated deep learning-based quantification of intermediate age-related macular degeneration features for the prediction of geographic atrophy and neovascularization	Stephen M McNamara
17:00 -17:50	CLINICCAI panel: How technology enables access to health-care in the global south	

Augmented Reality-guided External Ventricular Drain Placement: A Technical Note on a Clinically Fully Integrable System

Authors:

J.A.M. van Doormaal MD[1], Tim Fick MD[2], J.W. Meulstee MSc PhD[2], E.W. Hoving MD PhD[2], P.A.J.T. Robe MD PhD[1], T.P.C. van Doormaal MD PhD[2,3]

Affiliations:

[1] Department of Neurosurgery, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

[2] Department of Neuro-oncology, Princess Máxima Centre for Pediatric Oncology, Heidelberglaan 25, 3584 CS Utrecht, The Netherlands

[3] Department of Neurosurgery, University Hospital of Zürich, Rämistrasse 100, 8091 Zürich, Switzerland

Presenting author:

Name: J.A.M. van Doormaal MD

Affiliation: Department of Neurosurgery, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

E-mail address: j.a.m.vandoormaal-2@umcutrecht.nl

Phone number: +31619975015

Keywords: Augmented Reality, External Ventricular Drain, Image-Guided Neurosurgery

Key information:

1. Research question: Is integration of an Augmented Reality system for guiding External Ventricular Drain placement technically feasible?
2. Findings: Augmented Reality guided External Ventricular Drain placement demonstrated technical feasibility, achieving optimal placement in three out of an initial series of four simulated procedures. The mean workflow time was 22:45 mm:ss (SD ± 11:38 mm:ss) and mean Fiducial Registration Accuracy was 4.1 mm (SD ± 1.7 mm).

3. Meaning: Integrating an augmented reality system for External Ventricular Drain shows potential to improve the accuracy of the procedure while not substantially increasing procedural cost, time and complexity.

MANUSCRIPT

Introduction

External ventricular drain (EVD) placement, a common neurosurgical procedure for managing increased intracranial pressure, traditionally relies on a freehand approach guided by anatomical landmarks[1]. Suboptimal drain placement occurs in 26% of cases, leading to a higher likelihood of EVD dysfunction and reoperation[2]. Repeated attempts at cannulation likely increase the risk of hemorrhage and infection. Although image-guided methods reduce the rate of suboptimal placement under 10%[2], such methods are limited by their cost and technical complexity. Augmented Reality (AR) technology facilitates real-time interaction with virtual overlays superimposed onto the real patient, providing significant potential for assisting EVD procedures. Previously described AR systems show feasibility, but are not readily integrable in clinical workflows[3–6].

For this study, we co-developed an EVD guidance system for AR Head-Mounted Displays (AR-HMDs) that is readily compatible with existing hospital infrastructure. We assessed the feasibility of the system through several AR-assisted EVD placements in anatomical phantoms.

Material and methods

We developed an AR-EVD application incorporated as a module within our previously described medical AR, segmentation and cloud platform (Lumi, Augmedit, Naarden, The Netherlands)[7,8]. The AR-EVD module guides a user step-by-step through the intraoperative phase. First, an EVD trajectory with a target point inside a 3-D model of the ventricles is indicated. Optionally, 6 to 8 fiducials are planned on the 3-D model of the skin to facilitate point-based matching. Subsequently, point-based or contour-based image-to-patient registration is performed using a custom stainless-steel pointer and reference device, both equipped with laser-engraved optical markers. Continuous guidance of the EVD is provided using a custom tracking tool. A calibration step automatically determines the position of the EVD tip relative to the tracking tool. During EVD placement, the software continuously displays the distance to the target point and the angular offset between the EVD and the target point.

To test the feasibility of the workflow, two neurosurgical residents performed bilateral AR-guided EVD placements on two anatomical phantoms. Post-operatively, the phantoms were scanned using CT to determine placement accuracy using the Kakarla grading scale[9].

Results

No significant functional errors, bugs or crashes were observed during preliminary experiments. AR-guided EVD placement was completed in 4 out of 4 placement attempts. 3 out of 4 placements were

graded as a Kakarla 1, while 1 was graded as a Kakarla 2 due to placement in the contralateral ventricle. The average Euclidian distance from the intended target point to the actual drain tip was 9.50 mm (SD ± 3.96 mm). The mean operational time of the complete procedure was 22:45 mm:ss (SD ± 11:38 mm:ss). Feedback indicated challenges regarding holographic interfacing and maintaining marker tracking.

Discussion and Conclusion

The AR-based system shows potential for enhancing the accuracy of EVD placements through real-time, interactive guidance. Advantages include the direct superimposition of holographic information, tracking of the relative position of the EVD, the possibility to pre-operatively evaluate the EVD's trajectory and the limited required resource investment compared to conventional neuronavigation. The system is readily integrable with existing digital hospital infrastructure.

However, despite 3 out of 4 feasibility procedures achieving EVD placement within the ipsilateral frontal horn, the deviation from the target point was still relatively large. Furthermore, using the system adds several minutes to the procedure. Lastly, challenges with user engagement in the virtual interface indicated suboptimal human-computer interaction. Feasibility testing was limited due to the small number of conducted procedures and the tactile and technical differences between the phantoms and real patients.

The potential for this technology to reduce suboptimal placements and improve clinical outcomes is promising, but further validation is necessary. Future work will focus on refining the system's accuracy, human-computer elements and expanding trial sizes.

References

1. Rosenbaum BP, Vadera S, Kelly ML, Kshettry VR, Weil RJ. Ventriculostomy: Frequency, length of stay and in-hospital mortality in the United States of America, 1988-2010. *J Clin Neurosci Off J Neurosurg Soc Australas*. 2014;21(4):623-632. doi:10.1016/j.jocn.2013.09.001

2. Stuart MJ, Antony J, Withers TK, Ng W. Systematic review and meta-analysis of external ventricular drain placement accuracy and narrative review of guidance devices. *J Clin Neurosci Off J Neurosurg Soc Australas*. 2021;94:140-151. doi:10.1016/j.jocn.2021.10.014

3. Li Y, Chen X, Wang N, et al. A wearable mixed-reality holographic computer for guiding external ventricular drain insertion at the bedside. *J Neurosurg*. 2019;131(5):1599-1606. doi:10.3171/2018.4.JNS18124

4. Schneider M, Kunz C, Pal'a A, Wirtz CR, Mathis-Ullrich F, Hlaváč M. Augmented reality–assisted ventriculostomy. *Neurosurg Focus*. 2021;50(1):E16. doi:10.3171/2020.10.FOCUS20779

5. Van Gestel F, Frantz T, Vannerom C, et al. The effect of augmented reality on the accuracy and learning curve of external ventricular drain placement. *Neurosurg Focus*. 2021;51(2):E8. doi:10.3171/2021.5.FOCUS21215

6. Benmahdjoub M, Thabit A, Van Veelen MLC, Niessen WJ, Wolvius EB, Walsum TV. Evaluation of AR visualization approaches for catheter insertion into the ventricle cavity. *IEEE Trans Vis Comput Graph*. 2023;29(5):2434-2445. doi:10.1109/TVCG.2023.3247042

7. Fick T, Doormaal JAM van, Tosic L, et al. Fully automatic brain tumor segmentation for 3D evaluation in augmented reality. *Neurosurg Focus*. 2021;51(2):E14. doi:10.3171/2021.5.FOCUS21200

8. van Doormaal JAM, Fick T, Ali M, Köllen M, van der Kuijp V, van Doormaal TPC. Fully Automatic Adaptive Meshing Based Segmentation of the Ventricular System for Augmented Reality Visualization and Navigation. *World Neurosurg*. 2021;156:e9-e24. doi:10.1016/j.wneu.2021.07.099

9. Kakarla UK, Kim LJ, Chang SW, Theodore N, Spetzler RF. Safety and accuracy of bedside external ventricular drain placement. *Neurosurgery*. 2008;63(1 Suppl 1):ONS162-166; discussion ONS166-167. doi:10.1227/01.neu.0000335031.23521.d0

Disclosures

-T.P.C. van Doormaal is the founder and CMO of Augmedit B.V.

-J.W. Meulstee is a senior product developer for Augmedit B.V.

APPENDIX

Due to pending patent applications, figures pertaining to the hardware and software described in this study are currently withheld. These figures are expected to be publicly released prior to the time of CLINICCAI.

Figure 1: Point-based image-to-patient registration using AR-EVD module

Figure 2: Active EVD guidance using AR-EVD module

Introducing ClinIQ: A Non-Invasive Clinical Tool for Molecular Profiling of Brain Gliomas

Joseph A. Maldjian[1], Benjamin C. Wagner[1], Chandan Ganesh B. Y.[1], Nghi C. D. Truong[1], James M. Holcomb[1], Divya D. Reddy[1], Niloufar Saadat[1], Kimmo J. Hatanpaa[2], Toral R. Patel[3], Baowei Fei[4], Matthew D. Lee[5], Rajan Jain[5,6], Richard J. Bruce[7], Marco C. Pinho[1], Ananth J. Madhuranthakam

[1]Department of Radiology, University of Texas Southwestern Medical Center, [2]Department of Pathology, University of Texas Southwestern Medical Center, [3]Department of Neurological Surgery, University of Texas Southwestern Medical Center, [4]Department of Bioengineering, University of Texas Dallas, [5]Department of Radiology, NYU Grossman School of Medicine, [6]Department of Neurosurgery, NYU Grossman School of Medicine, [7]Department of Radiology, University of Wisconsin School of Medicine

Presenting author:

Joseph A. Maldjian, MD, Department of Radiology, UT Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75390-9178 (Joseph.Maldjian@utsouthwestern.edu) Office: 214-648-6751, Fax: 214-648-3904

Keywords:

Non-invasive molecular profiling, IDH-status, Deep learning, neuro-oncology, Image-based genetic subtyping, Clinical integration of AI diagnostics.

Key information:

Research Question: How can non-invasive image-based deep-learning tools accurately predict molecular markers in brain tumors, facilitating better therapeutic strategies and prognostic outcomes?

Findings: The MRI-based deep learning network, ***MC-net*** achieved a high accuracy of 96.4% in predicting IDH status. ***MC-net*** and ***Bayesian-Logistic-Regression*** model demonstrated a reliable bimodal distribution of mutation predictions.

Meaning: This study introduces a non-invasive MRI-based clinical tool that significantly enhances the accuracy of glioma profiling, promising to shift current paradigms in brain glioma diagnosis and management.

Introduction

Genetic-subtyping and molecular-profiling of brain tumors can transform therapeutic strategies and enhance prognostic accuracy.[1] Currently, molecular-markers are determined by obtaining glioma tissue via an invasive brain biopsy. Thus, non-invasive molecular-profiling has significant implications in therapy and prognosis. Recent studies show that non-invasive, image-based deep-learning-networks (DLN) can predict molecular-markers.[2-5] However, clinical-translation of these methods can be limited due to challenges in clinical-usability, extensive pre-processing and a gap between DLN-predictions and clinical-relevance. This study developed a clinical-tool (***ClinIQ***) for non-invasive molecular-profiling, designed for seamless integration with imaging-database-systems (XNAT/Flywheel) and capable of exporting outputs directly to PACS systems.

Material and methods

We developed an MRI-based-DLN (***MC-net***) for non-invasive prediction of IDH-status using a UNet. It was implemented for a voxel-wise dual-class-segmentation of the whole-tumor and IDH-prediction. ***MC-net*** was trained on 1082 cases from TCIA, UTSW and UCSF and evaluated on 1236 cases from NYU, UWM, EGD and UPENN.[6-9] We employed a Bayesian logistic regression (***BLR***) model to calculate confidence-scores (CS) for each prediction, using predicted voxel-percentages and ground truth from the test-data. ***MC-net+BLR*** was further tested and validated on 196 additional held-out UTSW cases.

The developed pipeline is as shown in Fig1. MRI from the scanner is converted from DICOM to NIftI using the Dcm2niix tool. Subsequently, a multi-contrast sequence picker selects the required sequences (T1-pre, T1GD, T2w & T2-Flair). These selected images are processed through the FeTS platform to produce template-registered and skull-stripped MRI, leading to IDH prediction using ***MC-net***. The ***BLR*** model then estimates the CS for each prediction. The output predictions including molecular-profiling and CS are stored in a JSON file. The C3d-tool (ITK) transforms the predicted segmentations back to native-space NIfTI. Custom Python-code converts the NIfTI segmentations into visually appealing color images and PyDicom converts them back to DICOM. Finally, the DCMTK[10]-tool transfers this data to PACS systems for evaluation with the remainder of the subject imaging data.

Results

MC-net achieved an excellent test-accuracy of 96.4%. Its voxel-wise prediction showed a bimodal-distribution with most cases classified either at very-low (0-5%) or very-high (95-100%) percentages of mutated-voxels (Fig 1). This distribution suggests that ***MC-net*** is highly effective in predicting IDH-status. ***MC-net+BLR*** achieved an accuracy of 94.4% on additional held-out test-cases providing proof of principle. Notably, the CS methodology allows for a higher likelihood than the overall accuracy for the IDH-wildtype class (approaching 98%).

Discussion and Conclusion

The developed tool for non-invasive molecular-profiling marks a substantial advancement in neuro-oncology. Combining a deep learning network (***MC-net)*** and a ***BLR*** model for molecular-profiling and estimating CS is an innovative approach to circumvent the invasive nature of current glioma biopsies. Notably, the high classification accuracy of ***MC-net*** emphasizes the potential of image-based molecular-profiling as a reliable adjunct to traditional invasive methods. This is further substantiated by the bimodal

distribution of predicted voxel-percentages, indicating a very high confidence in non-invasive molecular-profiling.

The ability to integrate ***ClinIQ*** with existing imaging database systems and export outputs directly to PACS reinforces its practicality and potential for widespread clinical adoption. The pipeline circumvents the traditional challenges associated with deep learning networks, such as extensive preprocessing and DLN-predictions' relevance, by optimizing the workflow for clinical usability and significance.

In conclusion, we present a clinical-tool for non-invasive molecular-profiling designed for seamless integration with XNAT/Flywheel and capable of exporting outputs directly to PACS. ***MC-net***'s high classification accuracy and CS from the ***BLR*** model demonstrate its efficacy and utility. This tool addresses the limitations of current approaches and has the potential to significantly improve patient outcomes and treatment paradigms. Future studies will be important in expanding the applicability in clinical settings, potentially establishing a new paradigm in diagnosing and managing brain tumors.

References

1. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology*. 2021;23(8):1231-1251.
2. Zhang B, Chang K, Ramkissoon S, et al. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro-oncology*. 2017;19(1):109-117.
3. Chang P, Grinband J, Weinberg B, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *American Journal of Neuroradiology*. 2018;39(7):1201-1207.
4. Chakrabarty S, LaMontagne P, Shimony J, Marcus DS, Sotiras A. MRI-based classification of IDH mutation and 1p/19q codeletion status of gliomas using a 2.5 D hybrid multi-task convolutional neural network. *Neuro-Oncology Advances*. 2023;5(1):vdad023.
5. Bangalore Yogananda CG, Wagner BC, Truong NC, et al. MRI-Based Deep Learning Method for Classification of IDH Mutation Status. *Bioengineering*. 2023;10(9):1045.
6. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*. 2013;26:1045-1057.
7. van der Voort SR, Incekara F, Wijnenga MM, et al. The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. *Data in brief*. 2021;37:107191.
8. Bakas S, Sako C, Akbari H, et al. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Scientific data*. 2022;9(1):453.
9. Calabrese E, Villanueva-Meyer JE, Rudie JD, et al. The university of california san francisco preoperative diffuse glioma mri dataset. *Radiology: Artificial Intelligence*. 2022;4(6):e220058.
10. Eichelberg M, Riesmeier J, Wilkens T, Hewett AJ, Barth A, Jensch P. Ten years of medical imaging standardization and prototypical implementation: the DICOM standard and the OFFIS DICOM toolkit (DCMTK). SPIE; 2004:57-68.

Disclosures

No Disclosures

APPENDIX

*Table 1: **MC-net** IDH Classification Results.*

	NYU	UWM	EGD	Upenn	Overall
Accuracy	96.0	93.6	96.3	98.0	**96.4**
Sensitivity	93.6	81.2	93.3	81.8	92.0
Specificity	96.9	94.7	97.7	98.5	97.3

*Figure 1: Flow chart of the developed pipeline **ClinIQ***

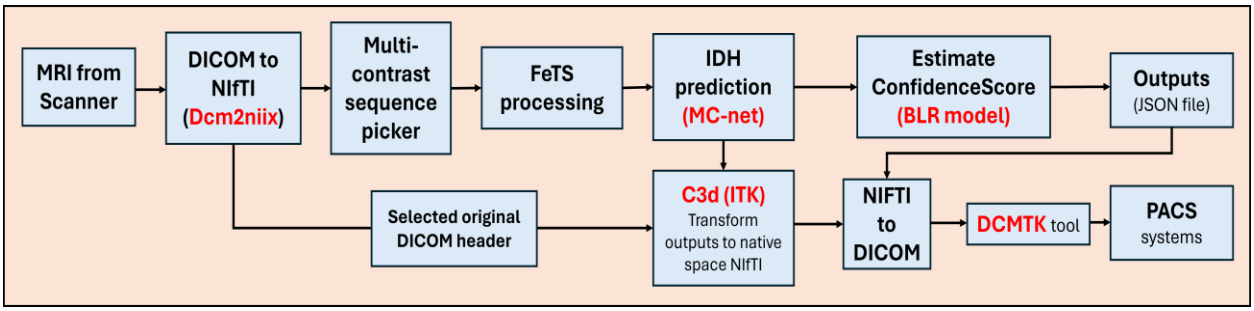
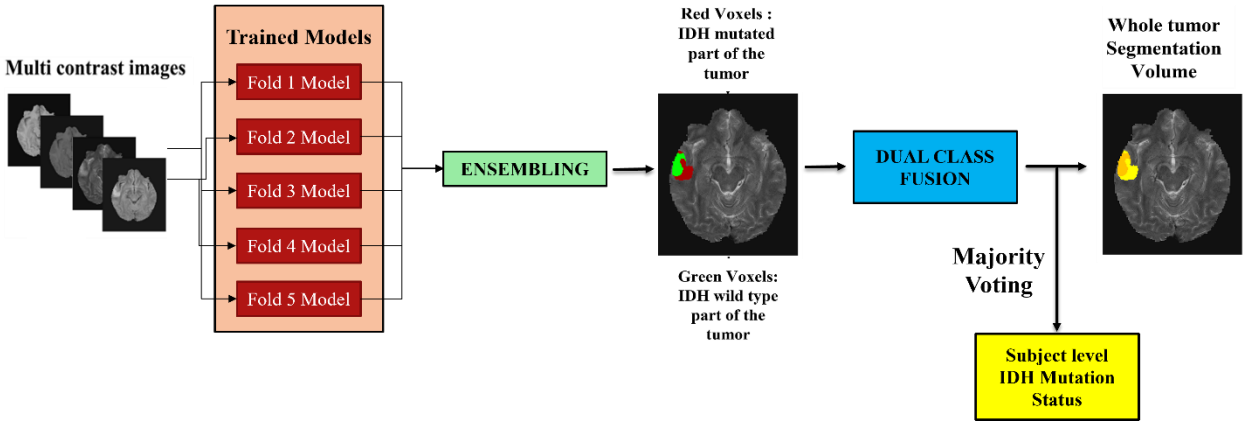


*Figure 2: **MC-net** Testing procedure.*



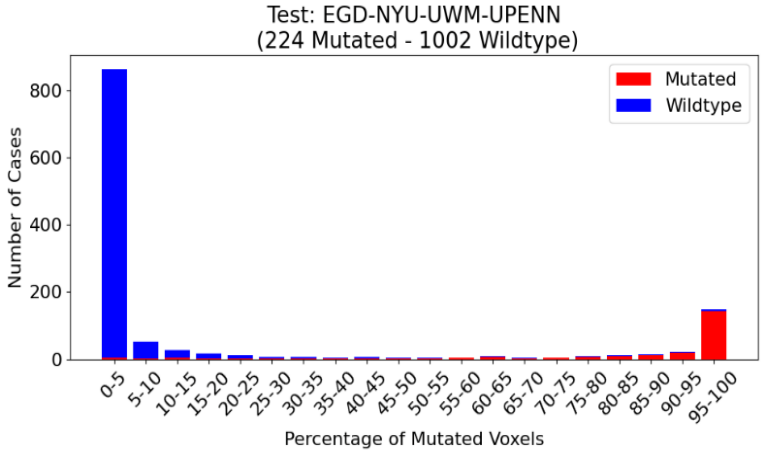


Figure 3: Distribution of percent mutant voxels on test dataset. The distribution shows very low (0-5%) or very high (95-100%) percentages of predicted mutated voxels, suggesting a clear differentiation between mutated and wildtype gliomas.

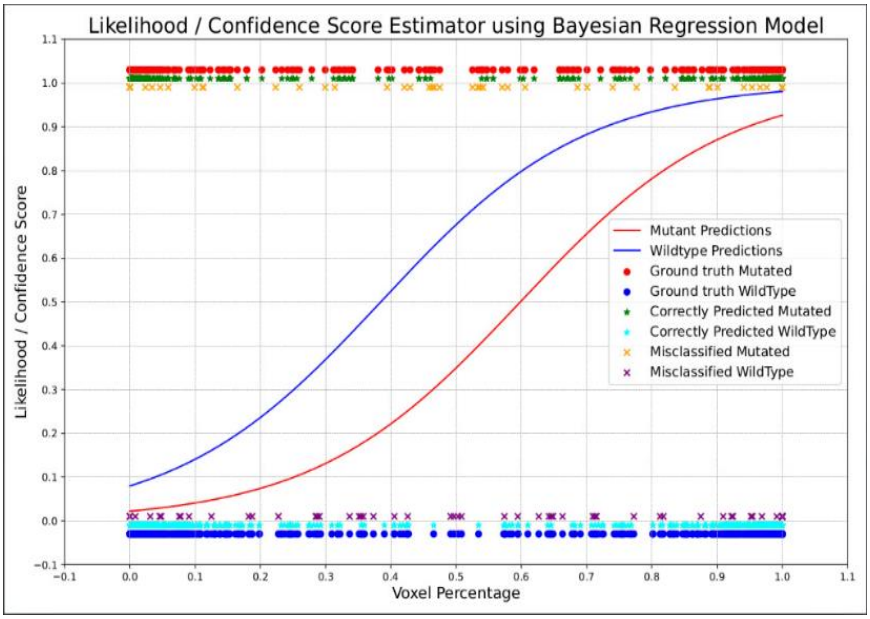


Figure 4: BLR curves to estimate Confidence Scores.

Determining cerebral white matter integrity and volume of brain structures in patients with primary Sjögren's syndrome in DTI and FLAIR with AI tools.

Authors: Michał Sobański[1], Adrian Korbecki[1,4], Tomasz Konopczyński[4], Piotr Gródek[4], Miłosz Gajowczyk, Patryk Rygiel[2], Justyna Korbecka[5], Kamil Litwinowicz[1,4], Martyna Sobańska[3], Agata Sebastian[6], Joanna Bladowska[7,8]

Affiliations:

1. Department of General Radiology, Interventional Radiology and Neuroradiology, Wroclaw Medical University, Wroclaw, Poland

2. Mathematics of Imaging and AI, Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands

3. Department of Maxillofacial Orthopaedics and Orthodontics, Wroclaw Medical University, Wroclaw, Poland

4. Hetalox sp. z o.o., Wroclaw, Poland

5. Clinical Department of Neurology, Medical University, Wroclaw, Poland

6. Clinical Department of Rheumatology and Internal Medicine, Wroclaw Medical University,Wroclaw, Poland

7. Faculty of Medicine, Wroclaw University of Science and Technology, Wroclaw, Poland

8. Department of Radiology, Wroclaw 4th Military Clinical Hospital, Wroclaw, Poland

Presenting author:

Michał Sobański, Department of General Radiology, Interventional Radiology and Neuroradiology, Wroclaw Medical University, Wroclaw, Poland. Contact phone: +48605603107. e-mail: michalsobanski11@wp.pl

Keywords:

Neuroradiology, DTI, AI, Sjögren's syndrome

Key information:

1. Research question: Do patients with Sjögren's syndrome present abnormal features in brain MRI? Can these abnormalities be described with the usefulness of AI tools?
2. Findings: Our study confirmed multiple disrupted white matter tracts and reduced volumes in particular brain areas, including novel observations in cerebellar peduncles and optic radiations. These findings are correlated by other observations of significantly lowered volumes of cerebellar cortex of base of the frontal lobes.

3. Meaning: AI tools provide fast and accurate assessment of integrity of white matter tracts and volumes of brain structures, which can help focus on patients with higher risk of central nervous system involvement. DTI and FLAIR could assist in the decision to start, monitor or change/intensify treatment.

MANUSCRIPT

Introduction

Primary Sjögren's syndrome (pSS) is an autoimmune disease in which central nervous system (CNS) involvement may occur. Incidence and pathogenesis of changes related to CNS in pSS is not well understood. The diffusion tensor imaging (DTI) can be used to demonstrate impaired integrity of cerebral white matter (CWM), especially by quantitative values such as fractional anisotrophy (FA). Moreover high resolution 3D sequences, such as 3D FLAIR images, can be used for assessment volumes of brain structures.

Material and methods

A study group containing 33 patients with primary Sjögren's syndrome and a control group containing 26 healthy patients were studied by performing DTI and 3D FLAIR sequences. The study group consisted of patients who met the 2016 ACR/EULAR classification criteria for pSS. The exclusion criteria consisted of: coexisting of other connective tissue diseases, infectious diseases (including viral infection 3 months due to MRI examination, also COVID-19), any neurologic and psychiatric diseases.

The DTI data was processed by a TractSeg[1] algorithm, which is a fully convolutional neural network (FCNN), resulting in a segmentation of 72 white matter (WM) tracts per patient, as well as value of FA for each tract. The segmented WM tracts are presented in Fig. 1.

The cerebral cortex was segmented on 3D FLAIR images into 68 parcells and the deep brain structures and intracranial spaces were segmented to 33 structures by a Hetalox brain segmentation algorithm[2]. Segmentation labels follow the FreeSurfer[3] classification style, corresponding to the Desikan-Killiany atlas labels with minor modifications are presented on Fig. 2.

Results

Considering all WM tracts collectively ('TOTAL') we observed a significant difference between the study group and control group. Numerous areas with a significant reduction in FA parameter values in the study group relative to the control group were shown in (Fig. 3). The novel observations of altered values of FA in cerebellar peduncles and optic radiations were presented. The areas with disrupted WM tracts correlated with significantly lower volumes of cortex in cerebellum and base of frontal lobes (Tab. 1).

Discussion and Conclusion

We found a significant ($p<0.002$) reduction of FA value in all cerebellar peduncles and a significant ($p<0,04$) reduction of cerebellar cortex volume in pSS, to the best of our knowledge such findings have never been reported before. Cerebellum damage is typically associated with locomotor disturbances. Our study revealed reduced FA

values in both fronto-pontine tracts ($p<0.0005$), suggesting predictive pathological alterations in fronto-pontine-cerebellar tracts. Disruption in these tracts may impair procedural memory processes, affecting the learning and execution of complex skills.[4]

Furthermore, we observed a significant ($p<0.02$) reduction in FA values in both optic radiations, a previously unreported observation. Similar reductions in FA value have been found in autoimmune diseases like neuromyelitis optica spectrum disorders (NMOSD),[5] which is suggested to be within the spectrum of CNS diseases in pSS patients.[6] The reduction in FA value is likely due to secondary Wallerian degeneration, possibly arising from changes in anatomical locations typical of alteration observed in NMOSD, particularly optic nerve involvement.[5]

In our study patients with pSS exhibited a significant ($p<0.05$) reduction in FA value and volume in prefrontal area which can, affect higher brain functions as: emotional processes, memory disorders, cognitive processes, habituation.[7]

Concluding, DTI and FLAIR of the brain are non-contrast agent studie that could be a very useful tools used for screening and monitoring the severity of neuro-pSS. The usefulness of automated methods such as TractSeg's algorithm and Hetalox brain segmentation, allows for rapid and reproducible collection of qualitative and quantitative data. Most importantly, the method requires minimal intervention and achieves automatic, reproducible segmentation.

References

1. Wasserthal, J., Neher, P., & Maier-Hein, K. H. (2018). TractSeg - Fast and accurate white matter tract segmentation. *NeuroImage*, *183*, 239–253. https://doi.org/10.1016/j.neuroimage.2018.07.070
2. Hetalox brain segmentation [Computer software]. (2024). Retrieved from http://hetalox.com
3. Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/J.NEUROIMAGE.2012.01.021
4. Janusz Moryś, J. N. (2021). *MÓZG CZŁOWIEKA. ANATOMIA CZYNNOŚCIOWA MÓZGOWIA. TOM 2* (8th ed., Vol. 2).
5. Jeantroux, J., Kremer, S., Lin, X. Z., Collongues, N., Chanson, J. B., Bourre, B., Fleury, M., Blanc, F., Dietemann, J. L., & de Seze, J. (2012). Diffusion tensor imaging of normal-appearing white matter in neuromyelitis optica. *Journal of Neuroradiology*, *39*(5), 295–300. https://doi.org/10.1016/j.neurad.2011.10.003
6. Morreale, M., Marchione, P., Giacomini, P., Pontecorvo, S., Marianetti, M., Vento, C., Tinelli, E., & Francia, A. (2014). Neurological involvement in primary Sjögren syndrome: A focus on central nervous system. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0084605

7. Fenlon, L. R., Suarez, R., Lynton, Z., & Richards, L. J. (2021). The evolution, formation and connectivity of the anterior commissure. In *Seminars in Cell and Developmental Biology* (Vol. 118, pp. 50–59). Elsevier Ltd. https://doi.org/10.1016/j.semcdb.2021.04.009

Disclosures

Adrian Korbecki, Kamil Litwinowicz and Tomasz Konopczyński are with Hetalox sp. z o.o., which provided the algorithm to calculate cerebral cortex parcellations and deep brain structures and intracranial spaces segmentation from FLAIR sequences. All other authors have nothing to declare.

APPENDIX

Figures

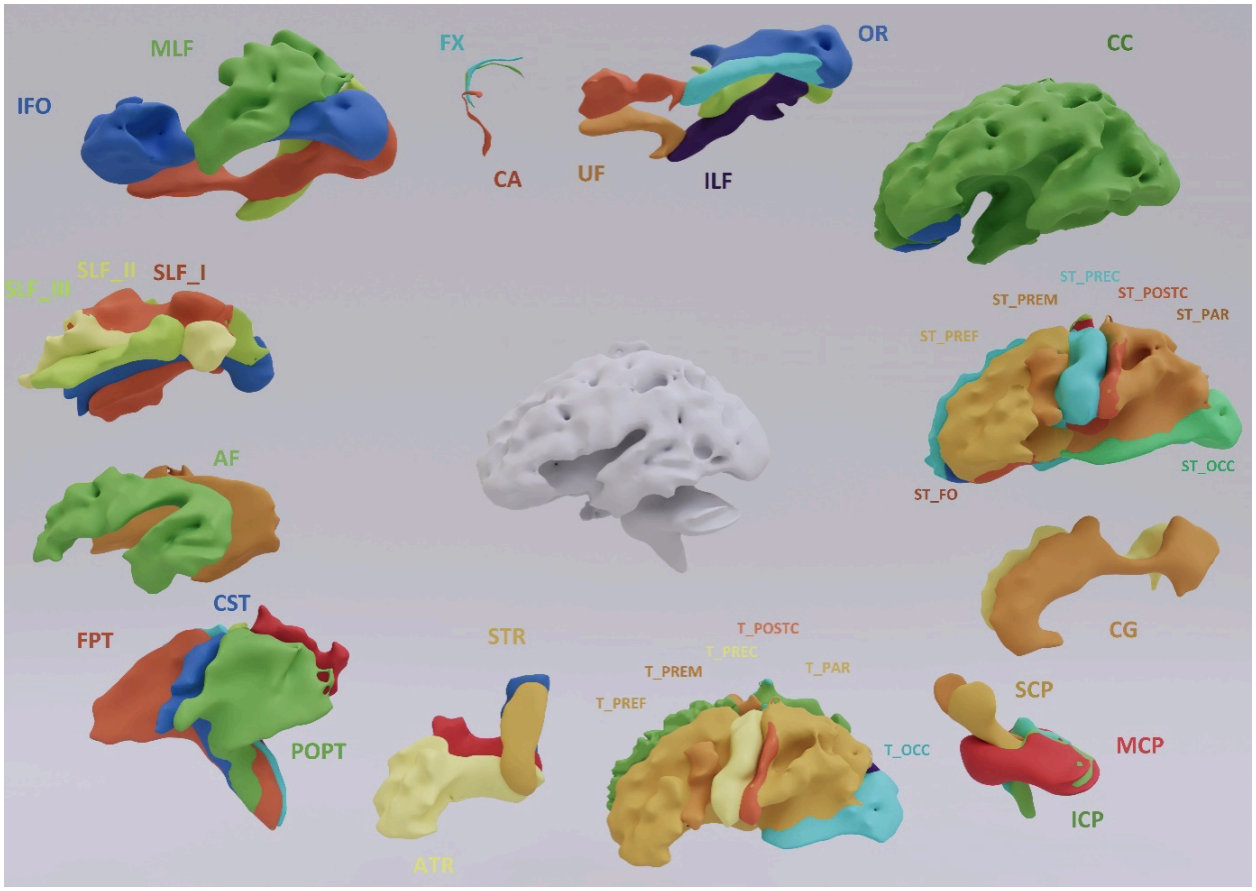


Fig. 1 Overview of all 72 WM tracts segmented by TractSeg of the brain (for tracts that occur in both the left and right cerebral hemispheres, only the right-tract is shown in the figure). WM tracts abbreviations arcuate fascicle- AF; anterior thalamic radiation – ATR; commissure anterior – CA; corpus callosum – CC (rostrum – CC1, genu – CC2, rostral body – CC3, anterior body – CC4, posterior body – CC5, isthmus – CC 6, splenium - CC7), anterior thalamic radiation – ATR; cingulum – CG; corticospinal tract – CST; middle longitudinal fascicle – MLF; fronto-pontine tract – FPT; fornix – FX; inferior cerebellar peduncle – ICP; inferior occipito-frontal fascicle – IFOF; inferior longitudinal fascicle – ILF; . parieto-occipital pontine tract – POPT; superior cerebellar peduncle – SCP; superior longitudinal fascicle I – SLF I; superior longitudinal fascicle I – SLF II; superior longitudinal fascicle III – SLF III; superior thalamic radiation – STR; uncinate fascicle – UF; thalamo-prefrontal tract – T_PREF; thalamo-premotor tract – T_PREM; thalamo-precentral tract – T_PREC; thalamo-postcentral tract – T_POSTC; thalamo-parietal tract – T_PAR; thalamo-occipital tract – T_OCC; ang. striato-fronto-orbital tract – ST_FO; striato-prefrontal tract – ST_PREF; striato-premotor tract – ST_PREM; striato-precentral tract – ST_PREC; striato-postcentral tract – ST_POSTC; striato-parietal tract – ST_PAR; striato-occipital tract – ST_OCC.

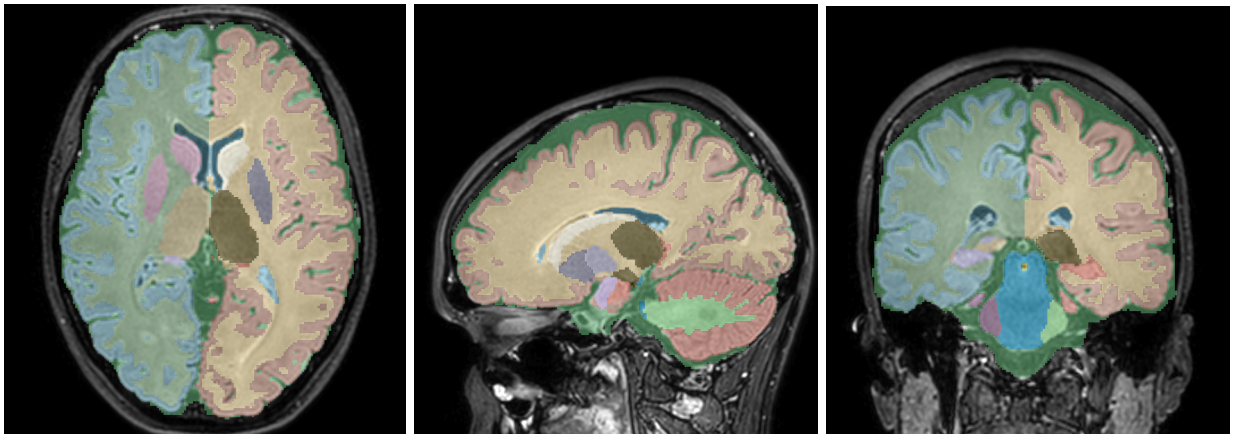


Fig. 2 Example brain segmentation of FLAIR sequence using Hetalox segmentation software.

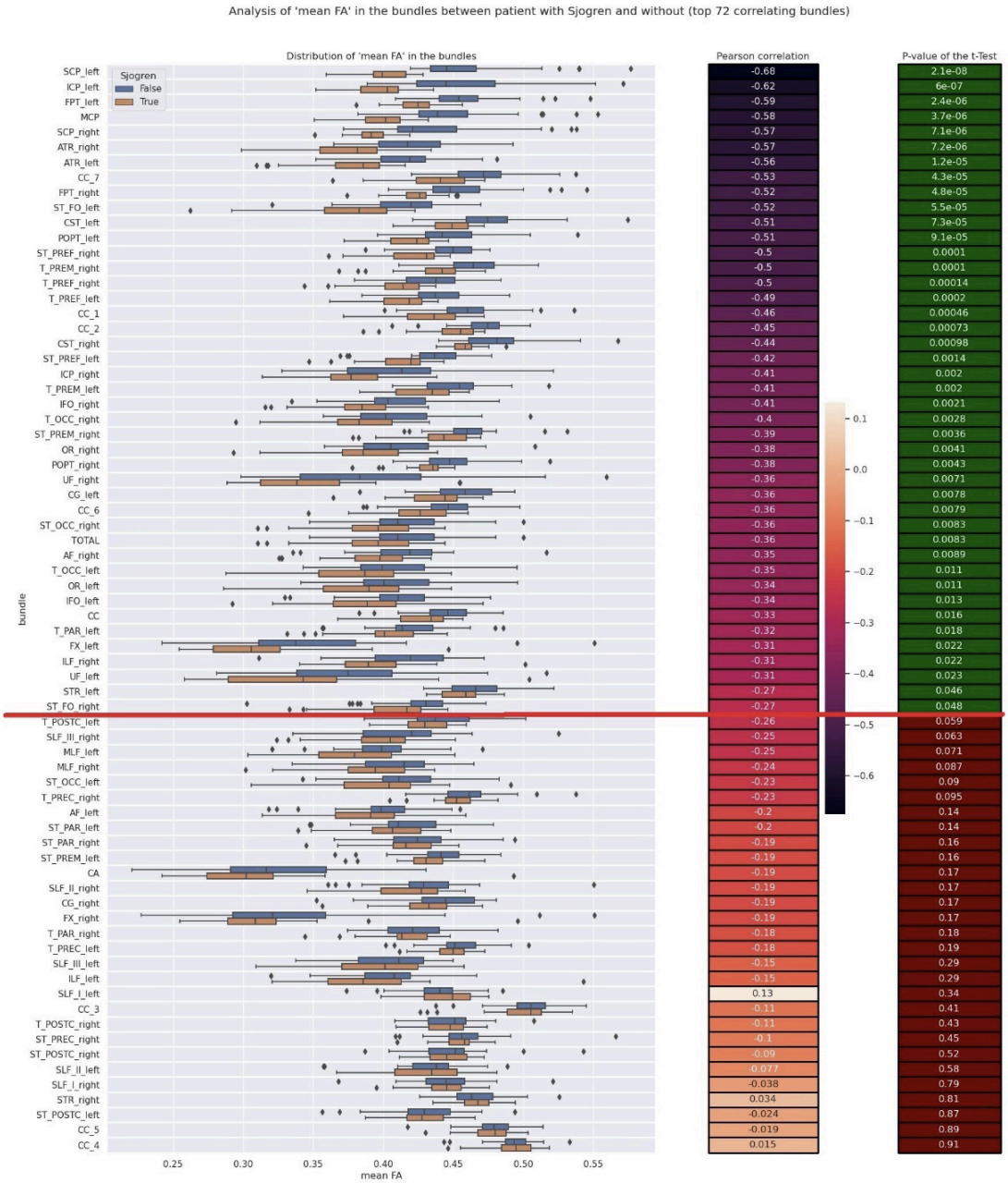


Fig. 3 Tracts with statistically significant difference between study and control group are shown above the red line. Pearson's correlation (r) and p-value are placed next to box-plots. WM tracts abbreviations: arcuate fascicle- AF; anterior thalamic radiation – ATR; commissure anterior – CA; corpus callosum – CC

(rostrum – CC1, genu – CC2, rostral body – CC3, anterior body – CC4, posterior body – CC5, isthmus – CC 6, splenium - CC7), anterior thalamic radiation – ATR; cingulum – CG; corticospinal tract – CST; middle longitudinal fascicle – MLF; fronto-pontine tract – FPT; fornix – FX; inferior cerebellar peduncle – ICP; inferior occipito-frontal fascicle – IFOF; inferior longitudinal fascicle – ILF; parieto-occipital pontine tract – POPT; superior cerebellar peduncle – SCP; superior longitudinal fascicle I – SLF I; superior longitudinal fascicle I – SLF II; superior longitudinal fascicle III – SLF III; superior thalamic radiation – STR; uncinate fascicle – UF; thalamo-prefrontal tract – T_PREF; thalamo-premotor tract – T_PREM; thalamo-precentral tract – T_PREC; thalamo-postcentral tract – T_POSTC; thalamo-parietal tract – T_PAR; thalamo-occipital tract – T_OCC; ang. striato-fronto-orbital tract – ST_FO; striato-prefrontal tract – ST_PREF; striato-premotor tract – ST_PREM; striato-precentral tract – ST_PREC; striato-postcentral tract – ST_POSTC; striato-parietal tract – ST_PAR; striato-occipital tract – ST_OCC.

Tables

Table 1: Brain structures with significantly lower volumes in pSS patients compared to control group.

Structure	p value
left inferior lateral ventricle	0.0172
left cerebellum cortex	0.0285
left accumbens area	0.0294
right cerebellum cortex	0.0378
rh parsopercularis	0.0023
rh superiorfrontal	0.0110

Generating Computed Tomography Angiography from Time-of-Flight Magnetic Resonance Angiography Using Diffusion-Based Models

Authors:
Dietmar Frey, MD[1,2], Alexander Koch[1], Adam Hilbert[1], Dimitrios Rallios[1], Orhun Utku Aydin, MD[1]

Affiliations:
[1]CLAIM – Charité Lab for AI In Medicine, Charité Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany

[2]Department of Neurosurgery, Charité Universitätsmedizin Berlin

Presenting author:
Dietmar Frey
dietmar.frey@charite.de

Keywords:
Machine Learning, Cross-modality image synthesis, Diffusion, TOF-MRA

Key information:

1. Research question:
   Can synthetic computed tomography angiography (CTA) images be generated from time-of-flight magnetic resonance angiography (TOF-MRA) images using paired image-to-image translation?
2. Findings:
   Realistic synthetic CTA images were generated based on the corresponding co-registered TOF-MRA images. While a baseline U-Net architecture produces suboptimal results, the application of a diffusion-based model significantly improves the perceived visual quality of the synthetic CTA images.
3. Meaning:
   The ability to generate realistic synthetic CTA from TOF-MRA could alleviate the scarcity of open-source CTA data and synthetic data be used for the development of AI computer vision models for clinical use.

MANUSCRIPT

Introduction

Vessel neuroimaging techniques such as Time-of-Flight Magnetic Resonance Angiography (TOF-MRA) and Computed Tomography Angiography (CTA) provide unique visual information crucial for diagnosis, treatment and monitoring of cerebrovascular disease. While CTA might offer unique advantages with respect to diagnostic sensitivity, cost and acquisition time, it is also associated with limitations such as radiation exposure and potential side effects or contraindications of contrast agents. In addition, most AI-based imaging applications in acute stroke are developed for CTA workflows, whereas

the availability of open-source CTA data of stroke patients is highly limited. Image-to-image translation is a promising option used to generate synthetic data from an input modality. Synthetic complementary CTA vessel imaging data can be used to improve the robustness and generalization of clinical AI applications. In this work, we generate CTA images from TOF-MRA using denoising-diffusion probabilistic models[1].

Material and Methods

We use data from the Topology-Aware Anatomical Segmentation of the Circle of Willis (TopCow)[2] challenge. The dataset comprises patients admitted to the Stroke Center of the University Hospital Zurich and provides paired CTA and TOF-MRA imaging. The available data is already anonymised and defaced.

We co-register the TOF-MRA images to the CTA images using FSL-FLIRT[3]. Afterwards, the images are resampled to (256,256,256) with an isotropic voxel spacing of 0.6mm. We extract 2D slices, 80 per patient. Overall we use 82 patients and split them into training (n=65) and test sets (n=17).

Two different paired image-to-image translation methods are trained and compared to learn a mapping from TOF-MRA to CTA: a U-Net[4] baseline and the same U-Net architecture reframed as a diffusion-based model with the source image as conditioning information. We use diffusion with v-objective parametrization[5] and Min-SNR weighting[6]. We augment using random flipping, translation and cut-outs. Random cut-outs are only applied to the source image.

We evaluate the model using Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity Index Measure (MSSIM, SSIM) and Learned Perceptual Image Patch Similarity (LPIPS)[7] perceptual loss using AlexNet[8]. Moreover, we evaluate the model on a downstream task, training segmentation models of the Circle of Willis on the synthetically generated data. The nnUnet[9] framework was used to train the models with 5-fold cross-validation. The test dataset for segmentation (n=17) was split into training (n=10) and test (n=7) sets. The segmentations were compared using the Dice coefficient and the (balanced) average Hausdorff distance.

Results

In the qualitative evaluation, the U-Net baseline provides images with high structural integrity but with limited perceived visual quality. CTA images created by the diffusion-based model look visually more realistic and the synthetic texture of the brain tissue resembles the original CT more closely. In the quantitative evaluation, the diffusion model outperformed the U-Net baseline in terms of perceptual loss but showed inferior results on other measures indicating a discrepancy to the visual analysis.

In the downstream segmentation analysis, the models trained on CTA data synthesized from TOF-MRA data could be applied to real CTA images to segment arteries of the Circle of Willis. Training on synthetic data generated by the diffusion model achieved better segmentation results compared to synthetic data of the U-Net baseline model.

Discussion and Conclusion

We showed a promising way of performing modality conversion in vessel neuroimaging. The proposed method can be used to generate synthetic CTA data from available TOF-MRA datasets to be used in downstream AI applications such as aneurysm segmentation, occlusion detection and automated collateral score assessment. This has the potential to improve the accuracy and generalization of AI models in medical imaging. Future works should assess the potential benefits and use cases of synthetic CTA images in clinical practice.

References

1. Ho J, Jain A, Abbeel P. Denoising Diffusion Probabilistic Models. In: *Advances in Neural Information Processing Systems*. Vol 33. Curran Associates, Inc.; 2020:6840-6851. https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. Accessed April 25, 2024.
2. Yang K, Musio F, Ma Y, et al. TopCoW: Benchmarking Topology-Aware Anatomical Segmentation of the Circle of Willis (CoW) for CTA and MRA. 2023.
3. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *NeuroImage*. 2012;62(2):782-790. doi:10.1016/j.neuroimage.2011.09.015
4. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing; 2015:234-241. doi:10.1007/978-3-319-24574-4_28
5. Salimans T, Ho J. Progressive Distillation for Fast Sampling of Diffusion Models. In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net; 2022. https://openreview.net/forum?id=TIdIXIpzhoI.
6. Hang T, Gu S, Li C, et al. Efficient Diffusion Training via Min-SNR Weighting Strategy. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE; 2023:7407-7417. doi:10.1109/ICCV51070.2023.00684
7. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE; 2018:586-595. doi:10.1109/CVPR.2018.00068
8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*. Vol 25. Curran Associates, Inc.; 2012. https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html. Accessed April 25, 2024.
9. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z

Disclosures

No conflict of interest.

Appendix

Figure 1: Baseline U-Net model performance and conditional diffusion model performance on the validation set. The diffusion model was sampled for 100 steps. The first row shows the TOF-MRA image, the second row the target CTA image, the third row the generated image by the Diffusion model, the fourth row the image produced by the U-Net baseline

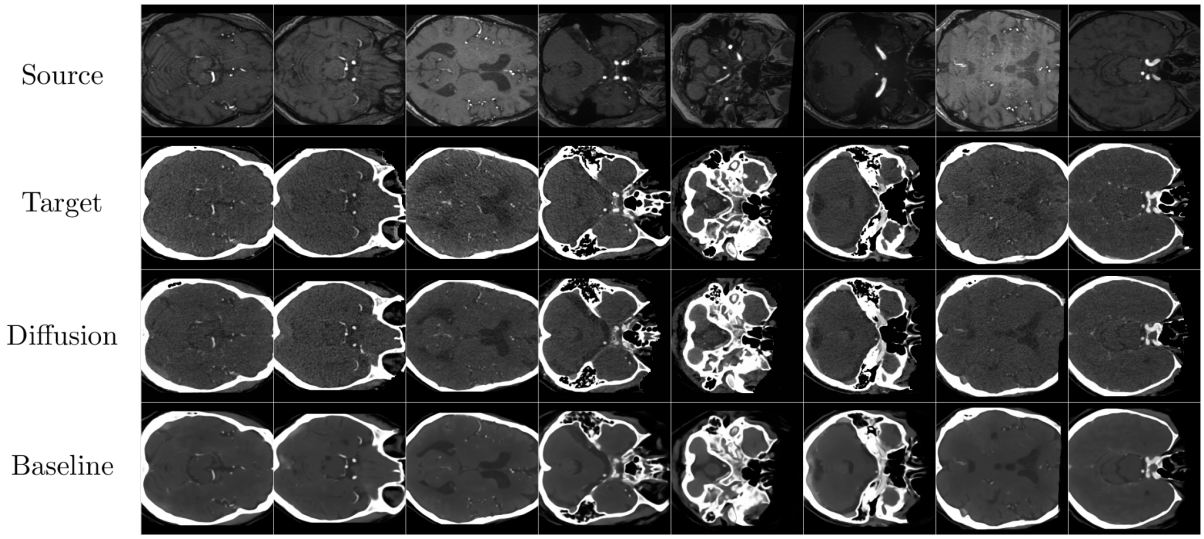


Table 1: Comparison of Metrics for the two methods (down arrow: lower is better, up arrow: higher is better). The diffusion model is sampled for 100 sampling steps.

Method	val loss (↓)	SSIM (↑)	MSSIM (↑)	PSNR (↑)	perceptual (↓)
Baseline	0.0695	0.5941	0.7796	17.6020	0.3087
Diffusion	0.0958	0.5216	0.7239	16.2069	0.1685

Figure 2: Downstream vessel segmentation performance using real and translated synthetic CTA data.

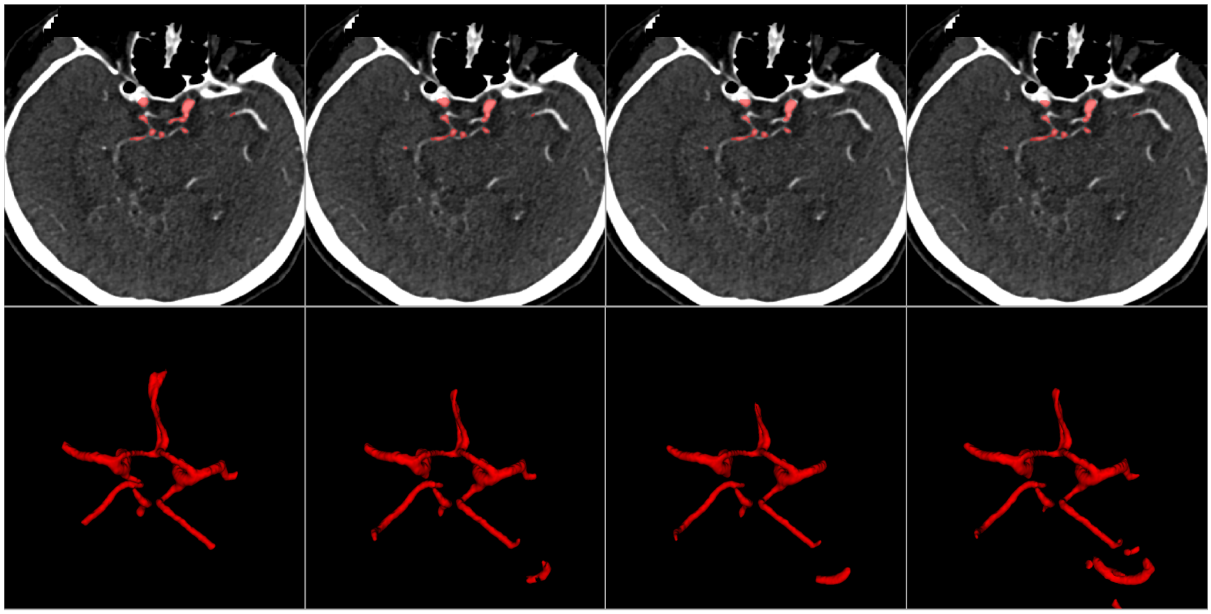


Table 2: Downstream segmentation task analysis (down arrow: lower is better, up arrow: higher is better)

Training Data	Dice (↑)	AHD (↓)	bAHD (↓)
Real data	0.831	1.493	1.559
Baseline	0.744	4.360	5.052
Diffusion	0.764	3.310	3.398

AI-based Outcome Prediction in Ischemic Stroke: A Study on Robust External Validation Principles

Authors:

Dietmar Frey, MD[1,2], Jana Rieger[1], Orhun Utku Aydin[1], Jaeyoung Cho[3], Luis M Lopez-Ramos[4], John Kelleher[5], Frank te Nijenhuis[6], Marta Rubiera[7], Susanne Bonekamp[3], Ronen Leker[8], Satoru Tanioka ,MD[1], Theo van Walsum[6], Aad van der Lugt[9], Adrian Mak[1], Sam Payabvash[10], Anna Matsulevits[1], Thomas Tourdias[11], Christian H Nolte, MD[12], Adam Hilbert[1]
on behalf of the VALIDATE, MRCLEAN and ICPSR investigators

Affiliations:

[1] Charité Lab for AI In Medicine (CLAIM), Charité Universitätsmedizin Berlin
[2] Department of Neurosurgery, Charité Universitätsmedizin Berlin
[3] Department of Neuroradiology, Heidelberg University Hospital, Heidelberg
[4] Simula Metropolitan Center for Digital Engineering, Oslo
[5] School of Computer Science and Statistics Trinity College Dublin, Dublin
[6] Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam
[7] Department of Neurology, Vall d'Hebrón University Hospital, Barcelona
[8] Department of Neurology, Hadassah Medical Center, Jerusalem
[9] Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam
[10] Department of Neuroradiology, Yale School of Medicine, New Haven
[11] Department of Neuroradiology, CHU Pellegrin, Bordeaux
[12] Department of Neurology with Experimental Neurology and Center for Stroke Research Berlin (CSB), Charité Universitätsmedizin Berlin, Berlin

Presenting author:

Dr. Dietmar Frey, Charité Lab for AI in Medicine (CLAIM) - Charité University Hospital, Berlin
dietmar.frey@charite.de, +49 30 450 560398

Keywords

Artificial intelligence, outcome prediction, decision support

Key Information

- **Research question**

The aim of the study is to evaluate the generalization capabilities of a Neural Network trained for predicting functional outcome on day 90 after ischemic stroke measured by the modified Rankin Scale.

- **Findings**

Model performance on the derivation cohort is strong and seems to persist across the analyzed external validation sets.

- **Meaning**

We show the most comprehensive external validation of an AI-based prediction model in stroke so far. The studied model has the potential to support treatment decisions across the globe. Our methodology can inspire future research to ensure reliable translation of AI into clinical practice.

Manuscript

Introduction

Stroke is a leading cause of disability and mortality worldwide. Modelling of prognosis and automation of biomarker extraction have the potential to aid treatment decisions and foster precision medicine in stroke. However, to date the AI models proposed for stroke decision support have struggled to retain their performance achieved on derivation cohorts when applied to unseen datasets or populations. Hence, external validation of AI-based approaches with robust testing methodology is a vital requirement to assess the generalization capabilities of models and validity in various patient groups; unfortunately, standardized AI guidelines for validation studies to assist researchers are scarce.

Materials and Methods

In the current project, we aim to investigate novel methodological principles for clinical validation of a Neural Network model trained on clinical tabular data of ischemic stroke patients to predict the full scale of 90 day modified Rankin Scale (mRS). We established a comprehensive validation cohort, containing retrospective data from 10+ centers on three continents. Classification performance is tested according to multiple metrics (e.g. sensitivity, specificity, receiver operating characteristic analysis) for the full-scale of the mRS score as well as in dichotomized (mRS<=2; mRS>2) and trichotomized (mRS<=2; 2<mRS<=4; mRS>4) scenarios. Furthermore, a novel multi-class model calibration approach and feature importance analysis are employed.

Besides general performance analysis, we also conduct a comprehensive sub-groups analysis concerning groups of age (e.g. <65, >80), gender, stroke severity (National Institutes of Health Stroke Scale - NIHSS>5, Alberta Stroke Program Early CT Score - ASPECTS<8), occluded artery (e.g. posterior circulation, ICA, M1/M2), medical history (e.g. previous stroke, cardiac record), and treatment methods (intravenous thrombolysis, mechanical thrombectomy).

Results

To date, we have finalized data inclusion and managed to include patients from the German Stroke Registry (GSR, n=5,412), MR CLEAN trial (n=500)[1], Hadassah Medical Center (HAD, n=944), University Hospital Heidelberg (HDB, n=1,254), University Hospital Vall d'Hebron (n=2,170), MR CLEAN Registry (MRCLEAN-R, n=4,737), Bordeaux University Hospital (n=237), Mie Chuo Medical Center (n=30), Mie University Hospital (n=37), Mie Prefectural General Medical Center (n=117), Suzuka Kaisei Hospital (n=97), Yale New Haven Hospital (n=445) and an open source dataset by the Inter-university Consortium for Political and Social Research (n=2,888)[2].

The model was trained on the combination of the GSR and MR CLEAN trial data, and then consecutively fine-tuned on data from HDB and HAD in two separate steps. 29 features including demographics, clinical examination and lab results (NIHSS, blood-pressure, serum glucose, etc.), medical history (stroke, atrial fibrillation, etc.), image-derived information (ASPECTS, Collateral score, occluded artery) and workflow variables (onset-to-admission/imaging) were used. Model training and fine-tuning involved 4-fold cross validation and selection of the best model with respect to test set performance (distinct from the training set) in each step. All remaining datasets are used for external validation.

The model achieved high prediction performance in the test sets of the training and fine-tuning cohorts (dichotomized/trichotomized/full-scale AUCs: 0.75-0.88/0.68-0.84/0.66-0.76) and showed good calibration in calibration analysis (Expected Calibration Error averaged across increasing mRS thresholds was 0.038). The model ranks age, NIHSS and pre-stroke mRS consistently as the 3 most important features.

Preliminary analysis of the combined Japanese cohort (JPA) and the MRCLEAN-R showed good generalization and minor deviation of classification performance only in full-scale mRS predictions (JPA dichotomized/trichotomized/full-scale AUC: 0.80/0.72/0.67, MRCLEAN-R dichotomized/trichotomized/full-scale AUC: 0.76/0.70/0.65). The analysis is ongoing and will be concluded before the conference date.

Discussion and Conclusion

We present the largest scale external validation of an AI-based prediction model in stroke up to date. Our baseline results show strong prediction performance of the model, and its generalization capabilities seem highly promising. Moreover, our study serves as a strong methodological blueprint for reliable translation of AI models into clinical practice.

References:

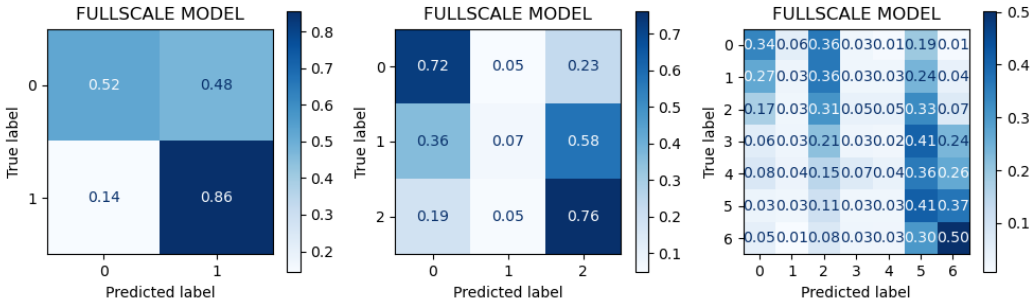
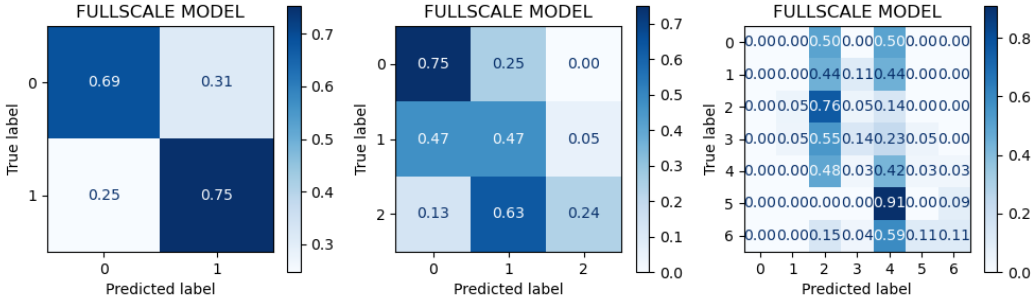
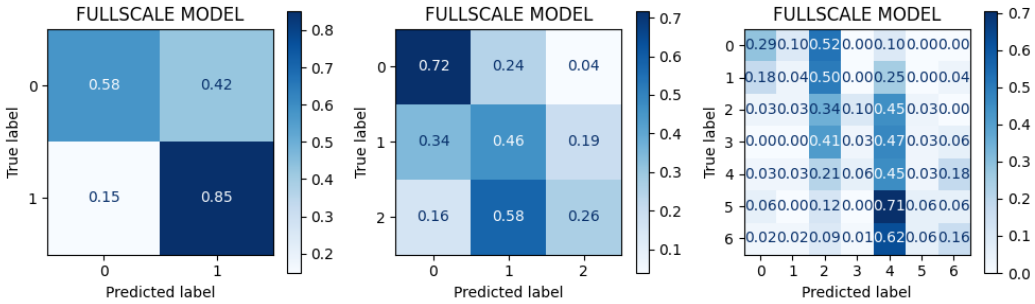
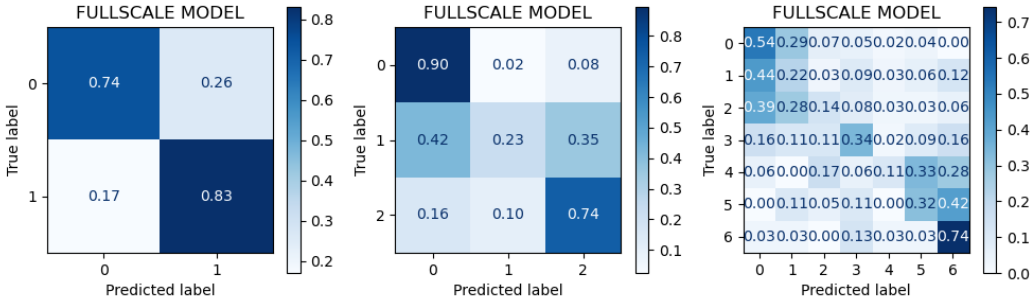
1. Berkhemer OA, Fransen PSS, Beumer D, et al. A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. *N Engl J Med*. 2015;372(1):11-20. doi:10.1056/NEJMoa1411587

2. Faria AV. Annotated Clinical MRIs and Linked Metadata of Patients with Acute Stroke, Baltimore, Maryland, 2009-2019: Version 5. Published online 2022. doi:10.3886/ICPSR38464.V5

Disclosures:

No conflict of interest.

Appendix

Evaluation on	Confusion matrices
Derivation datasets	
GSR	
MR CLEAN trial	
HAD	
HDB	
External validation sets	

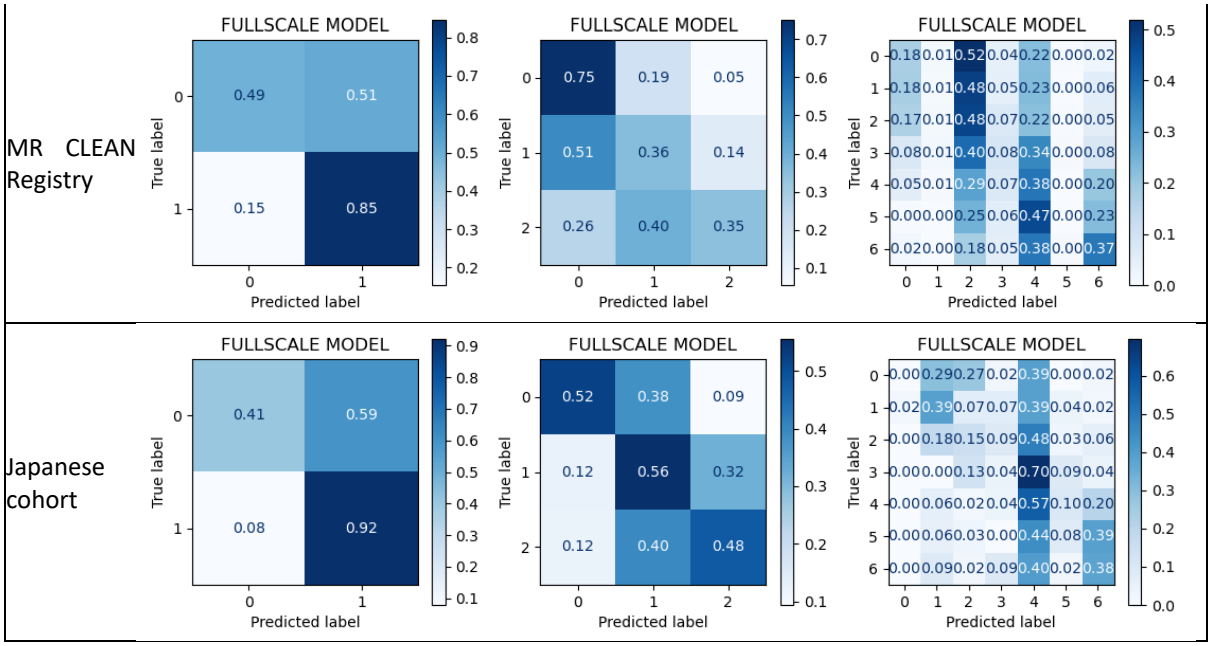
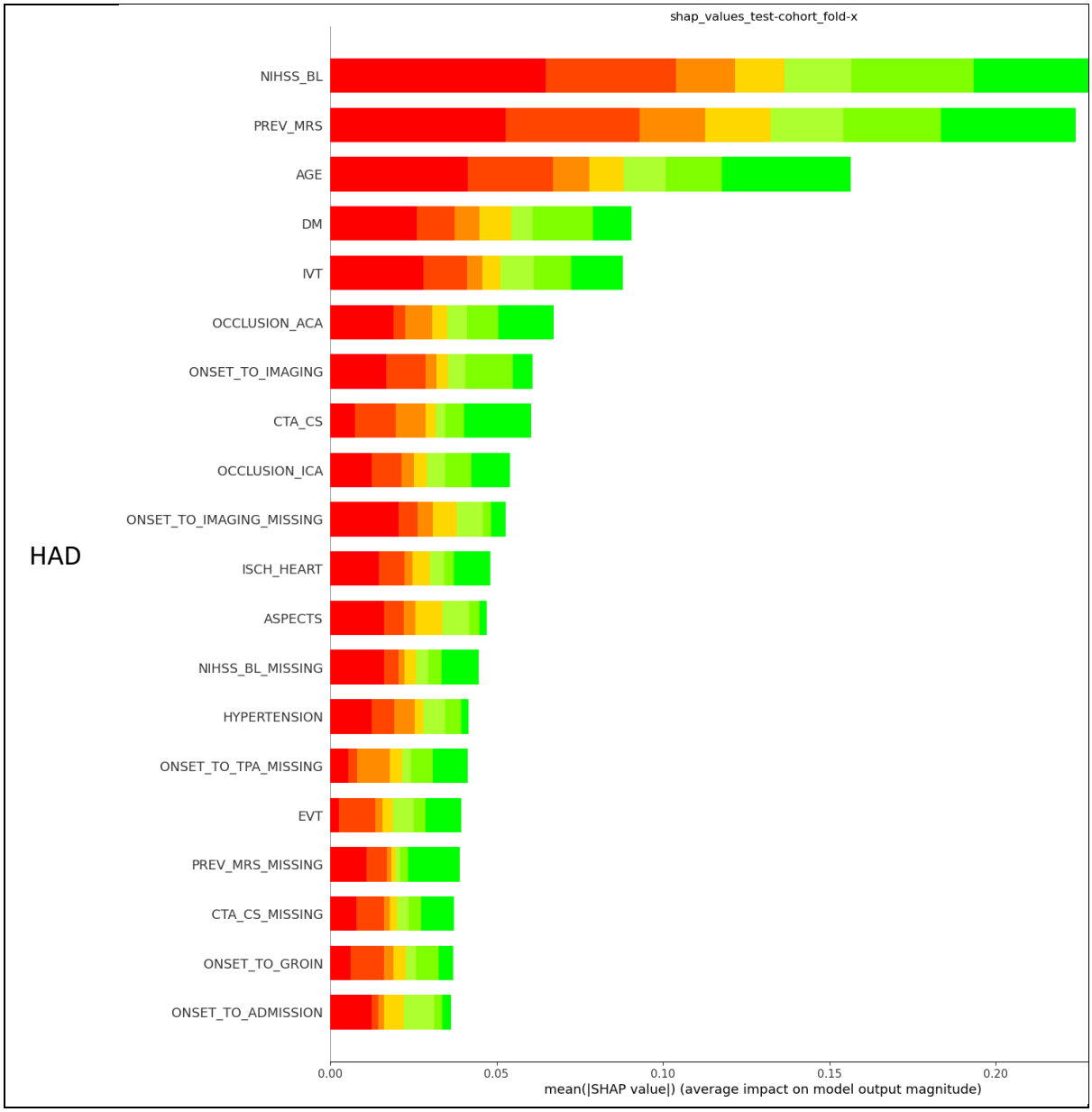


Figure 1. Confusion matrices in dichotomized, trichotomized and full-scale scenarios. First 4 rows show performance on derivation datasets (used for training, finetuning), last 2 rows show the first results on 2 the external validation datasets.


shap_values_test-cohort_fold-x
NIHSS_BL
PREV_MRS
AGE
DM
IVT
OCCLUSION_ACA
ONSET_TO_IMAGING
CTA_CS
OCCLUSION_ICA
ONSET_TO_IMAGING_MISSING
ISCH_HEART
ASPECTS
NIHSS_BL_MISSING
HYPERTENSION
ONSET_TO_TPA_MISSING
EVT
PREV_MRS_MISSING
CTA_CS_MISSING
ONSET_TO_GROIN
ONSET_TO_ADMISSION
0.00
0.05
0.10
0.15
0.20
mean(|SHAP value|) (average impact on model output magnitude)
HAD

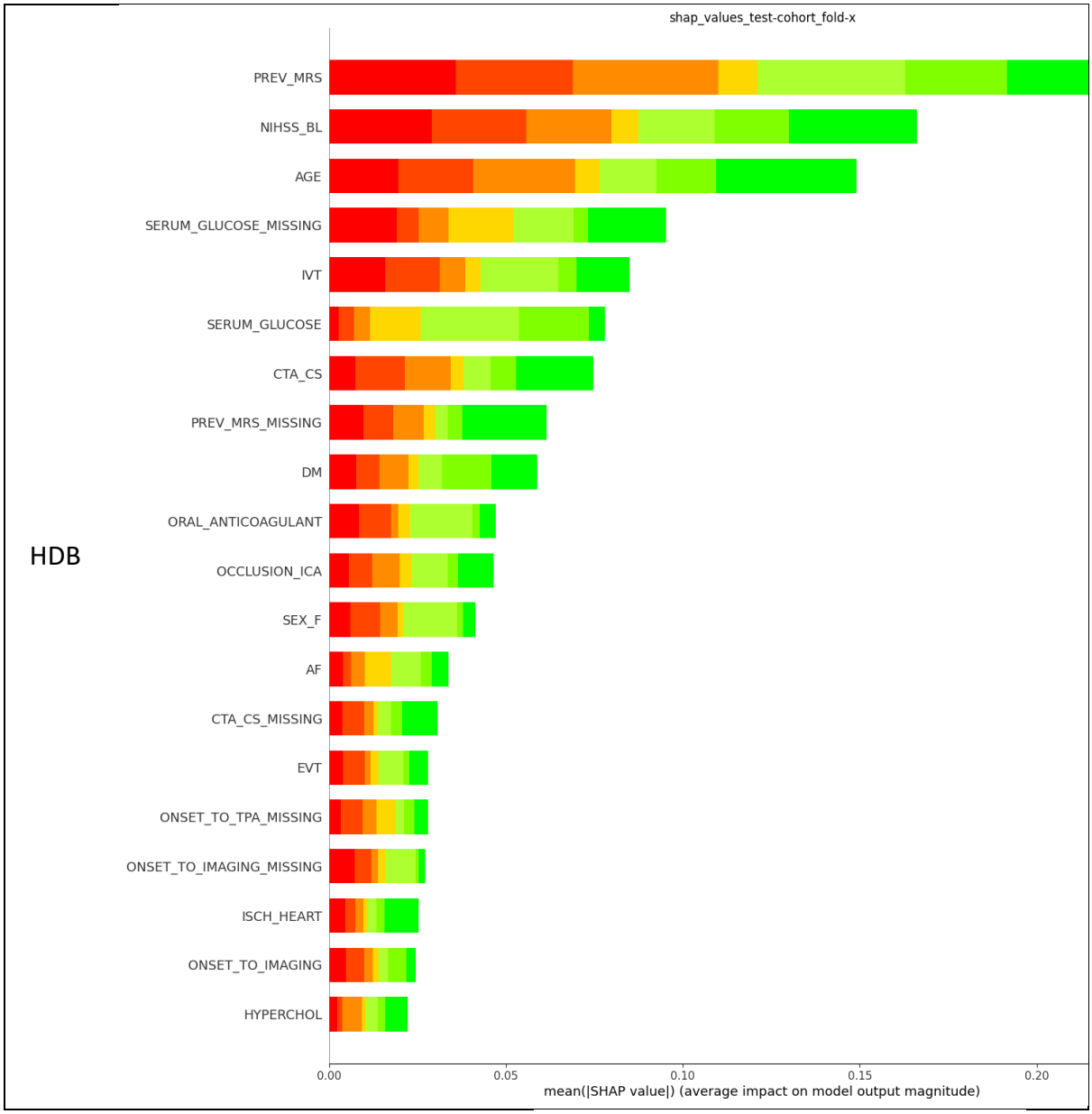


Figure 2. Feature importance ranking examples. Features are ordered by absolute magnitude of contribution and colors indicate the contribution for predicting the various classes of mRS.

DEEP LEARNING-ENABLED MENINGITIS SCREENING IN YOUNG INFANTS BASED ON A NOVEL NON-INVASIVE TRANSFONTANELLAR DEVICE: INITIAL PERFORMANCE RESULTS

Authors:

Sara Ajanovic[1,3], Beatrice Jobst[2], Javier Jiménez[2], Rita Quesada[2], Fabião Santos[2], Francesc Carandell[2], Manuela Lopez-Azorín[4], Eva Valverde[5], Marta Ybarra[5], M. Carmen Bravo[5], Paula Petrone[6], Hassan Sial[6], David Muñoz[7], Thais Agut[7], Barbara Salas[7], Nuria Carreras[8], Ana Alarcón[8], Martín Iriondo[8], Carles Luaces[7], Muhammad Sidat [9], Mastalina Zandamela [3], Paula Rodrigues[10], Dulce Graça[10], Sebastião Ngovene[10], Justina Bramugy [3], Anelsio Cossa[3], Campos Mucasse [3], W. Chris Buck[11], Sara Arias [1], Chaymae El Abbass [12,13], Najat Amalik[12,13], Imane Zizi[12,13], Houssain Tligui[12,13], Amina Barkat [12,13], Alberto Ibáñez[14], Montserrat Parrilla[14], Luis Elvira[14], Cristina Calvo [15,16,17,18], Adelina Pellicer[5,18], Fernando Cabañas*[4,19], Quique Bassat*[1,3,,20, 21,22]

Affiliations:

[1] Barcelona Institute for Global Health [ISGlobal] - Hospital Clínic, Universitat de Barcelona, Barcelona, (Spain)

[2] Kriba, Barcelona Science Park, Barcelona (Spain)

[3] Centro de Investigação em Saúde de Manhiça (CISM) Mozambique

[4] Department of Pediatrics and Neonatology Quironsalud Madrid University Hospital. Madrid (Spain)

[5] Neonatology Department, La Paz University Hospital - IdiPaz (Hospital La Paz Institute for Health Research), Madrid (Spain)

[6] Biomedical Data Science Team, Barcelona Institute for Global Health [ISGlobal], Barcelona, (Spain)

[7] Emergency department, Sant Joan de Déu Hospital, Institut de Recerca Sant Joan de Déu, Universitat de Barcelona, Barcelona (Spain)

[8] Neonatology department, Sant Joan de Déu Hospital, Institut de Recerca Sant Joan de Déu, Universitat de Barcelona, Barcelona (Spain)

[9] Universidade Eduardo Mondlane, Faculdade de Medicina, Maputo, (Mozambique)

[10] Maputo Central Hospital, Maputo, (Mozambique)

[11] University of California Los Angeles David Geffen School of Medicine, Los Angeles, (USA)

[12] Hôpital d'Enfant-Centre Hospitalier Universitaire Ibn Sina, Morocco

[13] Faculté de Médecine et Pharmacie, Université Mohammed V Rabat, Morocco

[14] Instituto de Tecnologías Físicas y de la Información (CSIC), Serrano 144, 28006 Madrid, (Spain).

[15] Pediatrics and Infectious Diseases Department, La Paz University Hospital, Fundación IdiPaz. Madrid (Spain).

[16] Biomedical Research Network Centre for Infectious Diseases (CIBERINFEC), Carlos III Health Institute, Madrid (Spain).

[17] Translational Research Network in Pediatric Infectious Diseases (RITIP), Madrid, (Spain)

[18] Universidad Autonoma de Madrid, Madrid (Spain).

[19] Biomedical Research Foundation, La Paz University Hospital-IDIPAZ, Madrid, Spain

[20] ICREA, Pg. Lluís Companys 23, 08010 Barcelona (Spain).

MANUSCRIPT

Introduction

Infant meningitis is a potentially life-threatening and disabling disease. Diagnosis typically requires a lumbar puncture (LP) to collect cerebrospinal fluid (CSF) for laboratory-based analysis[1]. The clinical presentation in young infants and newborns is often unspecific, hence making LPs part of a protocolized systematic screening approach in high-income settings. LPs are invasive, carry risks, are often blood-contaminated and most yield negative results, given the diseases' relatively low incidence[2,3]. However, in low-income settings, LPs and CSF exams are seldomly feasible, and suspected meningitis cases are mainly treated empirically[4]. This study aims to validate a novel, non-invasive transfontanellar CSF white blood cell (WBC) counter, used to classify between low and high levels of WBCs, to screen for meningitis, employing high-resolution ultrasound (HRUS) and deep learning (DL).

Material and methods

We enrolled patients under 24 months with suspected meningitis and an open anterior fontanelle who underwent LP within 24h from study enrolment, in 5 hospitals (3 Spanish, one Mozambican and one Moroccan) (2020-2023), after obtaining consent from legal guardians. We obtained images showing the backscatter pattern from CSF using a customized HRUS probe[5] (Neosonics®), positioned over the open anterior fontanelle region of the infants' head (Figure 1A). A DL model based on the Resnet50 architecture[6] was trained to classify CSF patterns according to WBC values obtained through the LP, setting a 30 cells/mm^3 threshold to differentiate controls from cases on an individual image-level. Patient-level results were obtained applying a soft voting strategy, by computing the maximum of the two median class probabilities over individual images (Figure 1B).

Results

We collected 2237 images from 34 patients with paired LPs WBC count results (10 $\geq$30 cells/mm^3, 24 <30 cells/mm^3) from Spanish and Mozambican hospitals to train the algorithm. Employing a leave-one-out methodology[7] for patient-level classification, we tested each patient while training the model on the remaining data. Additionally, for independent model testing, we gathered 1545 images from 42 patients alongside paired LPs results (8 $\geq$30 cells/mm^3, 34 <30 cells/mm^3) from the Moroccan hospital. The DL model achieved 94.7% accuracy with 94.4% sensitivity and 94.8% specificity for WBC levels below/above 30 cells/mm^3 at the patient level. When analysing the cohorts separately we obtained the following results: Spanish cohort: sensitivity/specificity 100%/80%; Mozambican cohort: sensitivity/specificity 100%/100%; Moroccan cohort: sensitivity 87.5 and specificity 97.1% (Table 1). The median fontanel thickness positive cases all cohorts is 3.75mm (IQR 3.43-4.5mm). Single image frames probabilities and cohort median values are shown in Figure 1C.

[21] Pediatrics Department, Hospital Sant Joan de Déu, Universitat de Barcelona, Barcelona (Spain)
[22] CIBER de Epidemiología y Salud Pública, Instituto de Salud Carlos III

Presenting author:

Sara Ajanovic Andelic

Barcelona Institute for Global Health (ISGlobal)

C/Rosselló 132, 5a (08036) Barcelona, Spain

sara.ajanovic@isglobal.org

+34 665535029

Keywords:

Infant Meningitis, Ultrasound, Deep Learning, Non-invasive Screening

Key information:

1. Research question: This study validates a novel, non-invasive transfontanellar cerebrospinal fluid (CSF) white blood cell counter for meningitis screening, employing high-resolution ultrasound (HRUS) and deep learning (DL).
2. Findings: Applying the DL model on images of CSF backscatter patterns on 76 patients (18 cases, 58 controls) we achieved 94.7% accuracy with 94.4% sensitivity and 94.8% specificity at the patient-level.
3. Meaning: We show the potential of HRUS and DL as a non-invasive screening method in young infants and neonates to modulate meningitis suspicion and indications for lumbar puncture, the current gold standard for meningitis diagnosis.

Discussion and Conclusion

These results show the potential utility of HRUS and DL as a screening method based on accurate CSF WBC level classification to modulate indications for LPs among neonates and young infants. While this encourages further validation of the device efficacy, its application in standard practice could have potentially spared up to 94.8% of negative LPs and detected 94.4% of positive cases. The single false negative case with 40 cells/mm^3 after correction for red blood cells (RBCs) contamination had a particularly thick fontanelle (4.8mm), possibly impacting sensitivity at such low WBC levels. Potential factors such as age and CSF volume or the thickness of the fontanel may affect signal saturation, as well as counting accuracy, respectively. These potential limitations guide future development to maintain sensitivity and robustness in those most challenging cases.

In conclusion, this technology could potentially be used as a quick and non-invasive screening tool in young infants and neonates modulating meningitis suspicion and indications for LPs. This could ultimately lead to a more agile meningitis diagnosis especially in low- and middle-income settings, allowing a more efficient use of resources and improve patient outcomes at both the individual and global health levels.

References

1. Obaro SK. Updating the diagnosis of bacterial meningitis. *Lancet Infectious Diseases*. 2019;19(11):1160-1161. doi:10.1016/s1473-3099(19)30549-3
2. Dalai R, Dutta S, Pal A, Sundaram V, Jayashree M. Is Lumbar Puncture Avoidable in Low-Risk Neonates with Suspected Sepsis? *American Journal of Perinatology.* 2020;39(01):099-105. doi:10.1055/s-0040-1714397
3. Bedetti L, Marrozzini L, Baraldi A, et al. Pitfalls in the diagnosis of meningitis in neonates and young infants: the role of lumbar puncture. *Journal of Maternal-fetal and Neonatal Medicine.* 2018;32(23):4029-4035. doi:10.1080/14767058.2018.1481031
4. Gudina EK, Tesfaye M, Adane A, et al. Challenges of bacterial meningitis case management in low income settings: an experience from Ethiopia. *TM & IH Tropical Medicine and International Health.* 2016;21(7):870-878. doi:10.1111/tmi.12720
5. Jiménez X, Shukla SK, Ortega I, et al. Quantification of very low concentrations of leukocyte suspensions in vitro by High-Frequency Ultrasound. *Ultrasound in Medicine & Biology.* 2016;42(7):1568-1573. doi:10.1016/j.ultrasmedbio.2016.01.027
6. Deep residual learning for image recognition. *IEEE Conference Publication | IEEE Xplore.* https://ieeexplore.ieee.org/document/7780459. Published June 1, 2016.

7. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning. Vol 2. *New York: Springer.* 2009. doi:10.1007/978-0-387-84858-7

Disclosures

B.J., J.J., R.Q., F.C., and F.S. are part of the team developing the technology presented in this abstract.

S.A., A.P., F.C., C.C., P.P. and Q.B. are clinical assessors for the company developing the technology.

APPENDIX

Acknowledgements

Mozambique: This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation INV-048197. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Morocco: Project PI16/01822 from PI ENRIQUE BASSAT ORELLANA, funded by Instituto de Salud Carlos III and co-funded by the European Union (FEDER) “Una manera de hacer Europa”. This publication has been funded by the Spanish Agency of Cooperation (AECID). Its content is Sara Ajanovic et al’s responsibility and does not necessarily reflect AECID's opinion. Kriba acknowledges support from the European Union’s Horizon Europe research and innovation program under project code 190155553 - NEOSONICS.

Figures

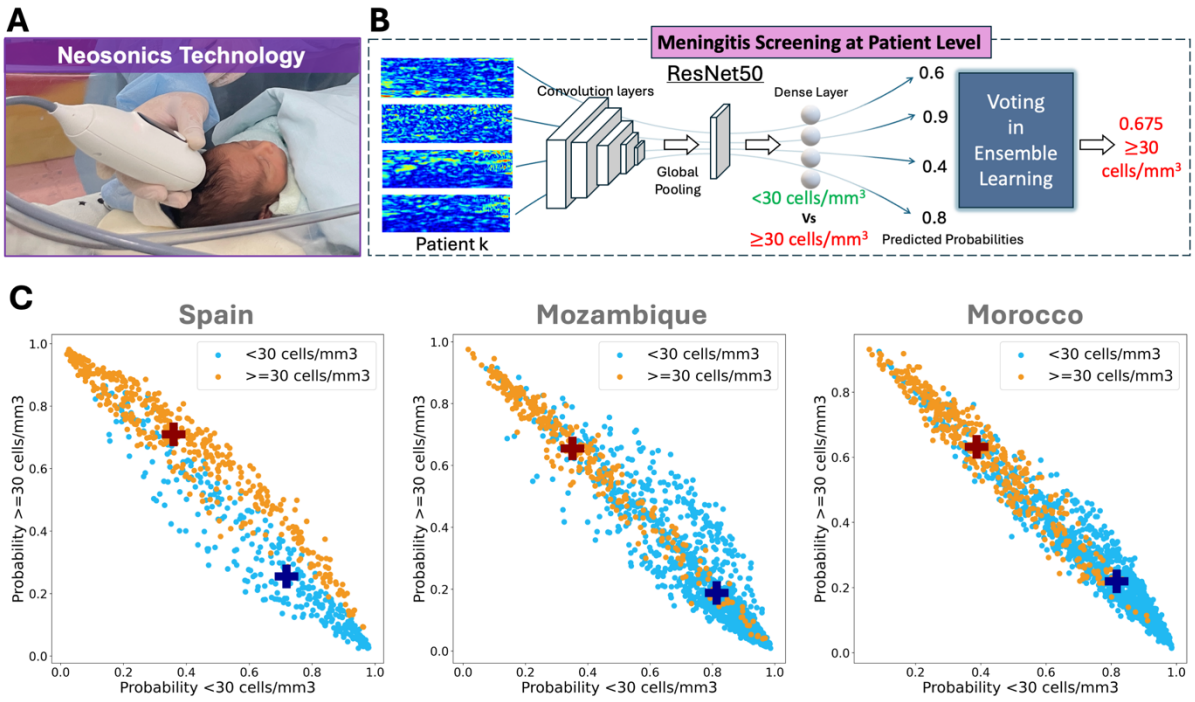


Figure 1: A. High-resolution ultrasonic (HRUS) probe (Neosonics®), positioning over the infants' head **B.** Deep learning binary classification model to distinguish between CSF images containing </≥ 30 cells/mm^3 and voting strategy for patient-level results. **C.** Single image frames probability for <30 cells/mm^3 images (blue) being classified as <30 cells/mm^3 or ≥30 cells/mm^3 and for ≥30 cells/mm^3 images (orange) being classified as <30 cells/mm^3 or ≥30 cells/mm^3 for Spanish, Mozambican and Moroccan cohorts. Cohort median values are represented as blue and dark orange crosses, respectively.

Tables

Table 1: Participants' clinical information and DL algorithm results. False classifications are marked in red.

Patient	Hospital	Age (days)	Weight (grams)	Sex	CSF appearance	Corrected WBC	CSF micro-biology	Blood micro-biology	**Result of the HRUS device**	# of image frames	Correct class probability
1	HULP	4	2035	F	turbulent	53	negative	*S. Hominis*	**≥30 WBC**	31	0.56
2	HULP	63	5000	F	clear	85	negative	*S. Agalactiae*	**≥30 WBC**	24	0.87
3	HULP	9	3900	F	clear	520	Enterovirus	negative	**≥30 WBC**	44	0.93

4	HULP	22	3370	M	turbulent	2400	Klebsiella	negative	**≥30 WBC**	50	0.89
5	HQM	93	4900	M	turbulent	300	Streptococcus agalactiae	*S. Agalactiae*	**≥30 WBC**	28	0.68
6	HQM	32	3600	M	turbulent	430	Streptococcus agalactiae	negative	**≥30 WBC**	36	0.83
7	HQM				turbulent	169	negative	negative	**≥30 WBC**	233	0.61
Spanish >30 WBC		28.5 [9-33]	4235 [3385-4775]			300 [153-558]					
8	HULP	64	4420	F	clear	7	negative	negative	**<30 WBC**	11	0.87
9	HULP	3	3290	F	clear	3	negative	negative	**<30 WBC**	19	0.76
10	HQM	16	3120	M	clear	3	negative	negative	**≥30 WBC**	15	0.38
11	HSJD	17	2965	M	clear	15	negative	negative	**<30 WBC**	14	0.81
12	HSJD	4	3890	F	clear	5	negative	negative	**<30 WBC**	24	0.73
13	HSJD	5	2690	F	hematic	9	negative	negative	**≥30 WBC**	34	0.36
14	HSJD	32	4000	M	hematic	18	negative	negative	**<30 WBC**	81	0.71
15	HSJD	61	4150	M	clear	0	negative	negative	**<30 WBC**	18	0.83
16	HSJD	337	8600	M	clear	11	negative	negative	**<30 WBC**	55	0.95
17	HSJD	23	4220	F	clear	0	negative	negative	**<30 WBC**	64	0.60
Spanish ≤30 WBC		17 [5-44.5]	3925 [3095-4925]			6 [3-12]					
Total Spanish participants		23 [8-33.5]	3922.5 [3035-4665]								
18	HCM	44	4200	F	NA	3520	Haemophilus spp	Hemophilus spp	**≥30 WBC**	140	0.71
19	HCM	40	4000	Male	hematic	25	negative	negative	**<30 WBC**	39	0.58
20	HCM	49	2400	F	turbulent	19200	negative	negative	**≥30 WBC**	57	0.61
21	HCM				turbulent	260	negative	NA	**≥30 WBC**	107	0.63
Mozambican >30 WBC		44 [40-44.5]	4000 [3400-4300]			1435 [260-7890]					
22	HCM	52	3500	M	clear	0	negative	negative	**<30 WBC**	45	0.91
23	HCM	311	5800	M	clear	0	negative	negative	**<30 WBC**	35	0.94
24	HCM	48	3800	F	clear	0	negative	negative	**<30 WBC**	116	0.86
25	HCM	137	7700	M	clear	0	negative	negative	**<30 WBC**	108	0.95
26	HCM	41	4600	M	clear	0	negative	negative	**<30 WBC**	69	0.89
27	HCM	197	7500	M	clear	0	negative	negative	**<30 WBC**	108	0.81
28	HCM	124	4400	F	clear	0	negative	negative	**<30 WBC**	101	0.88
29	HCM	32	5200	M	clear	0	negative	negative	**<30 WBC**	86	0.56
30	HCM	13	2900	F	clear	0	negative	negative	**<30 WBC**	22	0.67
31	HCM	3	4200	F	clear	0	negative	negative	**<30 WBC**	91	0.86

32	HCM	121	9200	M	clear	0	negative	negative	**<30 WBC**	123	0.78
33	HCM	131	7400	F	clear	0	negative	negative	**<30 WBC**	182	0.63
34	HCM	42	3600	M	clear	0	negative	negative	**<30 WBC**	27	0.76
Mozambican ≤30 WBC		52 [32 - 140]	4600 [3800-6950]			0 [0-0]					
Total Mozambican participants		47.5 [33-100.5]	4200 [3500-5862.5]								
35	HER	4	2180	M	turbulent	3725	*Serratia marcesens*	NA	**≥30 WBC**	60	0.73
36	HER	14	1210	F	clear	51	negative	contamination	**≥30 WBC**	49	0.70
37	HER	30	1315	F	clear	53	negative	negative	**≥30 WBC**	73	0.67
38	HER	28	3130	M	clear	39	*Human Herpes Virus 6*	negative	**≥30 WBC**	93	0.59
39	HER	10	1220	M	hematic	170	negative	negative	**≥30 WBC**	9	0.57
40	HER	56	4450	M	hematic	660	negative	NA	**≥30 WBC**	10	0.50
41	HER	2	4950	M	hematic	254	negative	NA	**≥30 WBC**	76	0.64
42	HER	3	3880	M	hematic	40	negative	NA	**<30 WBC**	33	0.28
Moroccan >30 WBC		12 [4-28]	2655 [1220-4450]			212 [2-46]					
43	HER	1	3200	F	xantochromic	3	negative	NA	**<30 WBC**	31	0.93
44	HER	23	3900	M	clear	7	negative	negative	**<30 WBC**	22	0.93
45	HER	0	3700	M	hematic	0	negative	negative	**<30 WBC**	71	0.86
46	HER	25	3500	M	xantochromic	8	negative	NA	**<30 WBC**	35	0.96
47	HER	1	3150	F	clear	10	negative	NA	**<30 WBC**	16	0.93
48	HER	2	1900	M	xantochromic	28	negative	negative	**<30 WBC**	24	0.75
49	HER	2	3120	M	clear	4	negative	negative	**<30 WBC**	9	0.96
50	HER	2	2800	M	hematic	7	negative	negative	**<30 WBC**	67	0.89
51	HER	6	1940	M	xantochromic	4	negative	positive	**<30 WBC**	29	0.87
52	HER	2	3700	M	hematic	11	*Haemophilus influenzae*	negative	**<30 WBC**	49	0.73
53	HER	1	2400	M	xantochromic	3	negative	negative	**<30 WBC**	12	0.85
54	HER	11	3029	M	xantochromic	8	negative	NA	**<30 WBC**	57	0.88
55	HER	4	1560	F	xantochromic	15	negative	negative	**<30 WBC**	30	0.83
56	HER	21	1350	M	clear	2	negative	negative	**<30 WBC**	7	0.95
57	HER	4	1395	F	hematic	4	negative	contamination	**<30 WBC**	53	0.74
58	HER	2	1350	M	hematic	0	negative	negative	**<30 WBC**	9	0.92

59	HER	2	3500	M	hematic	2	negative	negative	**<30 WBC**	13	0.88
60	HER	93	2600	F	xantochromic	5	*Staphylococcus aureus*	negative	**<30 WBC**	9	0.91
61	HER	2	3960	M	xantochromic	9	contaminated	contamination	**<30 WBC**	53	0.86
62	HER	3	3550	M	xantochromic	8	negative	NA	**<30 WBC**	18	0.78
63	HER	8	1013	F	hematic	14	negative	contamination	**<30 WBC**	22	0.54
64	HER	3	3050	M	xantochromic	4	negative	negative	**<30 WBC**	6	0.60
65	HER	5	1520	F	xantochromic	1	negative	negative	**<30 WBC**	16	0.73
66	HER	2	4185	M	xantochromic	2	negative	negative	**<30 WBC**	22	0.91
67	HER	2	3880	M	hematic	0	negative	negative	**≥30 WBC**	24	0.45
68	HER	17	1580	F	hematic	0	negative	NA	**<30 WBC**	10	0.61
69	HER	2	4800	M	xantochromic	25	negative	negative	**<30 WBC**	50	0.76
70	HER	6	2175	M	xantochromic	14	*Klebsiella pneumonaie*	negative	**<30 WBC**	49	0.71
71	HER	17	3700	F	hematic	1	*Entérovirus*	negative	**<30 WBC**	19	0.69
72	HER	3	3900	M	hematic	14	negative	negative	**<30 WBC**	18	0.55
73	HER	2	3760	M	hematic	0	negative	negative	**<30 WBC**	36	0.66
74	HER	7	4450	M	hematic	29	negative	negative	**<30 WBC**	144	0.75
75	HER	3	2870	M	xantochromic	1	negative	negative	**<30 WBC**	43	0.58
76	HER	13	3190	M	xantochromic	2	negative	negative	**<30 WBC**	69	0.85
Moroccan ≤30 WBC		3 [2-17]	3085 [1520-3900]			2 [8-24]					
Total Moroccan participants		3.5 [2-19]	3125 [1740-3925]								

Preoperative Rotator Cuff Tear Prediction from Shoulder Radiographs using a Convolutional Block Attention Module-Integrated Neural Network

Authors:

Chris Hyunchul Jo[1†], Jiwoong Yang[2†], Byunghwan Jeon[3], Hackjoon Shim[4, 5], Ikbeom Jang[3]*

Affiliations:

[1] Department of Orthopedic Surgery, Seoul National University College of Medicine, SMG-SNU Boramae Medical Center, Seoul, South Korea

[2] Artificial Intelligence Semiconductor, Hanyang University, Seoul, South Korea

[3] Division of Computer Engineering, Hankuk University of Foreign Studies, Yongin, South Korea

[4] Medical Imaging AI Research Center, Canon Medical Systems Korea, Seoul, South Korea

[5] ConnectAI Research Center, Yonsei University College of Medicine, Seoul, South Korea

Presenting author:

Full name: Chris Hyunchul Jo

Affiliation: Department of Orthopedic Surgery, Seoul National University College of Medicine, SMG-SNU Boramae Medical Center, Seoul, South Korea

Email: chrisjo@snu.ac.kr

Keywords:

Rotator cuff tear, Shoulder radiographs, Deep learning, CBAM

Key information:

Research question: We test whether a plane shoulder radiograph can be used together with deep learning methods to identify patients with rotator cuff tears as opposed to using an MRI in standard of care.

Findings: By integrating convolutional block attention modules into a deep neural network, our model demonstrates high accuracy in detecting patients with rotator cuff tears, achieving an average AUC of 0.889 and an accuracy of 0.831.

Meaning: This study validates the efficacy of our deep learning model to accurately detect rotation cuff tears from radiographs, offering a viable pre-assessment or alternative to more expensive imaging techniques such as MRI.

†: equal contribution, *: corresponding author

MANUSCRIPT

Introduction

Initial radiograph evaluations often fail to identify soft tissue injuries such as rotator cuff tears. It necessitates further imaging with more expensive MRI examinations, increasing healthcare costs. In this study, we show that a convolutional neural network with channel attention and spatial attention modules can significantly enhance the accuracy of rotator cuff tear detection, only using a single shoulder radiograph. All shoulder radiograph data used for training the deep learning model were collected from our local clinic. It may offer a viable pre-assessment or alternative to more expensive imaging techniques such as MRI.

Material and methods

Data

We retrospectively collected a dataset of shoulder radiographs from 99 patients from our clinic. The dataset comprises 50 patients with full-thickness rotator cuff tears (fRCT) and 49 without tears. We acquired radiographs in four angles – axial, glenoid, outlet, and anteroposterior (AP) – totaling 396 images. All the images were acquired before surgery. Regions of interest (ROIs) essential for fRCT diagnosis were identified and annotated with bounding boxes on all images. These annotations were used to train the YOLO v5 model to automatically crop ROIs from all radiographs, and to efficiently handle the annotation process for all future data. The results fo ROI extraction are shown in Figure 2 and Table 1. The dataset was divided based on subject IDs with each containing four view-specific images. We applied 5-fold cross-validation due to the limited size of the dataset, resulting in 316 training images from 79 subjects and 80 test images from 20 subjects, ensuring no subject overlap between folds.

Network Architecture and Training

The ROIs extracted were further processed using Contrast-Limited Adaptive Histogram Equalization (CLAHE) to enhance bone structures and edge visibility, facilitating more detailed recognition of fractures and structural integrity. We employed a ResNet50 model applied with a Convolutional Block Attention Module (CBAM) to diagnose rotator cuff tears. CBAM enhances the model's learning focus on essential features in medical image by sequentially applying channel attention and spatial attention. This method allows for a concentrated analysis of salient features crucial for accurately diagnosing rotator cuff tears. The pretrained ResNet50 was adapted to classify between the presence and absence of fRCT, with only two output classes in the final layer. The architecture of the model is shown in Figure 4. Due to the limited dataset of 99 subjects, we employ k-fold cross-validation to enhance the accuracy of the model's performance. Each fold in the k-fold cross-validation setup was balanced to have an equal ratio of fRCT and no-tear cases in both the train and validation datasets.

Implementation Details

We trained models on an NVIDIA RTX 3090 GPU using the SGD optimizer with a learning rate of 0.01 and a batch size of 8. To prevent overfitting, we employed various data augmentation techniques including rotation, horizontal flipping, random crop, scaling, translation, brightness adjustment, and inversion, as well as implementing a dropout rate of 0.2. All data was resized to 512x512 pixels prior to training. We

utilized a CrossEntropyLoss for the loss function and a CosineAnnealingWarmupRestarts scheduler to dynamically adjust the learning rate during the training process.

Results

Using k-fold cross-validation, we evaluated the performance of our model across each fold. The average accuracy achieved was 0.831 (329/396). The AUROC for the two classes was 0.889, indicating a high level of discriminative ability. The Positive Predictive Value (PPV) was 0.852 (161/189), and the Negative Predictive Value (NPV) was 0.812 (168 /207). This is shown in Figure 5. The findings demonstrate that radiographs alone can effectively classify patients with fRCT, underscoring its potential utility in diagnostic settings.

Discussion and Conclusion

Our proposed method demonstrates that radiographs alone can effectively diagnose fRCT. This approach calculates the probability of rotator cuff tears, thereby assisting in the decision-making process of whether to proceed with MRI imaging. To increase the generalizability of the model, we plan to collaborate with multiple centers to incorporate a larger and more diverse set of radiograph data. Future research will aim to expand the dataset, which was collected from our clinic, and analyze the impact of different radiographic views (axial, glenoid, outlet, and AP) on the diagnosis of rotator cuff tears to improve the current model's performance. In the long term, we aim to achieve a level of diagnostic accuracy close to that of MRI-based diagnostics for rotator cuff tears.

References

1. Youngjune K, Choi D, Lee KJ, et al. Ruling out rotator cuff tear in shoulder radiograph series using deep learning: Redefining the role of conventional radiograph. Eur Radiol. 2020;30(5):2843-2852.

2. Touvron H, Vedaldi A, Douze M, Jégou H. Fixing the train-test resolution discrepancy. Adv Neural Inf Process Syst. 2019;32.

3. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018;3-19.

4. SSelvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. The Institute of Electrical and Electronics Engineers, Inc.(IEEE) Conference Proceedings. 2017:618-626.

5. Cho Y, Jalics A, Lv D, et al. Predicting rotator cuff tear severity using radiographic images and machine learning techniques. In: Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition. 2021;237-241.

6. Mery D, Mery D. X-ray Testing. In: Computer Vision for X-Ray Testing: Imaging, Systems, Image Databases, and Algorithms. 2015;1-33.

7. Shariatnia MM, Ramazanian T, Sanchez-Sotelo J, Maradit Kremers H. Deep learning model for measurement of shoulder critical angle and acromion index on shoulder radiographs. JSES reviews, reports, and techniques. 2022;2(3):297-301.

APPENDIX

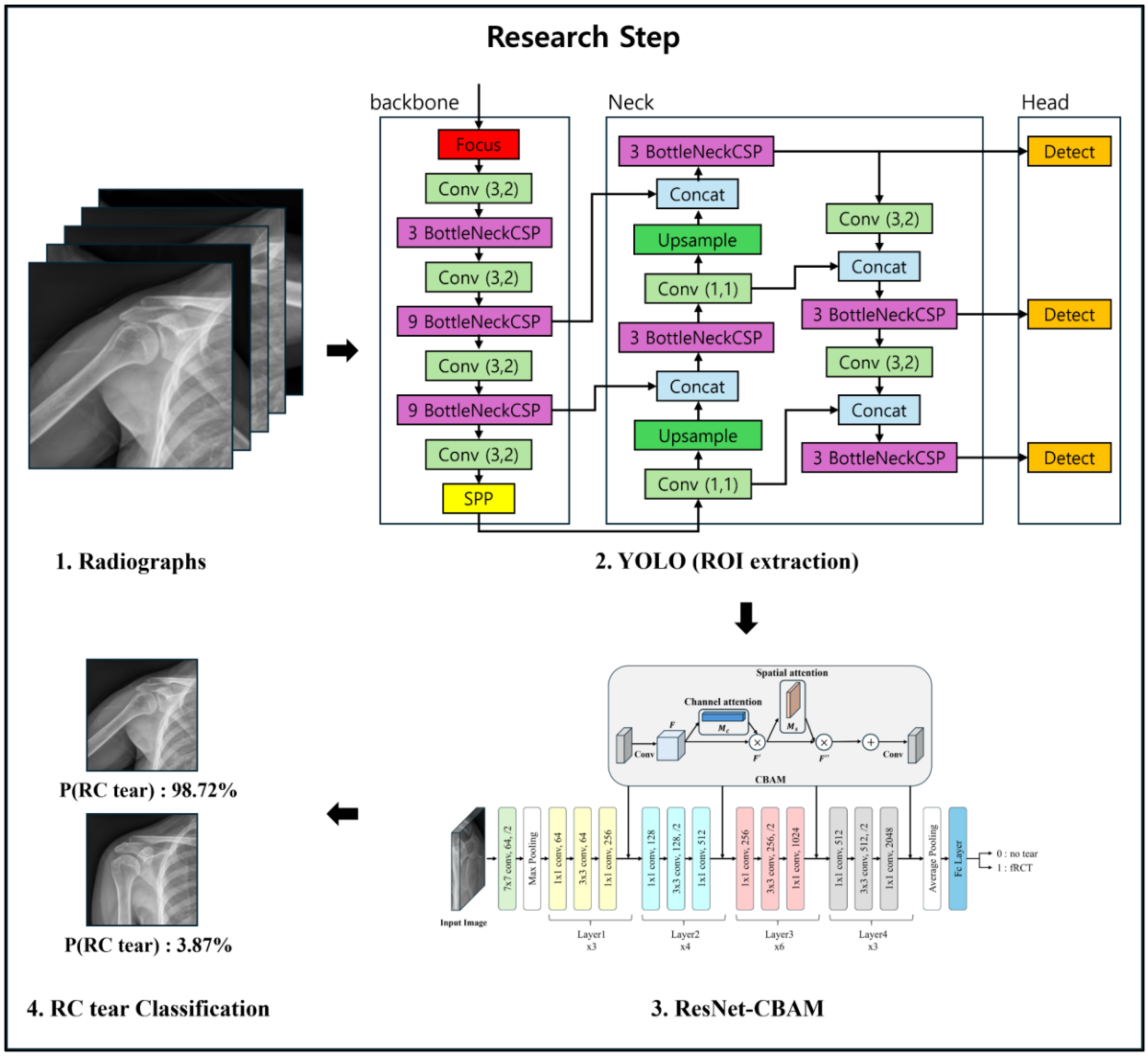


Figure 1: Stages of a study for diagnosing rotator cuff tears using radiographs. The radiographs are processed through Yolo v5 to extract essential Regions of Interest (ROI). These are then used to train the ResNet-CBAM model, enabling the classification of patients with rotator cuff tears. Our study proposes a diagnostic framework that predicts the presence of rotator cuff tears using only radiographs, a more cost-effective and accessible option. The proposed method calculates the probability of rotator cuff tears in patients, thereby assisting in the decision-making process of whether to proceed with MRI imaging.

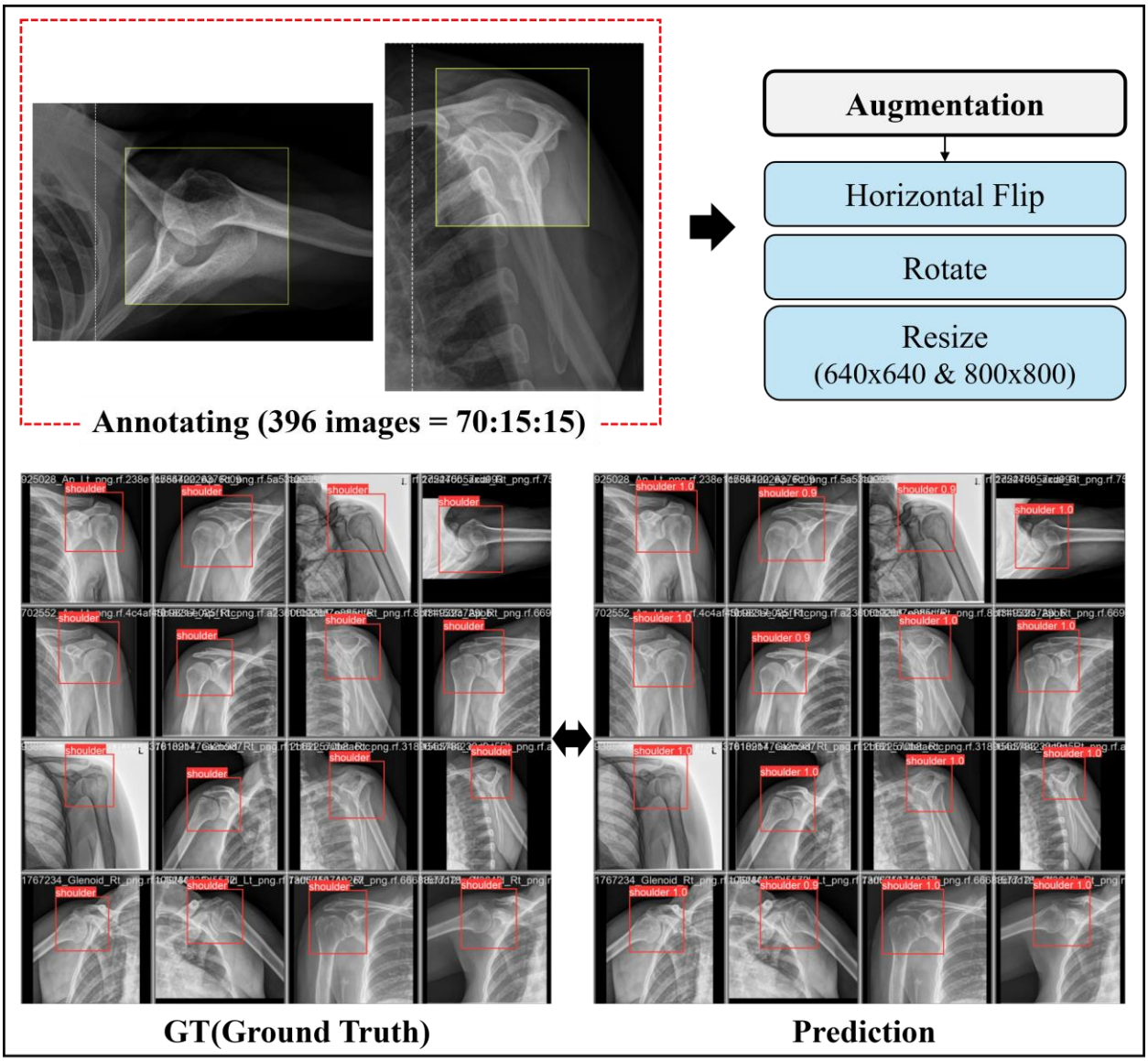


Figure 2: Training YOLO v5 for ROI Extraction. We conducted the process of annotating 396 radiograph images from 99 subjects, followed by training the YOLO v5 model to detect regions of interest (ROI). The data was divided into training, validation, and test sets in a 70:15:15 ratio. The annotated data was subjected to augmentation processes including horizontal flip, rotation, and resizing. For resizing, the original aspect ratio of the images was maintained, and the remaining space was filled with black edges. Two versions of the model were trained: version 1 (exp1) with images resized to 640x640 pixels and version 2 (exp2) with images resized to 800x800 pixels. The best-performing version, exp 2, was selected based on its superior performance in ROI extraction as indicated in Table 1. The ground truth (GT) and prediction results illustrate that the YOLO v5 model successfully detected th ROIs with high accuracy, demonstrating its effectiveness in this application.

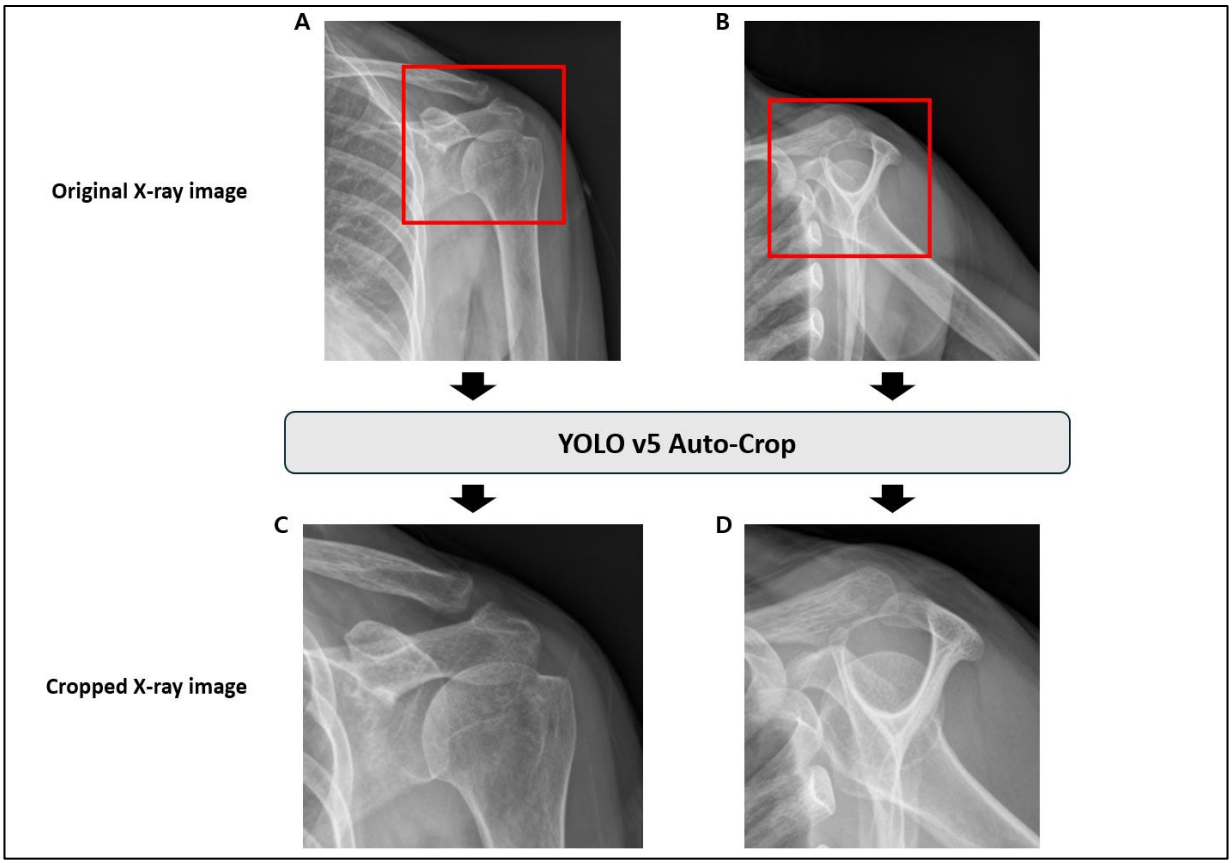


Figure 3: Automated ROI Cropping in radiographs Using YOLO v5. We train YOLO v5 with data labeled with bounding boxes for the ROI. When the saved best-weights (exp2) are loaded and applied, all radiographic data in the dataset are automatically cropped to include only the ROI. For the diagnosis of fRCT, specific regions of interest (ROIs) were cropped from the original-sized radiographs A and C to produce images B and D, respectively. These particular ROIs were then utilized as training data. This process allows the model to more effectively identify the features of full-thickness Rotator Cuff Tears (fRCT) by excluding unnecessary regions during training. Additionally, it reduces variations in image size and positioning that can occur depending on the individual who performed the radiograph, thereby aiding in generalization.

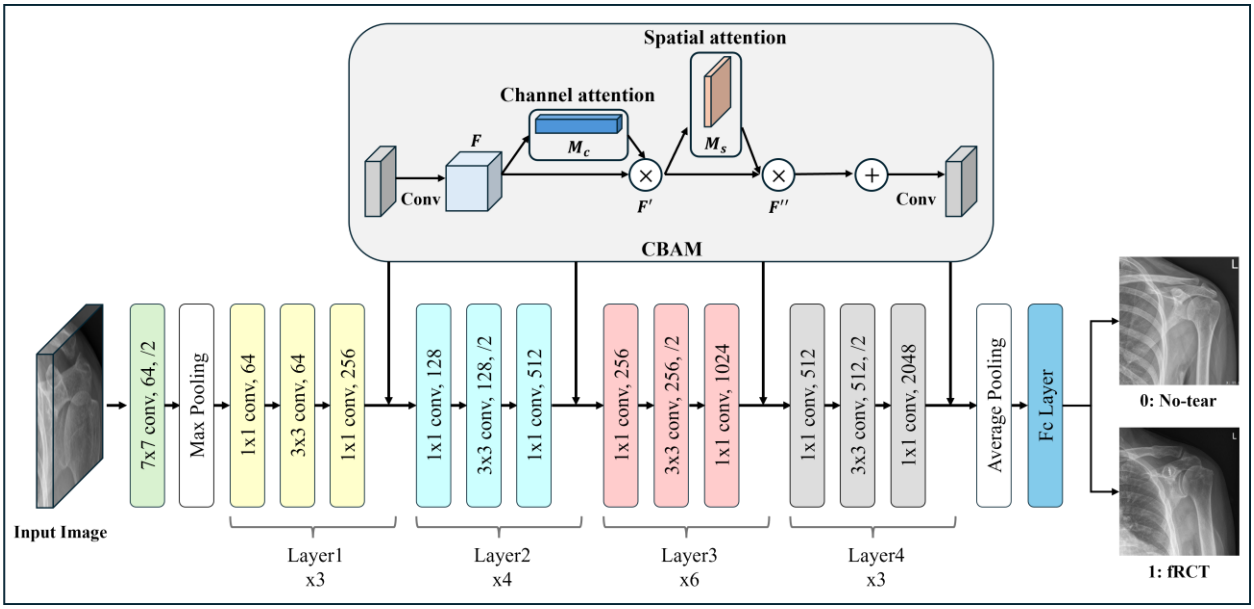


Figure 4: ResNet50-CBAM model structure. This figure is ResNet50 architecture with CBAM applied, enhancing feature representation by applying sequential channel and spatial attention mechanisms. It aids in enhancing performance by focusing on features necessary for the diagnosis of fRCT.

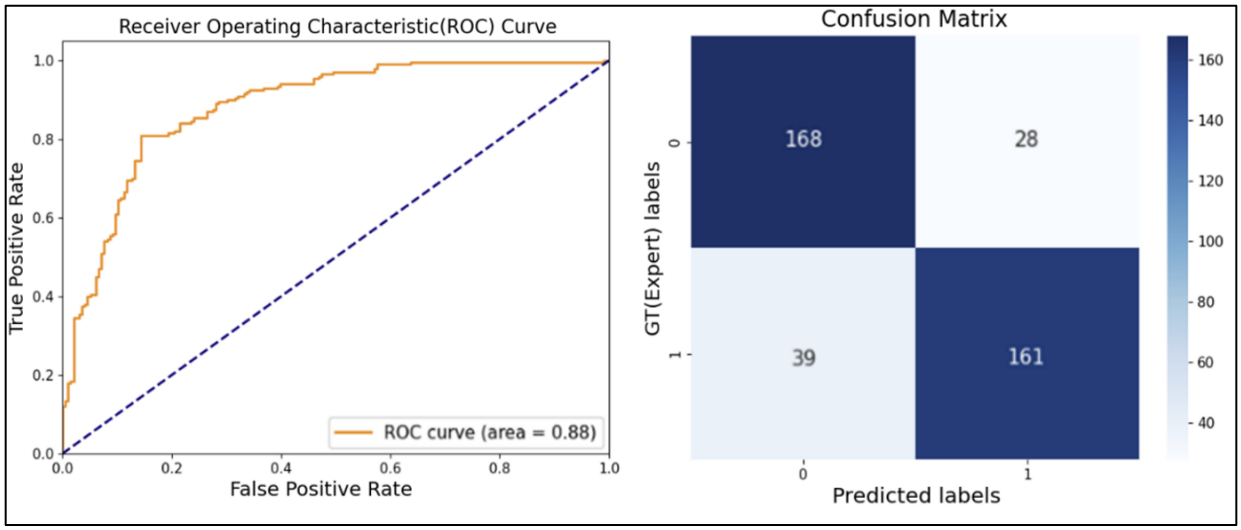


Figure 5: ROC curves(A) and confusion matrix(B) on the test set. In the confusion matrix, 0 represents 'no-tear' and 1 indicates 'fRCT'. During k-fold cross-validation, the cumulative results predicted by the model in each fold's test phase are presented in a confusion matrix. This is useful for assessing the average performance metrics of the dataset.

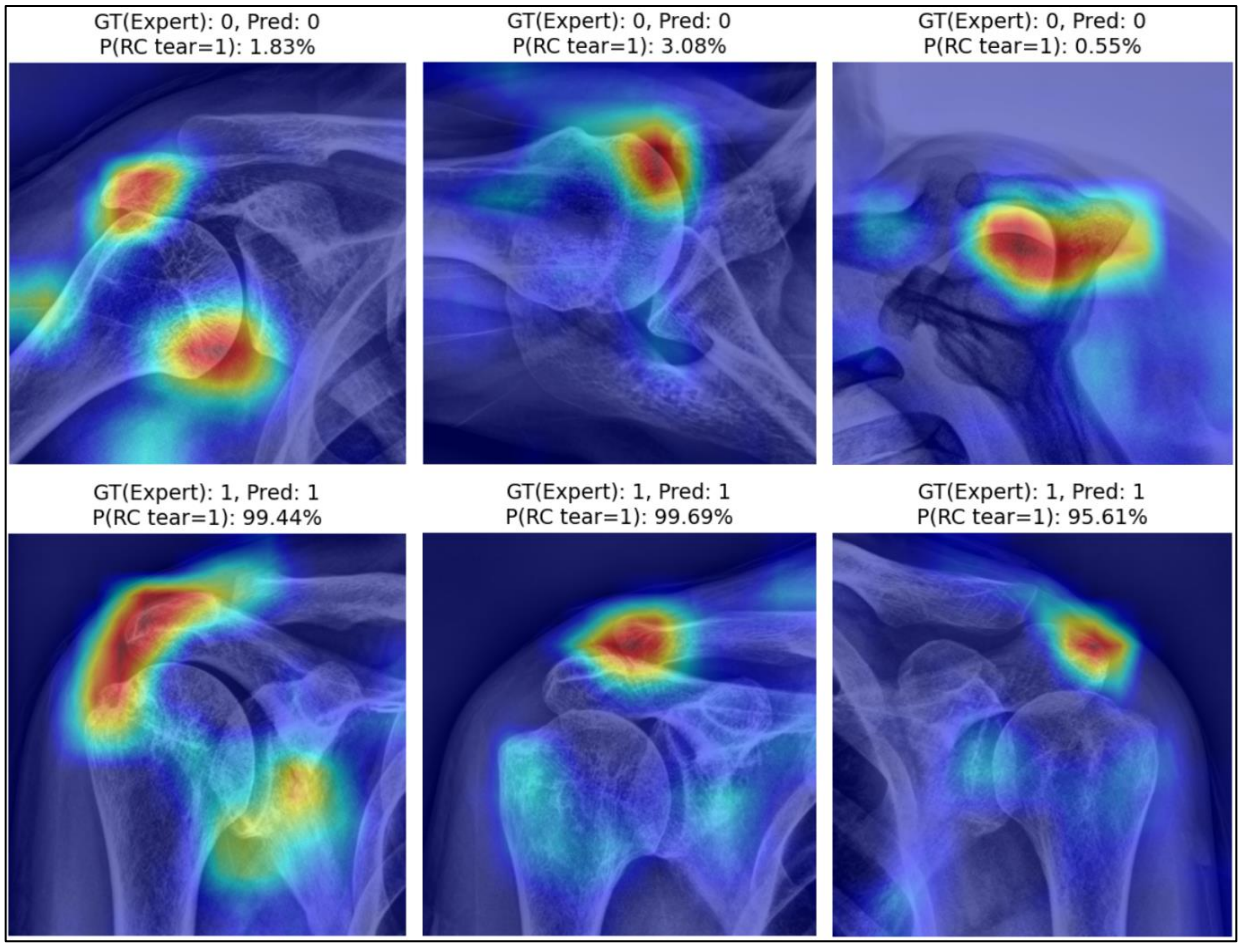


Figure 6: Visualization of Prediction Results Using Grad-CAM. We visualized which regions of the radiograph the model focused on during training using Grad-CAM. It was observed that the areas our model concentrated on through learning closely align with the regions that should be examined during the initial diagnosis stages of fRCT.

Table 1: Comparison of YOLO v5 Training Experiments.

The metrics used to evaluate the YOLO v5 model include Precision, Recall, mAP_0.5, and mAP_0.5:0.95. mAP_0.5 measures the mean Average Precision at an Intersection over Union (IoU) threshold of 0.5, while mAP_0.5:0.95 averages the mean Average Precision across multiple IoU thresholds ranging from 0.5 to 0.95. In our experiments, exp1 resized images to 640x640 pixels, and exp2 resized images to 800x800 pixels. As shown in Table 1, exp2 slightly outperformed exp1 in mAP_0.5:0.95, achieving 0.8003 compared to exp1's 0.7984. Therefore, we used the best weights from exp2 to perform the ROI crop on the images.

Experiment version	Precision	Recall	mAP_0.5	mAP_0.5:0.95
exp1	0.9975	1.000	0.9950	0.7984
exp2	**0.9986**	**1.000**	**0.9950**	**0.8003**

Fully Automated Volumetric Assessment of Abdominal Aortic Aneurysms

David Weiss[1], Thomas Hager, M.Sc.[2], Mariam Aboian, MD, PhD[1], MingDe Lin, PhD[3], Khaled Bousabarah, PhD[4], Daniel Renninghoff, M.Sc.[4], Wolfgang Holler, M.Sc.[4], Kathryn Simmons, B.S.[5], Sarah Loh, MD[5], Uwe Fischer, MD, PhD[5], Cornelius Deuschl, MD[6], Sanjay Aneja, MD[2], Edouard Aboian, MD[5]

[1]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA; [2]Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT, USA; [3]Visage Imaging, Inc., San Diego, CA, USA; [4]Visage Imaging, GmbH, Berlin, Germany; [5]Department of Vascular Surgery and Endovascular Therapy, Yale School of Medicine, New Haven, CT, USA; [6]Department of Diagnostic and Interventional Radiology and Neuroradiology, Essen University Hospital, Essen, Germany

Introduction

The estimated incidence of abdominal aortic aneurysms (AAA) is 0.4 to 0.67% in Western nations. Endovascular aortic repair (EVAR) emerged as a primary mode of aneurysm repair and gained widespread adoption [1]. Although the maximum aortic diameter (DMAX) presents the current reference standard for pre- and post-interventional assessment of patients with AAA, volumetric measurements may be more sensitive in detecting true aortic changes and identifying patients at risk [2, 3]. Importantly, the inter-rater variability of volumetric AAA measurements has been demonstrated to be less pronounced than that observed for DMAX [4]. However, semi-automatic volumetric measurement of AAAs presents a time-consuming method. Therefore, it is not widely established in clinical practice, and reliable automated approaches to volumetric AAA measurement are critically needed [5].

This study aims to train and validate a deep learning-based network that enables automated segmentation and volume determination of pre- and post-EVAR AAAs on computed tomography angiographies (CTA) both internally and externally. The clinical utility and efficiency gains facilitated by the tool were evaluated in comparison with vascular surgeons.

Material and Methods

In this HIPAA-compliant retrospective study, de-identified pre- and post-interventional CTAs of patients who underwent EVAR for management of infrarenal AAA between August 2017 and March 2023 at our institution were investigated. The training dataset comprised 80% of these CTAs, with 20% assigned to internal validation dataset. Pre- and post-EVAR CTAs from two external sites were used for external validation purposes [6]. Ground truth volumetric measurements were manually performed on all CTAs from lowest renal artery to aortic bifurcation and evaluated by two observers for consistency. Within preprocessing, all images were tailored to CTA segments containing measurements. A self-configuring nnU-Net model trained on our training dataset was validated on both internal and external validation datasets. In a single-fold cross-validation setup, the model was trained for 1,000 epochs for segmentation of total aneurysm; based on the observed performance, we elected 500 epochs for segmentation of lumen (Fig 1) [7]. AI-generated TA and LU volumes were correlated with determined ground truth values. Within a separate experiment, semi-automatic volumetric AAA segmentations were performed on internal validation dataset by two attending vascular surgeons. The times to perform a complete segmentation from lowest renal artery to aortic bifurcation were compared with generation time of the model. Baseline patient demographics were recorded.

Results

A total of 110 patients with 84 (76.4%) males and 26 (23.6%) females were included in our internal datasets; the mean age was 74.1 years. A total of 176 pre- and post-EVAR CTAs including 35,915 slices (204.06 per CTA) were utilized to train our network; the mean slice thickness was 0.774 millimeters. For TA and LU of the internal validation (44 CTAs), mean Dice similarity coefficients were 0.972±0.013 and 0.963±0.017, respectively. For the external validation (40 CTAs), mean Dice similarity coefficients were 0.959±0.042 and 0.955±0.029, respectively (Table 1, Fig 2). The generated TA and LU volumes showed a very strong correlation with determined ground truth values on both internal (r=0.998, P<0.001; r=0.993,

P<0.001) and external validation datasets (*r*=0.952, *P*<0.001; *r*=0.987, *P*<0.001) (Fig 3). Mean time savings of 59.2 (286.8vs116.9s) and 58.3% (276.3vs115.2s) provided by the model were demonstrated for TA and LU (Fig 4).

Discussion and Conclusion

Our network enables automated high-precision volumetric analysis of pre- and post-EVAR infrarenal AAAs. While related studies lack external validation, we presented an institution-agnostic algorithm facilitating substantial measurement workflow acceleration [8-10]. It can be incorporated into routine clinical practice to aid a more precise, efficient, and standardized morpho-volumetric evaluation of pre- and post-EVAR AAAs. The model is in the process of embedding into a research instance of our PACS (Visage Imaging, Inc., CA) allowing the application to non-preprocessed CTAs (Fig 5).

Keywords

Vascular surgery, aneurysm, volumetrics, deep learning, computed tomography angiography

Key Information (100 Words)

Research Question

Can a deep learning-based model improve the efficiency of morpho-volumetric analysis of AAAs by segmenting and measuring the total aneurysm and lumen volume of pre- and post-interventional infrarenal abdominal aortic aneurysms on internal and external CTAs?

Findings

The presented treatment-agnostic network facilitates high-precision segmentation and volume determination of both pre- and post-interventional infrarenal AAAs. The results were

reproducible on external CTAs. The model introduced significant time efficiency gains in comparison to experienced vascular surgeons.

Meaning

Our trained model enables accurate, reproducible, and efficient volumetric analysis of AAAs and is ready for clinical integration into PACS for sensitive AAA behavior monitoring.

References

1. Marcaccio, C.L. and M.L. Schermerhorn, *Epidemiology of abdominal aortic aneurysms.* Semin Vasc Surg, 2021. **34**(1): p. 29-37.
2. Chaikof, E.L., et al., *The Society for Vascular Surgery practice guidelines on the care of patients with an abdominal aortic aneurysm.* J Vasc Surg, 2018. **67**(1): p. 2-77 e2.
3. Vaitenas, G., et al., *Abdominal Aortic Aneurysm Diameter versus Volume: A Systematic Review.* Biomedicines, 2023. **11**(3).
4. van Prehn, J., et al., *Intra- and interobserver variability of aortic aneurysm volume measurement with fast CTA postprocessing software.* J Endovasc Ther, 2008. **15**(5): p. 504-10.
5. Schnitzbauer, M., et al., *CT after Endovascular Repair of Abdominal Aortic Aneurysms: Diagnostic Accuracy of Diameter Measurements for the Detection of Aneurysm Sac Enlargement.* J Vasc Interv Radiol, 2018. **29**(2): p. 178-187 e3.
6. Siriapisith, T., W. Kusakunniran, and P. Haddawy, *A retrospective study of 3D deep learning approach incorporating coordinate information to improve the segmentation of pre- and post-operative abdominal aortic aneurysm.* PeerJ Comput Sci, 2022. **8**: p. e1033.
7. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nat Methods, 2021. **18**(2): p. 203-211.
8. Caradu, C., et al., *Fully automatic volume segmentation of infrarenal abdominal aortic aneurysm computed tomography images with deep learning approaches versus physician controlled manual segmentation.* J Vasc Surg, 2021. **74**(1): p. 246-256 e6.
9. Lopez-Linares, K., et al., *Fully automatic detection and segmentation of abdominal aortic thrombus in post-operative CTA images using Deep Convolutional Neural Networks.* Med Image Anal, 2018. **46**: p. 202-214.

10. Caradu, C., et al., *Fully automatic volume segmentation using deep learning approaches to assess aneurysmal sac evolution after infrarenal endovascular aortic repair.* J Vasc Surg, 2022. **76**(3): p. 620-630 e3.

Tables and Figures

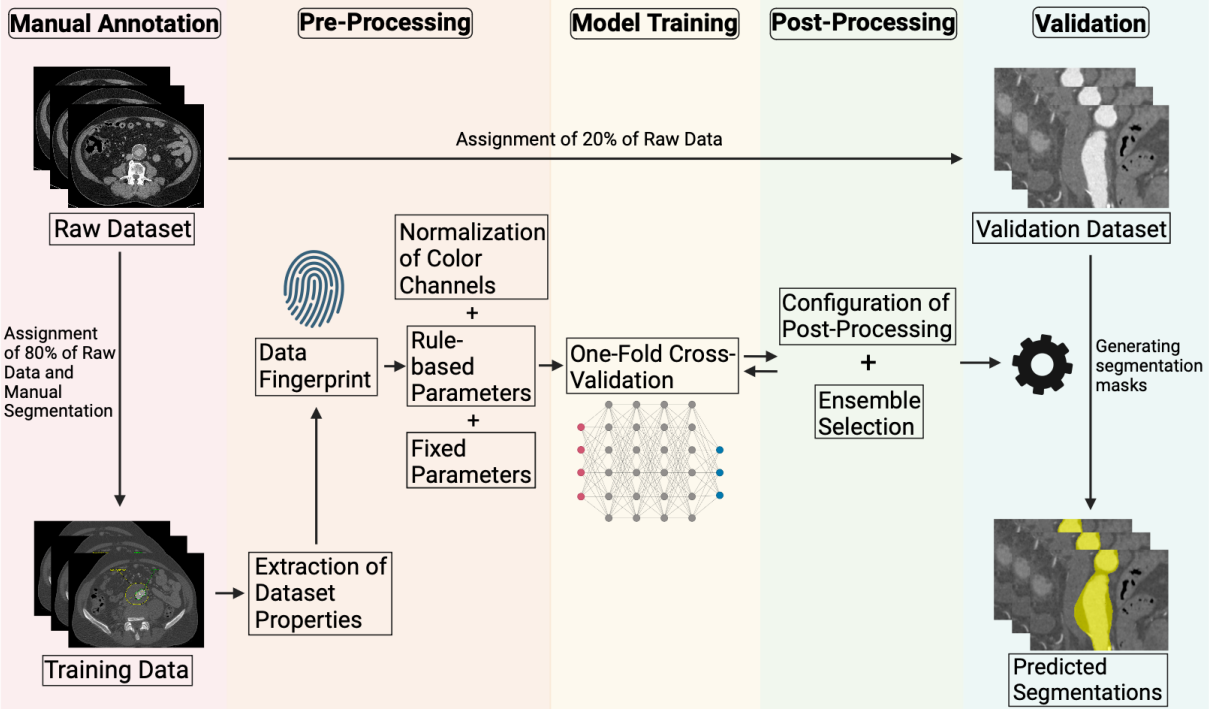


Figure 1. Modified nnU-Net workflow

The procedures of data annotation, network training, and validation are displayed. In addition to the existing model architecture provided by nnU-Net, the normalization of the CTA color channels was added to optimize the performance of our image analysis tool.

Total aneurysm

Dataset	Mean ± SD	Range	Median	IQR
Internal Validation (44 CTAs)				
Dice Similarity Coefficient	0.972 ± 0.013	0.917 - 0.985	0.976	0.971 - 0.989
Hausdorff Distance/Slice (in mm)	2.053 ± 2.379	1.0 - 15.339	1.414	1.207 - 1.747
External Validation (40 CTAs)				
Dice Similarity Coefficient	0.959 ± 0.042	0.747 - 0.983	0.968	0.962 - 0.978
Hausdorff Distance/Slice (in mm)	3.226 ± 9.166	1.0 - 52.292	1.0	1.0 - 1.207

Lumen

Dataset	Mean ± SD	Range	Median	IQR
Internal Validation (44 CTAs)				
Dice Similarity Coefficient	0.963 ± 0.017	0.905 - 0.986	0.969	0.957 - 0.974
Hausdorff Distance/Slice (in mm)	1.772 ± 1.614	1.0 - 9.225	1.207	1.0 - 1.640
External Validation (40 CTAs)				
Dice Similarity Coefficient	0.955 ± 0.029	0.809 - 0.979	0.960	0.957 - 0.968
Hausdorff Distance/Slice (in mm)	1.995 ± 3.785	1.0 - 21.108	1.0	1.0 - 1.104

Table 1. nnU-Net performance metrics – total aneurysm and lumen

The accuracy of the spatial segmentation masks generated by the network is quantitatively described. For the total aneurysm, median Dice Similarity Coefficients of 0.976 and 0.968 were observed in our internal and external validation datasets. For the lumen, mean Dice Similarity Coefficients of 0.963 and 0.955 were revealed in the internal and external validation datasets.

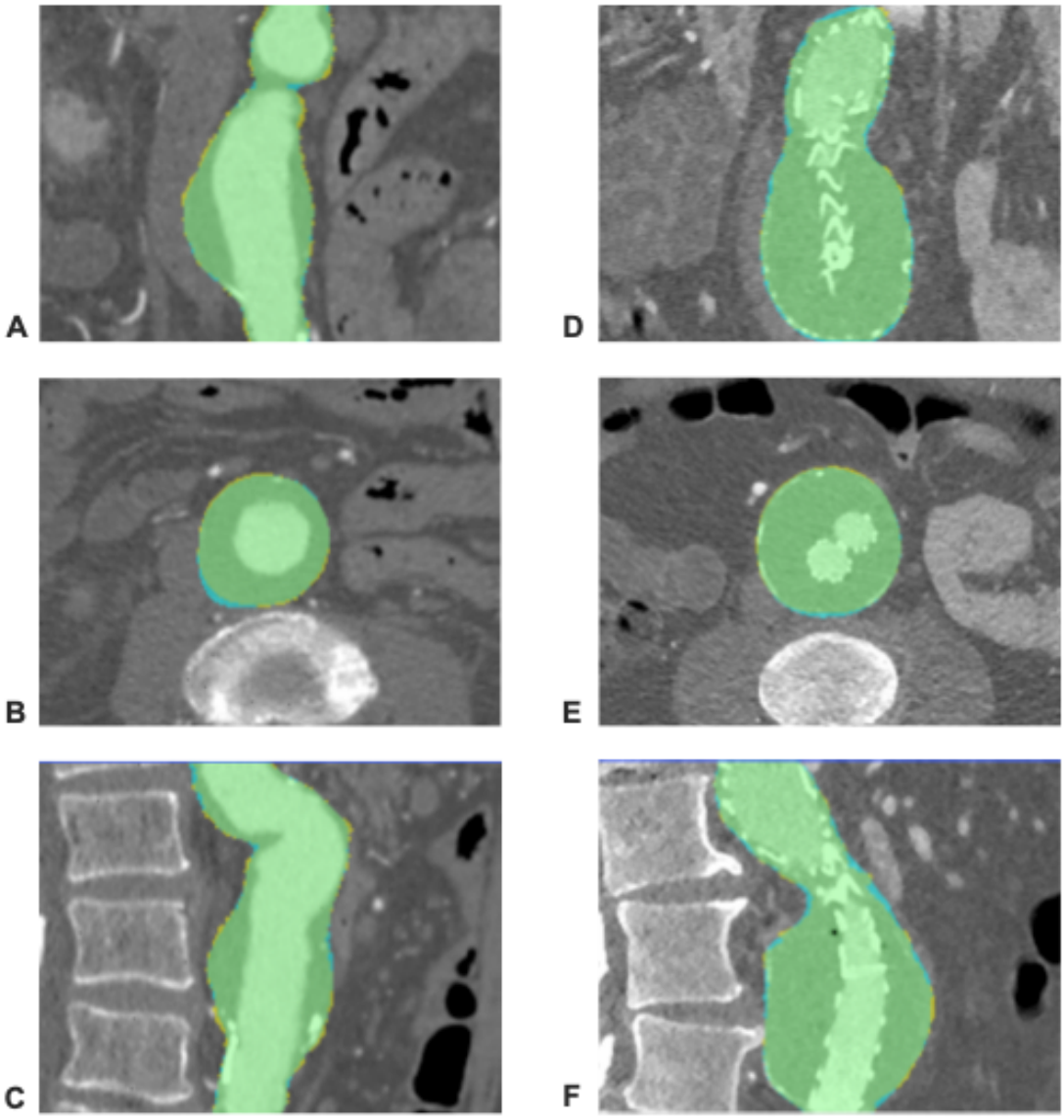


Figure 2. Representative CTA studies displaying AI-generated and ground truth segmentation masks of AAAs

A - C. The images of a pre-EVAR AAA of the external validation dataset including the AI-generated segmentation (blue), the ground truth segmentation (yellow), and their overlap (green) are depicted on coronal (**A**), axial (**B**), and sagittal (**C**) planes (DSC 0.968). The AI-generated volume was 95.48 and the ground truth volume 96.66 cm^3. **D - F.** The CTA study of a post-EVAR AAA of the internal validation dataset including the related segmentations and their overlap are shown on coronal (**D**), axial (**E**), and sagittal (**F**) planes (DSC 0.976). The AI-generated volume was 191.7 and the ground truth volume 187.3 cm^3.

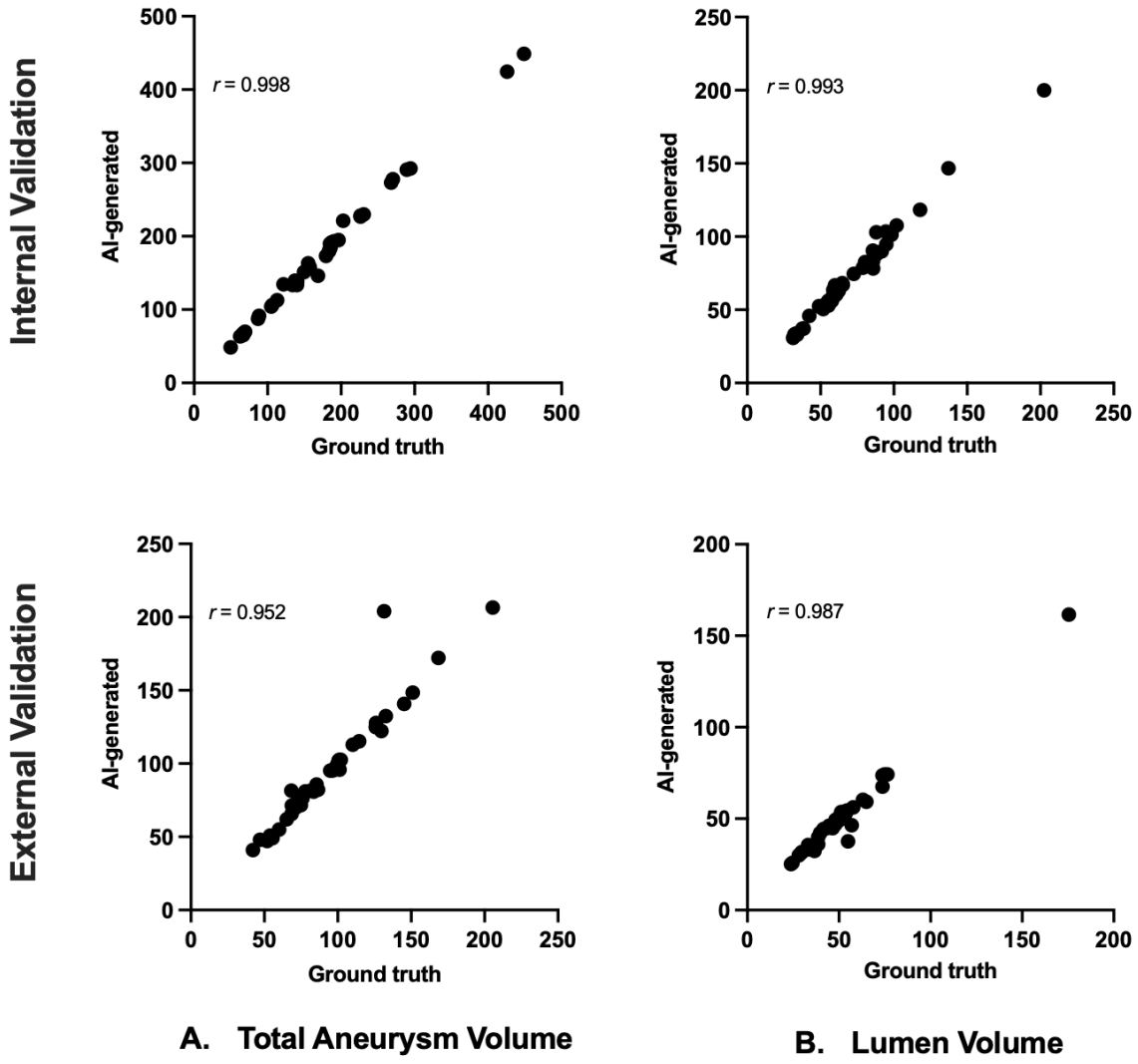


Figure 3. Correlation of AI-predicted and ground truth volumes

The volumes (in cm^3) of the total aneurysm (**A**) and lumen (**B**) of the internal and external validation datasets calculated by the trained network were correlated with the ground truth volumes. The predicted aneurysm sac volumes displayed on both internal ($r = 0.998$, $P < 0.001$) and external CTAs ($r = 0.952$, $P < 0.001$), the lumen volumes presented in the internal ($r = 0.993$, $P < 0.001$) and external validation datasets ($r = 0.987$, $P < 0.001$) demonstrated a very strong correlation with the determined ground truth volumes.

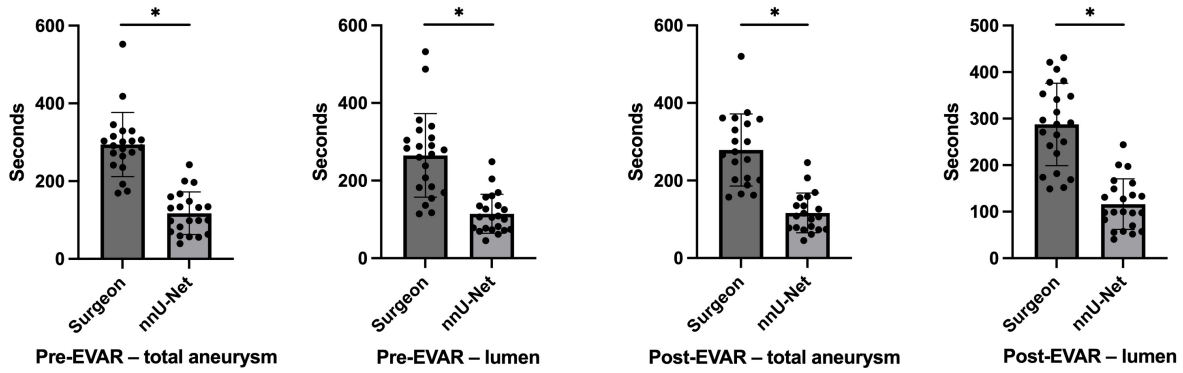


Figure 4. Time efficiency gains – nnU-Net vs vascular surgeons measuring AAA volumes

The time efficiency gains provided by the model in comparison with the manual measurements of two attending vascular surgeons are depicted. For total aneurysm segmentations, the model facilitated a mean time saving of 59.2 percent, whereas a mean time saving of 58.3 was observed for the lumen segmentations.

Artificial Intelligence for Prediction of Abdominal Aortic Aneurysm Status Using Multimodal Patient Data in the VASCULAID-RETRO-AAA Study

Authors:
L. Rijken[1,2,3], V. Ayyasomalayajula[1,2], J.M. Wolterink[4], K.K. Yeung[1,2]

Affiliations:

1. *Amsterdam University Medical Centers, Department of Vascular Surgery, Amsterdam, the Netherlands*
2. *Amsterdam Cardiovascular Sciences, Microcirculation, Amsterdam, the Netherlands*
3. *Amsterdam Public Health, Digital Health, Amsterdam, the Netherlands*
4. *Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, the Netherlands*

Presenting author:
Lotte Rijken[1,2,3]

1. *Amsterdam University Medical Centers, Department of Vascular Surgery, Amsterdam, the Netherlands*
2. *Amsterdam Cardiovascular Sciences, Microcirculation, Amsterdam, the Netherlands*
3. *Amsterdam Public Health, Digital Health, Amsterdam, the Netherlands*

Email: l.rijken@amsterdamumc.nl

Keywords:
Abdominal aortic aneurysm, disease progression, multimodal prediction models.

Key information:

1. Research question: An abdominal aortic aneurysm (AAA) is a vascular pathology with significant morbidity and mortality. We are investigating the role of multimodal AI models that take into account clinical, imaging, and multi-omics patient data for prediction of AAA disease progression.
2. Findings: Data infrastructure has been set up and clinical data collection is ongoing. Algorithms for automated segmentation of AAA geometry are developed. Proteomic and genomic analysis of patient biological samples is ongoing.
3. Meaning: Patient-specific risk stratification tools can help clinicians to forecast the risks of rupture and decide the optimal time and treatment to be performed.

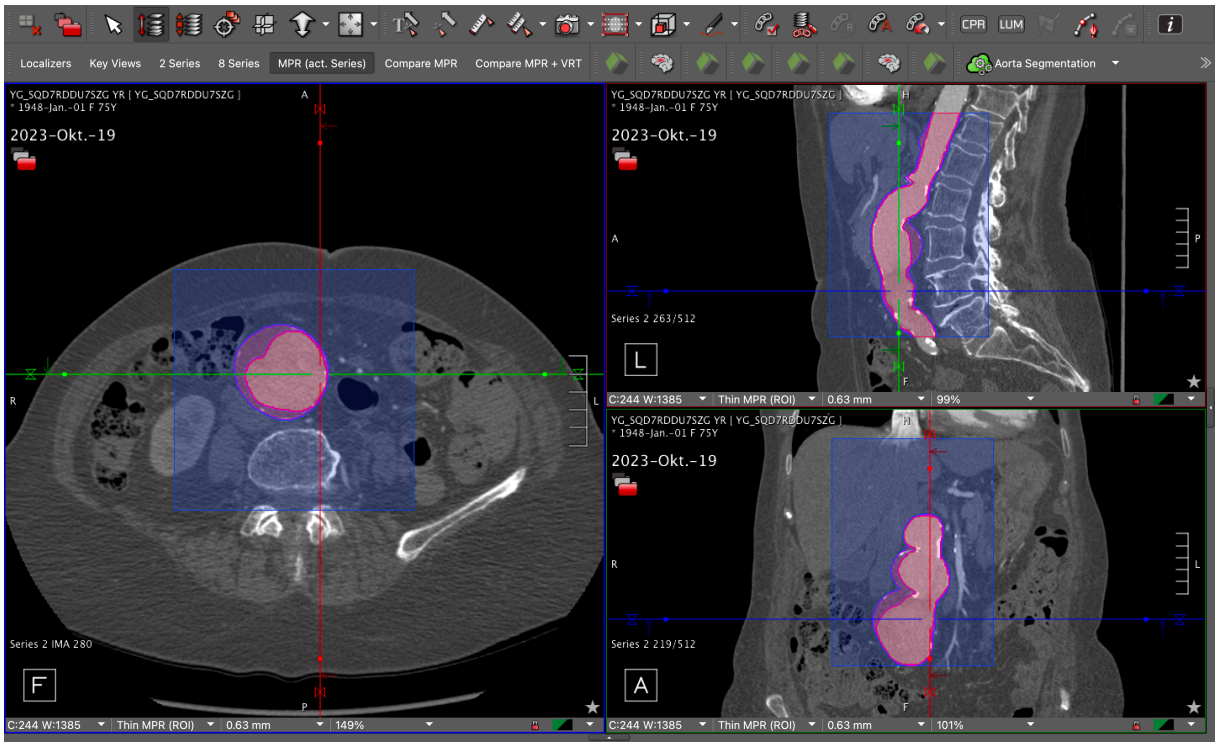


Figure 5. PACS-integrated AAA algorithm – MPR series

The layout of the volumetric AAA analysis tool embedded in our research PACS is displayed. Requiring the push of only one button ("Aorta Segmentation"), high-precision segmentations of the total aneurysm and lumen are automatically generated.

MANUSCRIPT (word count: 599)

Introduction

An abdominal aortic aneurysm (AAA) is a pathological dilatation of the aorta and is associated with significant mortality when ruptured. In current clinical practice, AAA diameter remains the most widely used criterion for intervention to prevent rupture[1]. In the era of personalized medicine, there is a need to go beyond this one-size-fits-all approach[2]. Furthermore, AAA is associated with significant heritability and complex mutational heterogeneity[3]. Moreover, approximately 30% of AAA patients are at high cardiovascular risk, necessitating optimal cardiovascular risk management[4]. In the VASCULAID-RETRO-AAA study, we aim to integrate multimodal information including clinical, imaging, and multi-omics data in robust machine learning pipelines to go beyond this one-size-fits-all approach and classify patients at high risk for AAA progression and cardiovascular disease.

Material and methods

The VASCULAID-RETRO-AAA study aims to leverage retrospectively collected data of 5,000 AAA patients from six European clinical centers to develop multimodal AI-algorithms. A data infrastructure network is established to collect standardized clinical and imaging data from all centers. This phase included developing a consensus-based variable set to select relevant variables from a clinical and ethical perspective. Moreover, blood and plasma samples from 250 patients from existing biobanks will be used for -omics analyses. Additionally, external databases and biobanks such as UK biobank, Finngen, Hunt, Kaiser Permanente are explored to enrich the VASCUL-AID database.

As part of this study, geometric deep learning models are trained on previously developed automated AAA segmentations to estimate hemodynamic parameters such as pressure and wall shear stress. Mass spectroscopy-based proteomic and lipidomic analysis of plasma samples will be undertaken to identify differentially secreted proteins and classify them into high vs low AAA progression. Targeted genotyping will be done to identify the SNP information at previously identified risk loci. Finally, clinical data, data from medical imaging analyses, and multi-omics data will be integrated to be used as input to multimodal prediction models based on graph neural networks (GNNs). The structure of the multimodal GNN is designed using a pilot-dataset.

Results

A robust data infrastructure has been established that incorporates Castor EDC for clinical data collection and a Health-RI XNAT-server for imaging data storage and retrieval. These platforms are operational and collection of clinical and imaging (CT, US, MRI) data of all 5,000 identified AAA patients has started. Through collaborative efforts, a comprehensive consensus-based variable list, of around 300 potential clinical input variables for the prediction models, has been formulated. Moreover, the R10-version of Finngen biobank data[5] with 4083 AAA patients is accessed for genomic analyses.

An automatic segmentation of the whole arterial tree to model complex vascular structures has been developed within our consortium[6]. Proteomic and genomic analysis of blood samples is currently ongoing. A knowledge graph of known gene associations with AAA including 590 genes and 141 significant-loci associated with AAA progression has been constructed based on literature review and open-source databases[7].

A pilot-dataset was used to create the preliminary design of the multimodal GNN model. Each patient will be represented as a node in a cohort level graph representation and each node embedding is derived from modality specific models. For instance, to generate aggregate feature inputs of each modality, the data is converted to a d-dimensional vector representation and fused together with other modalities. This model will be optimized to generate overall risk scores or probabilities for AAA progression and the risk of cardiovascular events.

Discussion and Conclusion

Extensive AI analysis of imaging and clinical patient data will be integrated with omics data for the first time to predict progression of AAA. Deep graph neural network models based on multimodal data will become a critical tool in advancing precision medicine.

References

1. Chaikof EL, Dalman RL, Eskandari MK, et al. The Society for Vascular Surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. *J Vasc Surg*. 2018;67(1):2-77.e2. doi:10.1016/j.jvs.2017.10.044
2. Darling RC, Messina CR, Brewster DC, Ottinger LW. Autopsy study of unoperated abdominal aortic aneurysms. The case for early resection. *Circulation*. 1977;56(3 Suppl):II161-4.
3. Li J, Pan C, Zhang S, Spin JM, Deng A, Leung LL, Dalman RL, Tsao PS, Snyder M. Decoding the genomics of abdominal aortic aneurysm. Cell. 2018 Sep 6;174(6):1361-72.
4. Tomee SM, Bulder RMA, Meijer CA, van Berkum I, Hinnen J-W, Schoones JW, et al. Excess Mortality for Abdominal Aortic Aneurysms and the Potential of Strict Implementation of Cardiovascular Risk Management: A Multifaceted Study Integrating Meta-Analysis, National Registry, and PHAST and TEDY Trial Data. Eur J Vasc Endovasc Surg Off J Eur Soc Vasc Surg. 2023 Mar;65(3):348–57.
5. Kurki, M.I., Karjalainen, J., Palta, P. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 613, 508–518 (2023). https://doi.org/10.1038/s41586-022-05473-8
6. Alblas D, Brune C, Yeung KK, Wolterink JM. Going off-grid: continuous implicit neural representations for 3D vascular modeling. InInternational Workshop on Statistical Atlases and Computational Models of the Heart 2022 Sep 18 (pp. 79-90). Cham: Springer Nature Switzerland.
7. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic acids research. 2020 Jan 8;48(D1):D845-55.

Disclosures

None

Machine learning-based prediction of spinal cord ischemia after complex endovascular aneurysm repair using clinical and imaging-derived features

Authors:

Kaj O. Kappe, MSc.[1,2], J. Engel, BSc.[1,2], Vincent Jongkind, MD, PhD[1,2], Jelmer M. Wolterink, PhD[3], Kak Khee Yeung, MD, PhD[1,2]

Affiliations:

[1] *Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Surgery, Amsterdam, the Netherlands*

[2] *Amsterdam Cardiovascular Sciences, Microcirculation, Amsterdam, the Netherlands*

[3] *Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, the Netherlands*

Presenting author:

Kaj O. Kappe

k.o.kappe@amsterdamumc.nl

[1] *Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Surgery, Amsterdam, the Netherlands*

[2] *Amsterdam Cardiovascular Sciences, Microcirculation, Amsterdam, the Netherlands*

Keywords:

Spinal cord ischemia, sarcopenia, image segmentation, classification, outcome prediction

Key information:

1. Research question: Spinal cord ischemia (SCI) after fenestrated and branched endovascular aneurysm repair (F/B-EVAR) remains a rare but devastating complication. The aim of this study was to evaluate the performance of classifiers for the prediction of SCI based on clinical and preoperative imaging-derived parameters.
2. Findings: The trained classifiers demonstrate potential for prediction of SCI after F/B-EVAR based on clinical and preoperative imaging-derived parameters.
3. Meaning: Preoperative prediction of SCI after F/B-EVAR could aid in optimization of the treatment strategy for patients with a thoracoabdominal aortic aneurysm to reduce the risk of SCI.

MANUSCRIPT *(included in the word count)*

Introduction

The introduction of fenestrated and branched endovascular aneurysm repair (F/B-EVAR) has increased the number of aneurysm varieties that can be treated in a minimally invasive fashion. Despite improved post-operative outcomes over the years, spinal cord ischemia (SCI), potentially leading to paraplegia, remains a devastating complication. SCI is characterized by a multifactorial etiology, making it difficult for clinicians to stratify high and low-risk patients and promptly apply preventive measures. Sarcopenia is characterized by skeletal muscle loss, and its severity is commonly assessed based on two-dimensional psoas muscle measurements on the axial computed tomography (CT) plane. Sarcopenia has, amongst others, been suggested as a predisposing factor for SCI.[1]

This study aimed to classify SCI and non-SCI patients using a machine learning model based on clinical features and sarcopenia-related anatomical features derived automatically from CT. Preoperative prediction of SCI could potentially alter the treatment strategy and benefit the patient by lowering the risk of SCI.

Material and methods

Within our international multicenter consortium, patient characteristics, preoperative CT, and post-operative SCI status were collected for patients who underwent F/B-EVAR for a thoracoabdominal aortic aneurysm. In all centers, patients with SCI were selected and matched with non-SCI patients.

Iliopsoas muscle segmentations were automatically derived from the preoperative CT as a surrogate for sarcopenia using the open-source TotalSegmentator tool.[2,3] Two-dimensional (2D) conventional measurements were obtained from the segmentations at the level of vertebra L3 based on the vertebra labels. The collected 2D parameters were the psoas muscle area at vertebra level L3 [cm^2], attenuation of the area (Hounsfield Units, [HU]), and lean psoas muscle area (LPMA), measured by multiplication of the psoas muscle area and attenuation [cm^2*HU]. Three-dimensional collected parameters were the iliopsoas muscle volume [cm^3], iliopsoas muscle volume attenuation [HU], and lean iliopsoas muscle volume (LIMV) in [cm^3*HU].

The collected clinical parameters included the patient's age, sex, BMI, ASA-classification, glomerular filtration rate, aneurysm classification, and maximum aneurysm diameter.

Multiple machine learning models were trained to predict SCI after F/B-EVAR based on the sarcopenia-related and clinical parameters. The trained models were Gaussian Naïves Bayes Classifier, Random Forest, Logistic Regression, and XG Boost. The performances of the trained models were based on the mean area under the curve (AUC) of the receiver operating characteristic (ROC) curve in a 5-fold cross-validation.

Results

Two hundred and five patients were included (mean age of 72.5 ± 7.7 years, 71.7% male) with 148 (72.1%) patients treated for a Crawford I-IV thoracoabdominal aortic aneurysm and 57 (27.8%) for a juxta/pararenal abdominal aortic aneurysm. Sixty patients (29.2%) had SCI postoperatively.

The best-performing machine learning model was the Gaussian Naïves Bayes classifier with a mean AUC of 0.82 ± 0.10 in the 5-fold cross-validation. The Random Forest reached an AUC of 0.70 ± 0.09, logistic regression reached an AUC of 0.79 ± 0.08 and the XG Boost classifier obtained an AUC of 0.73 ± 0.10.

Discussion and Conclusion

Preoperative SCI risk stratification remains a challenge for patients treated with F/B-EVAR for a thoracoabdominal aortic aneurysm. The trained machine learning models demonstrated promising results for the prediction of SCI based on sarcopenia-related and clinical parameters. These results suggest that machine learning models could potentially aid the physicians in better SCI risk stratification, and thereby enhance optimization of the treatment strategy. The current study was limited by the number of patients. Further research should focus on the inclusion of more patients and the addition of more imaging-derived and clinical parameters to enhance the performance of the machine learning models.

References

1. Kölbel T, Nana P, Torrealba JI, Panuccio G, Behrendt CA, Spanos K. The Association of Sarcopenia and ASA Score to Spinal Cord Ischemia in Patients Treated With the t-Branch Device. *J Endovasc Ther*. Jun 6 2023:15266028231179414. doi:10.1177/15266028231179414
2. Wasserthal J, Breit H-C, Meyer MT, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*. 2023;5(5):e230024. doi:10.1148/ryai.230024
3. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*. 2021/02/01 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z

Disclosures

None.

H&E-based cell prediction multi-classification models to capture morphologically distinct subpopulations of $CD8^+$ T cells

Authors:

Muzammil Arif Din Abdul Jabbar[1,2], Willa Wen-You Yim[2]*, Felicia Wee[2], Li Yen Chong[2], Joe Yeong[2,3]*

*Co-corresponding

Affiliations:

[1]University of Cambridge, Cambridge CB2 1TN, United Kingdom

[2]Institute of Molecular and Cell Biology (IMCB), Agency for Science, Technology and Research (A*STAR), Singapore 138673, Singapore

[3]Immunology & Serology Section, Department of Microbiology, Division of Pathology, Singapore General Hospital, Singapore 169856, Singapore

Presenting author:

Muzammil Arif Din Abdul Jabbar

University of Cambridge, Cambridge CB2 1TN, United Kingdom & Agency for Science, Technology and Research (A*STAR), Singapore 138673, Singapore

mads2@cam.ac.uk

+44 0 7724 280 477

Keywords:

Cancer, histopathology, digital pathology, machine learning, $CD8^+$ T cells

Key information:

1. Research question: Do morphological subpopulations exist within H&E images of $CD8^+$ T cells?
2. Findings: Feature extraction using the UNI model followed by hierarchical clustering demonstrated existence of clusters in $CD8^+$ T cells.
3. Meaning: These clusters have important implications for development of machine learning models for identification of $CD8^+$ T cells subsets in tumour H&E slides, allowing for characterisation of the immune profile of tumours without additional immunostaining.

MANUSCRIPT

Introduction

H&E2.0[1] refers to the identification of biomarker-expressing cells directly from hematoxylin and eosin (H&E) images through machine learning (ML) methods, without the need for immunohistological staining[2]. Such models are generally binary classification models[3,4]. However, morphologically distinct subpopulations may exist within the biomarker-positive class that would affect the performance of binary classification models and multiclass models may be more appropriate, Here, we demonstrate that morphologically sub-populations can be detected in H&E images of $CD8^+$ cells and that model prediction performance increases when a multi-classification model is used to identify all $CD8^+$ subpopulations and $CD8^+$ cells.

Material and methods

We employed H&E images obtained from four colorectal cancer tumors, which had been previously immunostained for $CD8^+$. We utilized the UNI model, a neural network model pre-trained on 100 million pathology images[5], and the Prov-GigaPath model, pretrained on 1.3 billion image tiles,[6] to extract features from 100 by 100 pixel patches of cells. Subsequently, we employed unsupervised hierarchical clustering to identify potential cell clusters within the $CD8^+$ population (Figure 1A). The features extracted were then used to train extreme gradient boosting (XGBoost) machine learning models to classify patches as $CD8^+$ or $CD8^-$ with 5-fold cross-validation (CV). Both a binary and multi-class model were trained and tested on the same training and testing datasets, based on a random 80:20 train:test split.

Results

Our analysis revealed the presence of distinct clusters within the $CD8^+$ population across the colorectal cancer tumors, irrespective of tumor origin (Figure 1B-D). Furthermore, we demonstrated the robustness of these findings, even when employing smaller image sizes of up to 60 by 60 pixels. Both the UNI and Prov-GigaPath identified features that suggested six distinct clusters within the $CD8^+$ images, with 52.3% concordance between both models, compared to 1 in 6 or 16.6% for chance, with most errors being at boundaries between both clusters (Figure 1C). The features extracted by the UNI vision transformer could be used to train an extreme gradient boosting (XGBoost) binary model with an AUROC score of 0.870±0.0097 (mean±SD of 5 CV iterations), whereas a multi-class model that predicted $CD8^+$ subpopulations and $CD8^-$ cells had an AUROC score of 0.875±0.0086 (Figure 1E), Similarly, the Prov-GigaPath model had an accuracy of 0.875±0.0037 and 0.887±0.0018 respectively.

Discussion and Conclusion

The discovery of morphologically distinct subpopulations within $CD8^+$ cells of H&E cell patches and the higher AUROC score achieved by the multi-classification model that captured all subtypes of the positive class indicates that multi-classification may be more appropriate than binary classification for H&E cell prediction models. Limitations of this work include the small number of tumors, four, that only one type of tumor, colorectal, was studied, and that only one biomarker was examined. We will address these limitations in future work along with testing other feature extraction methods.

References

1. Feng X, Zhang M, Yim W, et al. 1282 H&E 2.0 virtual staining of CD3+ T cells: improving cell labeling for precise model training. *Journal for ImmunoTherapy of Cancer*. 2023;11(Suppl 1). doi:https://doi.org/10.1136/jitc-2023-SITC2023.1282
2. Rood JE, Maartens A, Hupalowska A, Teichmann SA, Regev A. Impact of the Human Cell Atlas on medicine. *Nature Medicine*. 2022;28(12):2486-2496. doi:https://doi.org/10.1038/s41591-022-02104-7
3. Bai B, Yang X, Li Y, Zhang Y, Pillar N, Ozcan A. Deep learning-enabled virtual histological staining of biological samples. *Light, Science & Applications*. 2023;12:57. doi:https://doi.org/10.1038/s41377-023-01104-7
4. Lee K, Lockhart JH, Xie M, et al. Deep Learning of Histopathology Images at the Single Cell Level. *Frontiers in Artificial Intelligence*. 2021;4. doi:https://doi.org/10.3389/frai.2021.754641
5. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine. Published online March 19, 2024:1-13. doi:https://doi.org/10.1038/s41591-024-02857-3
6. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*. 2024;630(8015):181-188. doi:10.1038/s41586-024-07441-w

Disclosures

No conflict of interest declared.

APPENDIX

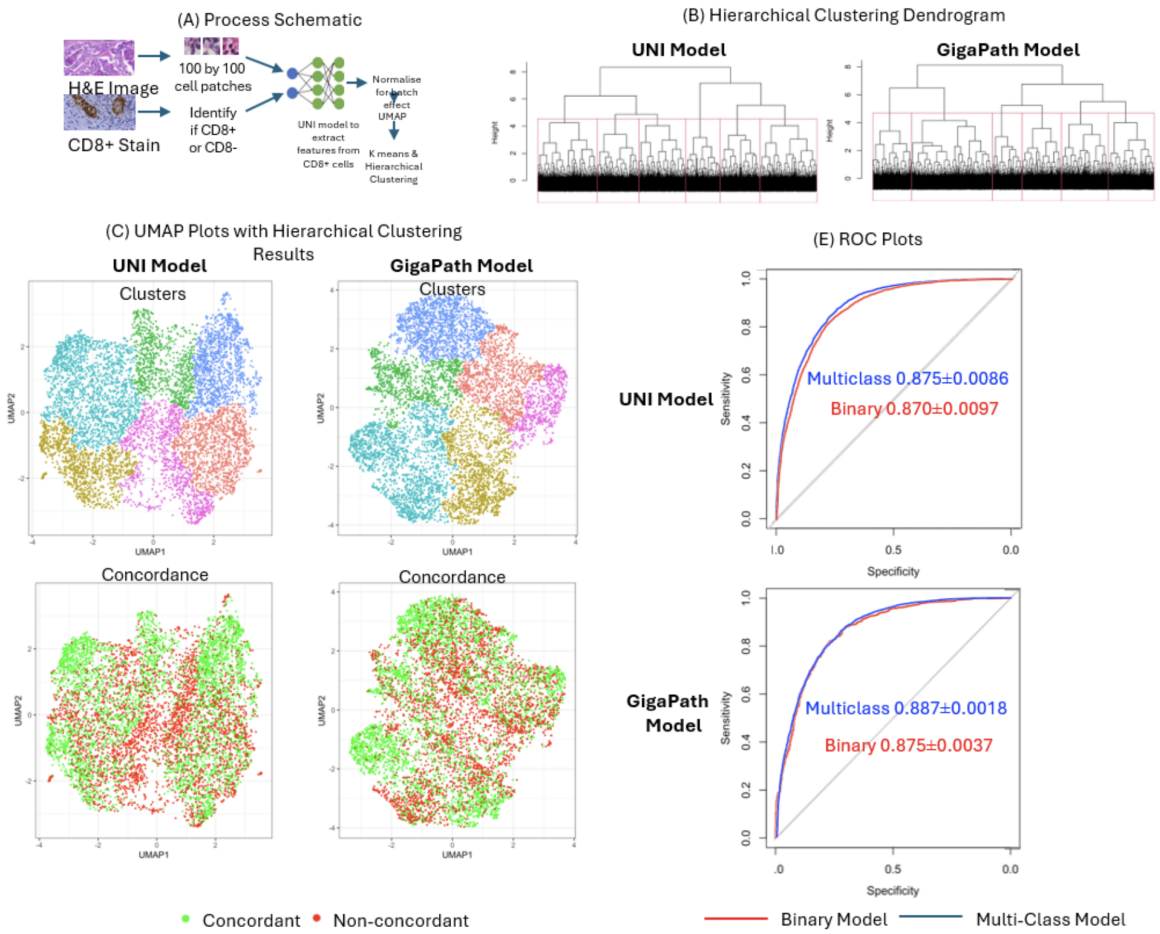


Figure 1: (A) Schematic of process of obtaining patches, feature extraction and clustering. (B) Hierarchical clustering dendrogram depicting six clusters obtained from both models. (C) UMAP Plots with hierarchical clustering results on top, and concordance of both clustering methods below. (D) ROC Plots of binary and multi-class classification models showing marginally increased accuracy for multiclass classifiers

From Concept to Clinic: Improving Colorectal Cancer Diagnosis through Deep Learning-Driven Lymph Node Screening in Clinical Routine

Authors:

Amjad Khan[1*], Bastian Dislich[1*], Stefan Reinhard[1], Aurel Perren[1], Inti Zlobec[1]

Affiliations:

[1]Institute of Tissue Medicine and Pathology, University of Bern, Switzerland
*These authors equally contributed to this work and share first authorship

Presenting author:

Amjad Khan, PhD
Institute of Tissue Medicine and Pathology, University of Bern, Switzerland
Murtenstrasse 31, 3008, Bern, Switzerland
E-Mail: amjad.khan@unibe.ch

Keywords:

AI-integration, colorectal cancer, lymph nodes, metastasis, histopathology

Key information:

1. Research question:

   The microscopic evaluation of lymph nodes for potential metastases can be a time-consuming and laborious task in the routine diagnosis of colorectal cancer. Can a deep learning model be used in a real-life clinical setting to assist pathologists in the diagnosis of lymph node metastases in cases of colorectal cancer?

2. Findings:

   To assist pathologists, we have recently developed and integrated a deep learning-based model for detecting lymph node metastases in colorectal cancer (CRC), called MetAssist. In this clinical quality control setup, MetAssist helped to improve diagnosis sensitivity from 98% to 100%.

3. Meaning:

   The pathologists have found the clinical integration of MetAssist to be supportive and provided valuable comments to improve the workflow. Under prospective clinical conditions, our analysis shows that MetAssist performed extremely well.

MANUSCRIPT

Introduction

The N-stage of a patient with colorectal cancer (CRC) is determined by the number of metastatic lymph nodes, including micro-metastases between 0.2 and 2.0 mm[1]. A minimum of 12 lymph nodes should be examined, although the actual number of resected and examined lymph nodes is often higher. The microscopic assessment of lymph node tissue for potential metastases can be a time-consuming task in routine diagnosis. To assist pathologists, we have recently developed and validated a deep learning-based model for the detection lymph node metastases in colorectal cancer (CRC), details for model development, architecture, training, and validation can be found here[2,3]. The model works in three steps: (1) lymph node tissues are first segmented on a low-resolution whole slide image (WSI), (2) segmented lymph node tissues are screened for metastases on higher resolution patches and (3) finally overlays of potential metastatic regions are generated and the patch level scores are aggregated to predict the slide score (e.g. positive or negative). For clinical quality control, the model has been integrated into the institute's diagnostic setup and has been named “MetAssist”. In this study, we aim to assess the prospective accuracy and feasibility of MetAssist when used by expert pathologists in the clinical practice for CRC cases in a tertiary care hospital.

Material and methods

In clinical pathology settings, the image management system is often used to deploy AI (artificial intelligence) algorithms. However, in this study, we have integrated an AI model for lymph node metastasis screening in the simplest form. Pathologists can immediately use the AI algorithms without any technical barriers. Our proposed integration pipeline uses a remote high-performance cluster (HPC) at the university with an automated workflow between the institute and the HPC, as well as a simplified web interface for pathologists’ interaction, result visualisation and comparison, review, and feedback (see Figure 1). After scanning lymph node slides with the institute's digital scanner, the automated workflow copies the WSI to the HPC and runs the model using a GPU node. Once the computational job is complete, the results are sent back to the institute in the form of overlays and WSI slide level scores for display on the web interface. Meanwhile, pathologists perform microscopic analysis on the glass slides to ensure that no case is left without human review and report their slide scores to the web interface. As the scores are fed in by the pathologists, the interface interactively displays similarities or differences and prompts a review request for discrepant cases by highlighting case in the interface. Pathologists can quickly review and leave comments if there is any false call, both from pathologist and the model. A total of 35 CRC resected specimens (427 slides with locoregional lymph nodes) have been diagnosed by 14 pathologists in this setting, who compared their diagnosis with the MetAssist predictions and reviewed cases with discrepancy.

Results

The technical implementation of the workflow proved to be stable. The web interface and user experience were also well-received and accepted by the pathologists. In terms of clinical diagnosis, the sensitivity and specificity of MetAssist compared to pathologists were approximately 0.981 and 0.769, respectively. Similarly, the accuracy (ACC) and the area under the curve of the receiver operating characteristic curve (AUCROC) were 79.6% and 0.875, respectively. The disparities in negative slides could be attributed to various tissue artefacts, mainly tissue folds and fragments of primary cancer that were detected, but

outside of lymph nodes. The pathologists commented on each false positive. The MetAssist indications prompted a review of the slides. As a result, the final AI-assisted pathology diagnostic sensitivity and specificity were improved to 1.0 and 0.770, respectively. Correspondingly, the ACC and AUCROC were 79.8% and 0.885 respectively.

Discussion and Conclusion

Overall, the pathologists have found the integration of MetAssist to be beneficial. They also provided valuable comments to improve both the model and the integration process. Under prospective clinical conditions, our analysis shows that MetAssist performed extremely well, reaching a sensitivity of 100%. Therefore, our workflow meets the desired diagnostics performance. At present, we are working on improving the model on false predictions, in particular to improve the segmentation step considering tissue folds and floating parts of primary tissue. In the future, we have plans to extend the model to other lymph node tissue types and to integrate other algorithms currently under development for more clinical applications.

References

1. Amin MB, Edge SB, Greene FL, et al. American Joint Committee on Cancer (AJCC). AJCC Cancer Staging Manual. *AJCC Cancer Staging Manual*. Published online 2017:211-212.
2. Khan A, Brouwer N, Blank A, et al. Computer-Assisted Diagnosis of Lymph Node Metastases in Colorectal Cancers Using Transfer Learning With an Ensemble Model. *Modern Pathology*. 2023;36(5):100118. doi:10.1016/J.MODPAT.2023.100118
3. Khan A, Janowczyk A, Müller F, et al. Impact of scanner variability on lymph node segmentation in computational pathology. *J Pathol Inform*. 2022;13:100127. doi:10.1016/j.jpi.2022.100127

Disclosures

The authors have no competing interest related to this study.

APPENDIX

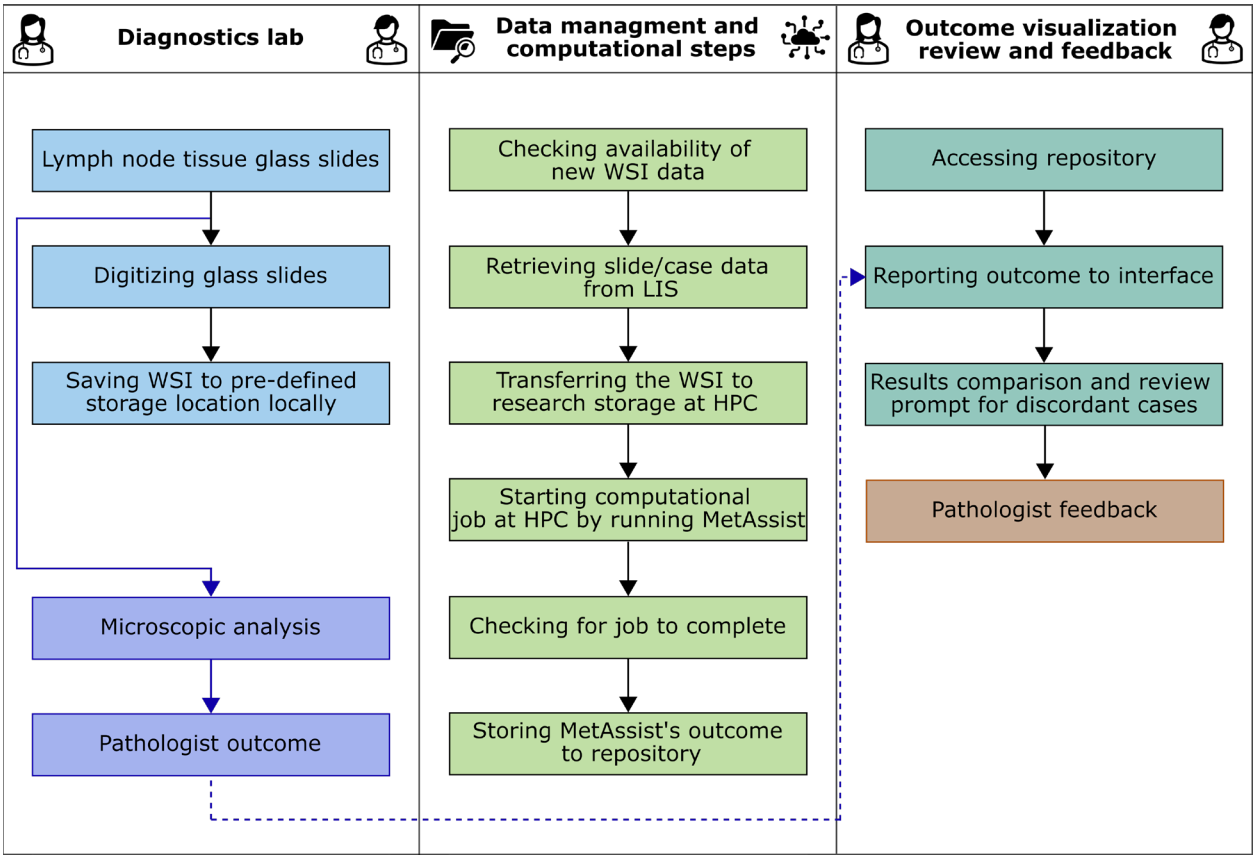


Figure 1: The block diagram shows all the steps and processes involved in the integration of MetAssist for Pathologists into the institute's lab. [LIS: laboratory information system, WSI: whole slide image, HPC: high-performance cluster]

Deep learning for intraoperative navigation in minimally invasive liver surgery: a multicenter external validation

Authors:

Namkee Oh, M.D.[1], Bogeun Kim, M.Sc.[2], Jongmin Shin, M.Sc[2], Sunmin Park, M.Sc[2], Manuel Lim, M.D.[3], Chan Woo Cho, M.D.[4], Ph.D., Gyu-Seong Choi, M.D., Ph.D.[1]

Affiliations:

[1]Department of Surgery, Samsung Medical Center, Seoul, Korea

[2]Research Institution for Future Medicine, Samsung Medical Center, Seoul, Korea

[3]Department of Surgery, Myongji Hospital, Goyang, Korea

[4]Department of Surgery, Yeungnam University Medical Center, Deagu, Korea

Presenting author:

Namkee Oh, M.D.

Department of Surgery, Samsung Medical Center, Seoul, Korea

e-mail: ngnyou@gmail.com

phone number: +81-10-5100-1905

Keywords:

Liver Surgery, Computer-Assisted Surgery, Deep Learning

Key information:

1. Research question: Does a deep learning-base model, using real-time segmentation, assist in intraoperative navigation during minimally invasive liver surgery?
2. Findings: The MiT-B3 model demonstrated robust performance for segmenting avascular planes and vascular structures, maintaining similar performance in external validation with multicenter video data.
3. Meaning: This study indicates that deep learning could enhance intraoperative navigation for liver mobilization in minimally invasive liver surgery, potentially increasing the adoption and safety of the procedures.

MANUSCRIPT (582 words)

Introduction

With the introduction of laparoscopic surgery, there has been a paradigm shift from open surgery to minimally invasive surgery (MIS) in many areas of surgery[1]. In liver surgery, the first laparoscopic wedge resection was reported in 1991, followed by laparoscopic major hepatectomy in the 2000s and laparoscopic donor hepatectomy in the 2010s[2-4]. Despite these advancements, minimally invasive liver surgery (MILS) has not seen global widespread adoption like other minimally invasive procedures[1]. This is because there is still a significant learning curve to be overcome to achieve proficiency in major MILS[5]. This study was designed to explore the potential of artificial intelligence (AI) to assist in the performance of major MILS by providing intraoperative navigation through real-time segmentation of the safe plane for dissection.

Material and methods

Data collection: A total of 48 videos of pure laparoscopic donor right hepatectomy (PLDRH) were collected from three institutions: Samsung Medical Center (SMC, n=40), Myong-Ji Hospital (MJH, n=5), and Yeungnam University Medical Center (YUMC, n=3). Frames were extracted every 10 seconds during the right liver mobilization process, resulting in 2,753 frames for analysis.

Annotation process: Three surgical scopists(registered nurses) annotated the avascular plane and vessels (inferior vena cava, inferior hepatic vein, short hepatic vein, and diaphragm vein) in these frames (Figure 1). Subsequently, three liver surgeons reviewed and revised these annotations for accuracy.

Model development: The segmentation model was developed using the U-Net architecture, with different encoders for performance comparison: ResNet and EfficientNet for convolutional neural network (CNN) architectures, and MiT (Mix Transformer) from SegFormer for transformer-based

encoding[6-9]. All images were resized to 256 x 256 pixels and underwent z-normalization for each channel. Data augmentation techniques such as rotation, flip, and transposition were applied to enhance model robustness.

Training and validation: The model was trained over 100 epochs with a batch size 8 using the AdamW optimizer and a learning rate of 1e-4. The Dice CE loss function, which combines the benefits of cross-entropy and the Dice coefficient while excluding background loss, was employed. A five-fold cross-validation was performed using the 40 videos from SMC. External validation was conducted with 8 videos from MJH and YUMC.

Performance Evaluation: Model performance was assessed using metrics such as the Dice similarity coefficient (DSC), intersection of union (IOU), precision, recall, and specificity.

Results

The performance of each model, trained and validated using 40 surgical videos from SMC, is presented in Figure 3 as the parameter size varies. Specifically, the models with approximately 45M parameter are summarized in the Table1. Among the tested models, the MiT-B3 model demonstrated superior performance. In the internal validation, it achieved a DSC of 0.680 ± 0.019 for the avascular plane and 0.704 ± 0.010 for the vascular structures. Upon application to the eight external datasets, the model maintained robust performance, showing a DSC of 0.656 ± 0.003 for the avascular plane and 0.707 ± 0.010 for the vascular structure (Table2).

Discussion and Conclusion

This study validated the use of artificial intelligence (AI), particularly the MiT-B3 model, for intraoperative navigation in MILS, demonstrating its effectiveness in accurately segmenting key anatomical structures. Its robust performance across different institution suggests potential for wider clinical application, enhancing surgical precision and safety. Despite promising results, the study's focus

on PLDRH and a small sample size limits generalizability. Future studies should aim to include a larger, more diverse set of videos from various surgical procedures and institution to enhance the robustness and generalizability of the AI model. In conclusion, this study highlights the potential of deep learning to improve intraoperative decisions, enhancing precision and safety.

References

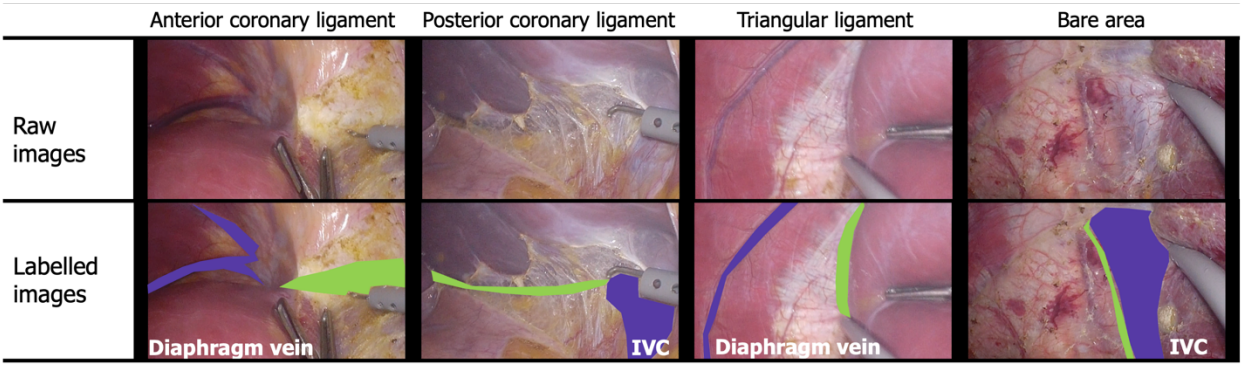
1. Ferrari D, Violante T, Novelli M, et al. The death of laparoscopy. *Surgical endoscopy*. 2024:1-12.
2. Reich H, McGlynn F, DeCaprio J, Budin R. Laparoscopic excision of benign liver lesions. *Obstetrics and gynecology*. 1991;78(5 Pt 2):956-958.
3. Cherqui D, Soubrane O, Husson E, et al. Laparoscopic living donor hepatectomy for liver transplantation in children. *The Lancet*. 2002;359(9304):392-396.
4. Rhu J, Choi GS, Kim JM, Kwon CHD, Joh JW. Complete transition from open surgery to laparoscopy: 8-year experience with more than 500 laparoscopic living donor hepatectomies. *Liver Transplantation*. 2022;28(7):1158-1172.
5. Samstein B, Griesemer A, Halazun K, et al. Pure laparoscopic donor hepatectomies: ready for widespread adoption? *Annals of Surgery*. 2018;268(4):602-609.
6. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Springer; 2015:234-241.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016:770-778.
8. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. PMLR; 2019:6105-6114.
9. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*. 2021;34:12077-12090.

Disclosures

All authors have no conflict of interests related to this work.

APPENDIX

Figure1. Pixel level segmentation of avascular plane and vascular structures of ground truth



* IVC, inferior vena cava

Figure2. Internal and external validation

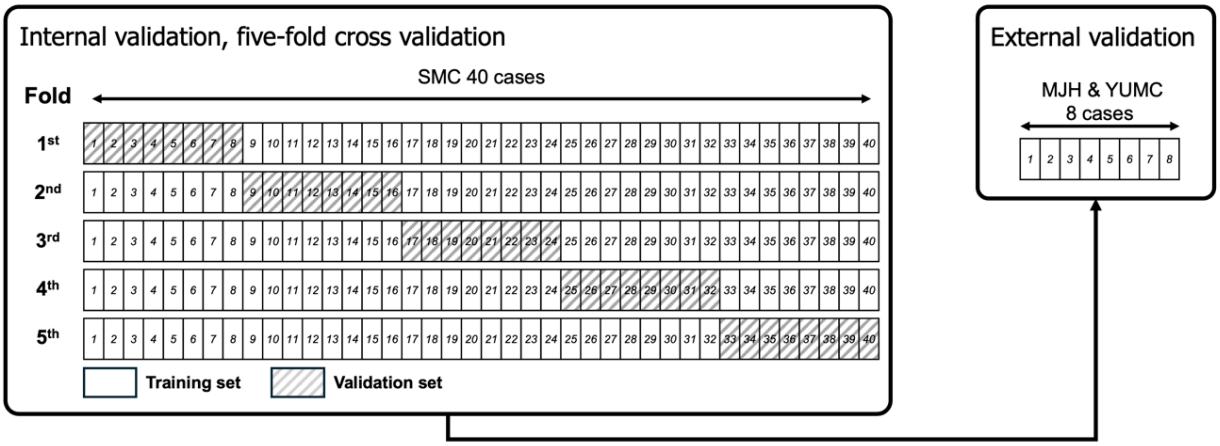


Figure3. Performance comparison for the models

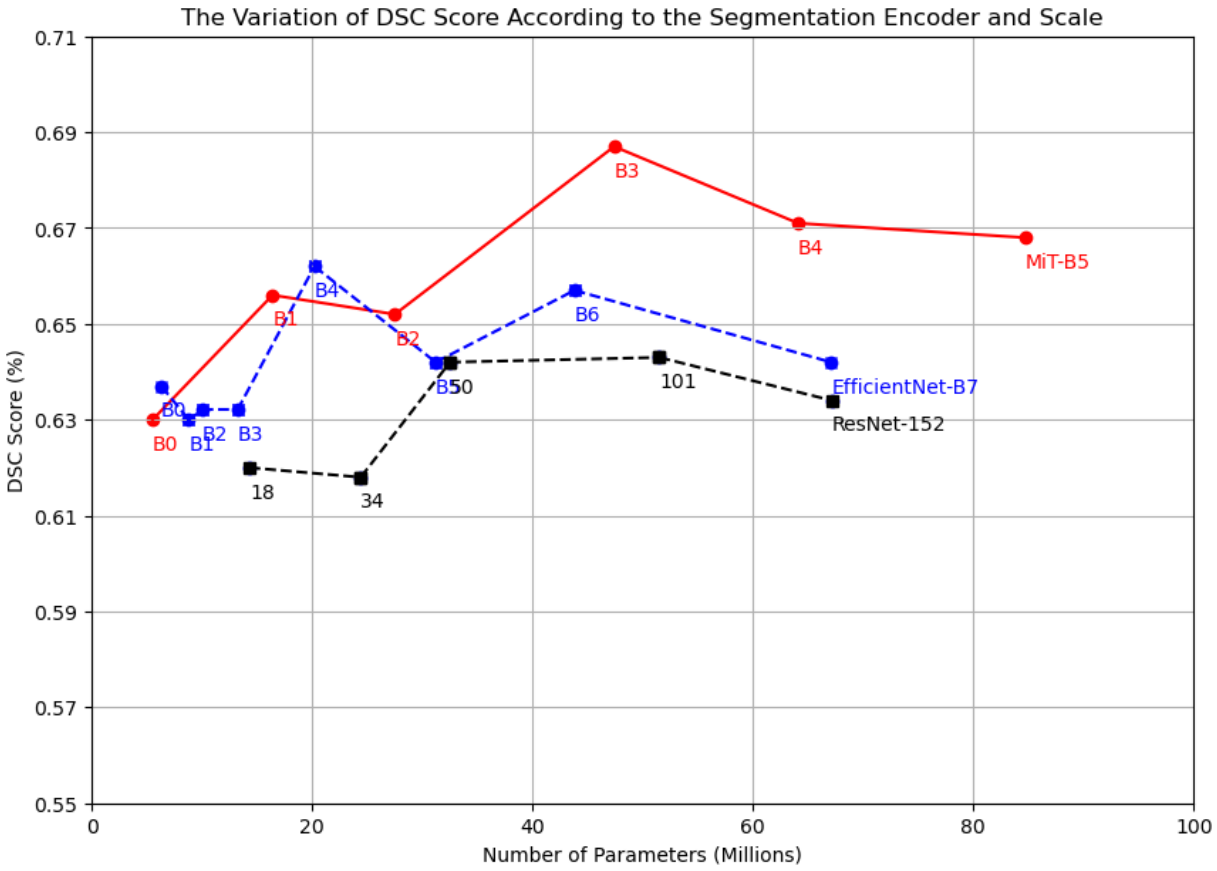


Figure4. Results of segmentation for avascular plane and vascular structures

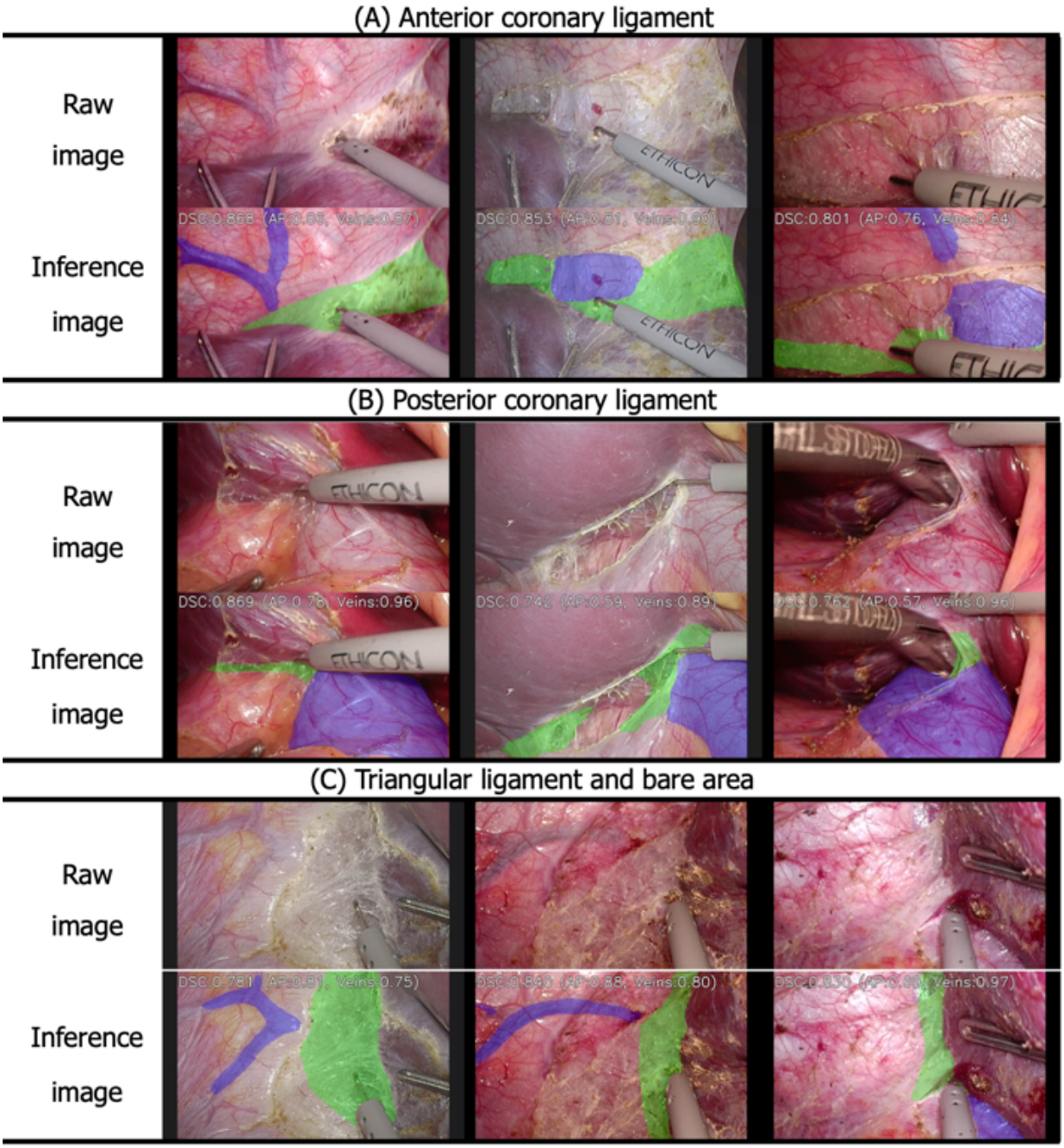


Table 1. Comparative performance of the models for internal and external validation with 8:2 datasplit

Internal validation						
		DSC	IOU	Precision	Recall	Specificity
MiT-b3	Avascular plane	0.661	0.527	0.699	0.698	0.979
	Vascular structures	0.713	0.643	0.673	0.652	0.990
EfficientNet-B6	Avascular plane	0.639	0.505	0.701	0.657	0.981
	Vascular structures	0.674	0.609	0.675	0.502	0.989
ResNet101	Avascular plane	0.648	0.518	0.656	0.729	0.972
	Vascular structures	0.637	0.567	0.583	0.617	0.988
External validation						
	DSC	IOU	Precision	Recall	Specificity	
MiT-b3	Avascular plane	0.654	0.512	0.606	0.788	0.965
	Vascular structures	0.711	0.662	0.640	0.507	0.994
EfficientNet-B6	Avascular plane	0.641	0.499	0.622	0.741	0.970
	Vascular structures	0.701	0.662	0.711	0.340	0.995
ResNet101	Avascular plane	0.602	0.463	0.562	0.747	0.958
	Vascular structures	0.677	0.636	0.691	0.369	0.996

DSC, Dice similarity coefficient; IOU, intersection of union

Table 2. DSC of 5-fold cross validation and external test data for the MiT-b3

		1st fold	2nd fold	3rd fold	4th fold	5th fold	Average (SD)	External validation (SD)
MiT-b3	Avascular plane	0.661	0.697	0.705	0.656	0.683	0.680 (0.019)	0.656 (0.003)
	Vascular structures	0.713	0.734	0.741	0.660	0.671	0.704 (0.032)	0.707 (0.010)

DSC, Dice similarity coefficient; SD, Standard deviation

Table S1. Hyperparameters

Items	Value	
Network Architecture	U-Net	
Backbone Network	MiT-B0~B5, EfficientNet-B0~B7, All of the encoders were pretrained with ImageNet	
Image Size	256 x 256 x 3	
Batch Size	8	
Learning Rate	1e-04	
Number of Epochs	100	
Optimizer	AdamW	
Loss Function	Combination of Dice Loss and Cross-entropy Loss: Dice loss + Cross-entropy loss (without background)	
Number of Classes	3 (Activation Layer: Softmax Function)	
Augmentation	At Training # Normalization *- Normal distribution (mean, std)* *- Resize (256 x 256)* # Geometric transformation *- RandomRotate90* *- Flip* *- Transpose* *- ShiftScaleRotate*	At Test (Inference) # Normalization *- Normal distribution (mean, std)* *- Resize (256 x 256)*

Validation of instance segmentation of instruments during laparoscopic cholecystectomy using MedSAM

Authors:

AS Soares[1,2], C Barata[3,5], C Santiago[3], J Godinho[4], TM Rodrigues[2], LT Castro[1], S Bano, R Rocha[1], P Alves[1], P Mira[1], J Costa[4], M Chand[1], D Stoyanov[1]

Affiliations:

1 – University College London, London, United Kingdom

2 – Hospital Prof. Dr. Fernando Fonseca, Lisbon, Portugal

3 – Instituto de Sistemas e Robótica, LARSyS, Instituto Superior Técnico, Lisbon, Portugal

4 – University of Lisbon, Lisbon, Portugal

5 – LisbonELLIS Unit

Presenting author:

Antonio Sampaio Soares, Hospital Prof. Dr. Fernando Fonseca, Lisbon, Portugal and University College London, United Kingdom; antonio.soares.17@ucl.ac.uk; +351963864555

Keywords:

Instance segmentation, MedSAM, instrument, laparoscopic cholecystectomy

Key information:

1. *Research question: Can medical students and core trainees achieve accurate instance segmentation of instruments using MedSAM?*
2. *Findings: A high dice similarity coefficient was achieved using MedSAM for instance segmentation. There was a significantly higher average time to completion of the annotation using this tool.*
3. Meaning: *MedSAM used by junior doctors shows good accuracy in instance segmentation of common surgical instruments. As such, it is a possible tool to scale high quality annotation of surgical video. Further work is warranted to reduce computational requirements.*

MANUSCRIPT

Introduction

Currently, the best performing instance segmentation algorithms in computer vision employ supervised training. It is very difficult to extend these methods to surgical video analysis, given constraints in surgeon hours and expertise. Therefore, it is desirable to the ability to scale the ability to create high quality annotations.

MedSAM (1) is a state-of-the-art computer vision model which provides an open-source graphic user interface (GUI) to perform semantic segmentation of medical images. It has been trained on 1 570 263 medical images, of which 27095 (1.7%) are laparoscopic images. To our knowledge, this model has not been validated in the task of instance segmentation. We aim to test this model on the task of instance segmentation of instruments used during laparoscopic cholecystectomy through the development of an open-source graphic user interface.

Material and methods

The GUI to deploy MedSAM was edited to add the following functionalities: labelling, the use of point-based segmentation, timestamping and the user experience optimised. This tool was used to segment instances of the various instruments used in laparoscopic cholecystectomy: dissector, grasper, hook, clipper, scissors, irrigator, and specimen bag, in an in-house dataset. Ground truth was annotated by a senior trainee using the Pixel Annotation Tool. MedSAM was used through the custom user interface by one medical student and one core trainee.

Annotation accuracy was measured through the Dice Similarity Coefficient according to the recommendations by Maier Hein et. Al (2), for every instrument class and compared. Number of missed annotations is counted per class.

Results

A total of 961 images were analysed, from 5 patients operated in Lisbon as part of the Surg_Cloud project. The number of instruments per image was as follows: 1 in 411 images, 2 in 481 and 3 in 23 images. The detailed breakdown of the similarity assessment can be seen in table 1 and figure 1. Manual annotation required approximately 37 seconds per image. Annotation using the MedSAM GUI on a cpu required approximately 60 seconds per image, of which approximately 20.2 seconds were for image encoding.

Discussion and Conclusion

The use of MedSAM GUI shows promise due to the high similarity coefficients achieved by doctors with less experience in identifying instruments. While for instruments with very constant shape this was uniform, a bigger variability was present for the fenestrated grasper, probably in relation with the change of the instrument shape. The use of the GUI is encouraged by providing the code to use this tool.

The difference in time taken for annotation is a clear limitation, which we hypothesize to be due to compute requirements. The deployment of these models in edge devices is a significant bottleneck. It warrants further study on the impact of GUI optimisation and different annotation instructions.

.

References

1 - Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, *15*(1), 654.

2 - Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., ... & Jäger, P. F. (2024). Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 1-18.

Disclosures

No relevant conflicts of interest to disclose

APPENDIX

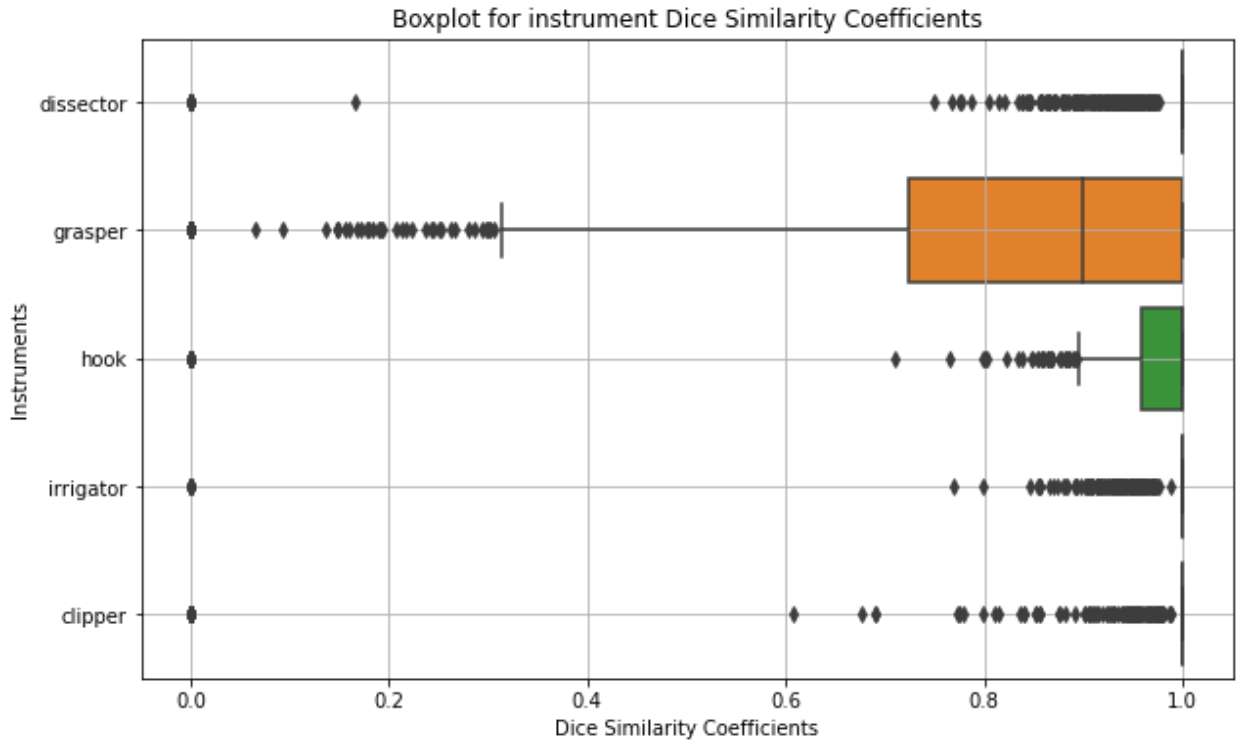


Figure 1: boxplot of the Dice similarity coefficient scores for each instrument

Table 1: Summary Statistics of Dice Similarity Coefficients (DSC) for Various Surgical Instruments

Instrument	n	Median DSC	IQR
Dissector	193	0.922	(0.870, 0.947)
Grasper	659	0.841	(0.283, 0.902)
Hook	346	0.942	(0.918, 0.961)
Irrigator	123	0.933	(0.902, 0.952)
Clipper	125	0.931	(0.778, 0.963)

Advancing Laparoscopic Surgery: Addressing Adoption Challenges with User-Centric Augmented Reality Solutions

Authors:

Pooja P Jain [1] [2], Jaswanth G [2], Chetan Shirvankar [2] [3], Anusuiya [2], Subhamoy Mandal [2]

Affiliations:

[1] B.C Roy Technology Hospital, Kharagpur, West Bengal, India
[2] School of Medical Science and Technology, Indian Institute of Technology Kharagpur, West Bengal, India
[3] Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, Canada

Presenting author:

Dr. Pooja P Jain

B.C Roy Technology Hospital, Kharagpur, School of Medical Science and Technology, Indian Institute of Technology Kharagpur

+91 9916041433

pooja.jain@kgpian.iitkgp.ac.in

Keywords:

laparoscopic surgery, precision surgery, augmented reality surgical navigation, artificial intelligence in surgery, Technology readiness index

Key information:

1. **Research question**: Can a user-friendly video-based AR surgical navigation system designed for ease of adoption by novice users address the challenges hindering AR adoption in Indian laparoscopic surgeries?
2. **Findings**: The study found surgeons in India face cost, user experience, and comfort barriers to AR adoption. Our video-based AR tool showed promise for ease of use, particularly in deep laparoscopic surgeries.
3. **Meaning**: The developed user-friendly video-based AR tool for laparoscopic surgery demonstrated promising initial usability, particularly for deep-seated lesions, suggesting its potential to address these identified hurdles and facilitate broader AR integration.

MANUSCRIPT *(included in the word count)*

Introduction

Augmented reality (AR) in surgery equips doctors with real-time visualizations of a patient's internal structures, guiding their decisions with the help of invaluable data. AR facilitates meticulous surgical planning and rehearsal, allowing surgeons to strategize solutions for potential complications virtually [1]. Although optical see-through methods of AR have gained increasing popularity, the cost involved, lack of real-world stimulus, and the discomfort associated with wearing AR/VR systems have led to a low level of acceptance among surgeons [2]. A previous clinical feasibility study by Dennler C et al. on hololens-based AR system utilizing a questionnaire based study showed that impaired sight caused by the device and surgical steps requiring other instruments were the main hindering factors for integration of the wearable systems in surgery. Keeping this in mind, a video-based AR system can be preferable as it ensures little effort for the surgeon to adapt to the technology, while a computer system does most of the work before the surgery. For widespread inception in a new setting, it becomes essential to first assess the requirements and limitations faced by the first line users and to address them. Previous studies have reported the preferences of surgeons for different visualization techniques and assessed their opinion of potential impact of surgical AR [3]. The goal of this study was to analyze the awareness and challenges related to AR adoption in surgical practice in India and to develop an intuitive user-friendly video-based AR tool experimented by novice users for a patient-specific surgical navigation in laparoscopic surgeries.

Material and methods

A cross-sectional user-study was conducted to analyze the adoption and fears relating to uptake of the AR systems among Indian doctors. It was a pilot study constituting a small study group of (36?) 23 participants, includingphysicians researching in technology(n=16), ,residents in training (n=9) and surgeons (n=5). Based on the feedback provided, especially the barriers towards adoption of AR, a simplified working phantom model of an AR surgical navigation system in the preclinical laboratory was developed and experimented by healthcare professionals. A similar setup is reported for performance evaluation of trainee surgeons in a phantom model with a Laparoscopic Augmentation System (LAS) added with a computer connected to a tracking system [4]. Our workflow was based on the video-based AR system and used the existing laparoscopy monitor to provide relevant anatomical information. The 3D rendered model of the organ of interest and the lesions' information was available to the doctor with keyboard-based controls as shown in Figure 1. It was coupled with a modified U-Net segmentation model [5] that provided object tracking of the laparoscopy instruments to ensure that the instrument did not enter the danger zone, or the area around the critical structures. Warnings regarding the surgical procedure were displayed on the laparoscopy monitor.

Results

Adapted Technology Readiness Index (TRI) was calculated to assess the perceived readiness of the present healthcare system to adopt AR technology in surgery by different predictor variables as visualized in Figure 2. ANOVA analysis gave a p-value of (>0.05) for each of the questions, proving that there is no significant difference in TRI means of the three subgroups (Table 1). Although statistically

insignificant , there were differences in subjective opinions among different subgroups of the study. Researchers were worried about cost and infrastructure, residents had moderate concerns about usability and cost, and surgeons prioritized cost-effectiveness but were enthusiastic about AR for training. After experimenting with the developed simulation setup, some of the healthcare professionals expressed the ease of use of the setup and confirmed its high usability in surgeries involving deep or subsurface lesions.

Discussion and Conclusion

This study explores key obstacles hindering the adoption of AR in Indian surgical setting and gives a potential base level solutions targeted to the novice users. The proposed model of the video-based AR system got a positive feedback from the participants of the study. An intuitive interface for all skill levels, combined with minimal modifications to existing surgical equipment, could accelerate the practical integration of AR technology. Future research should involve larger, quantitative user trials with case-specific adaptations of the simulation model.

References

[1] Vashishth TK, Sharma V, Sharma KK, Kumar B, Chaudhary S, Panwar R. Virtual reality (VR) and augmented reality (AR) transforming medical applications. In: AI and IoT-Based Technologies for Precision Medicine. IGI Global; 2023:324-348. doi:10.4018/979-8-3693-0876-9.ch020

[2] Edwards PJ`` E, Chand M, Birlo M, Stoyanov D. The Challenge of Augmented Reality in Surgery. In: Atallah S, ed. Digital Surgery. Springer International Publishing; 2021:121-135. doi:10.1007/978-3-030-49100-0_10

[3] Heiliger C, Heiliger T, Deodati A, et al. AR visualizations in laparoscopy: surgeon preferences and depth assessment of vascular anatomy. Minimally Invasive Therapy & Allied Technologies. 2023;32(4):190-198. doi:10.1080/13645706.2023.2219739

[4] Heiliger C, Heiliger T, Deodati A, et al. Phantom study on surgical performance in augmented reality laparoscopy. International Journal of Computer Assisted Radiology and Surgery. 2023;18(8):1345-1354. doi:10.1007/s11548-022-02809-7

[5] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; 2015:234-241. doi:10.1007/978-3-319-24574-4_28

[6] Dennler, C., Bauer, D.E., Scheibler, AG. et al. Augmented reality in the operating room: a clinical feasibility study. BMC Musculoskelet Disord 22, 451 (2021). https://doi.org/10.1186/s12891-021-04339-w

Disclosures

No conflicts of interest.

APPENDIX

Figures and Pictures

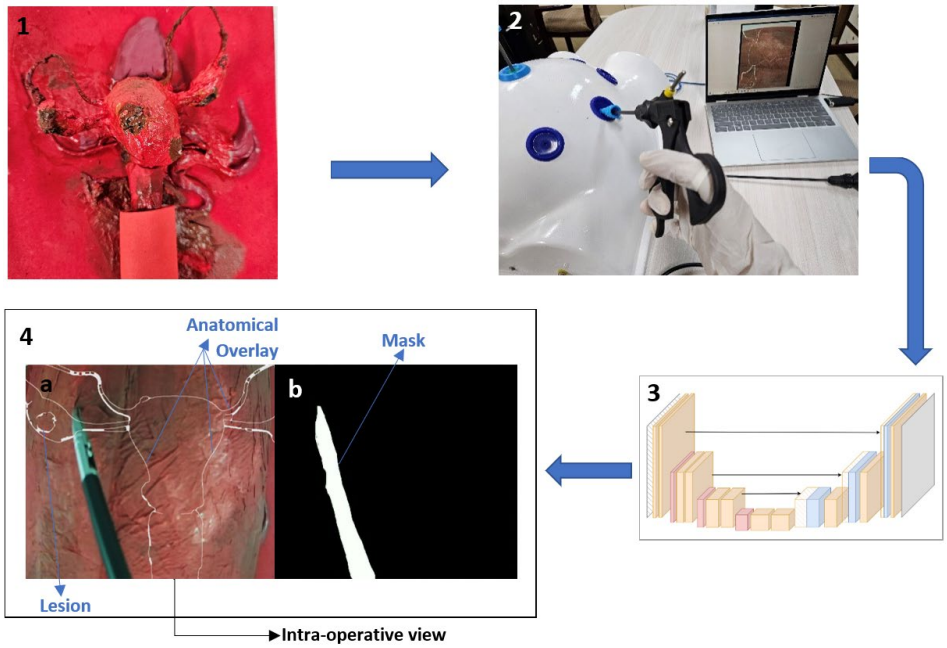


Figure 1: Workflow and setup of the system. (1) Phantom model to simulate the organ if interest (2) Setup which includes the laparoscopy simulator trainer, laparoscopy monitor connected to a processing unit, and laparoscopy instruments (3) U-Net segmentation algorithm used on the live video feed (4) (a) Intraoperative view on laparoscopy monitor of the edge-highlighted target organ superimposed on the surgical field of view (b) The segmented mask for object tracking by the AI algorithm

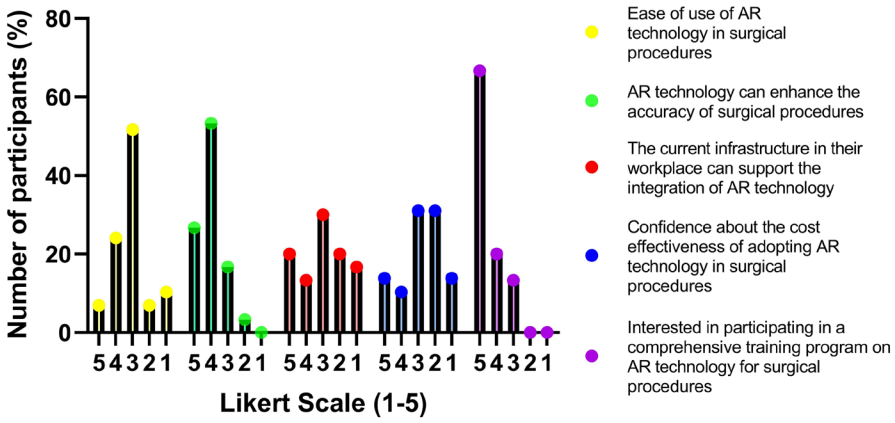


Figure 2: Predictor variables for AR variables assessed by TRI (Technology Readiness Index)

Tables

Table 1: Analysis of Variance (ANOVA) analysis for evaluations of the technological preparedness

Component of Technology Readiness	F-Statistic	P-Value
Ease of Use	1.118	0.342
Accuracy Enhancement	1.813	0.183
Infrastructure Support	0.697	0.507
Cost Effectiveness	1.684	0.205
Training And Education	0.126	0.882

There are no statistically significant differences among the mean scores of the three groups (Researcher, Resident, Surgeon) in any of the investigated categories. This indicates that the views or evaluations of the technological preparedness characteristics addressed in the survey are comparable among the three categories.

Bridging innovation and practice: the journey of FAROS from technical design to in-vivo animal validation

Authors: N.A. Cavalcanti[1], F. Carrillo[1], R. Li[2], K. van Assche[2], A. Davoodi[2], M. Tummers[2], M. Huber[3], F. Teyssere[4], J.A. Perez Velásquez[4,5], A. Massalimova[1], C.J. Laux[6], R. Sutter[7], M. Farshad[6], G. Borghesan[2,8], K. Denis[9], G. Morel[5], T. Chandanson[4], T. Vercauteren[3], E. Vander Poorten[2], P. Fürnstahl[1]

Affiliations:

1. Research in Orthopedic Computer Science, Balgrist University Hospital, University of Zurich, Zurich, 8008, Switzerland

2. Robot-Assisted Surgery Group, Department of Mechanical Engineering, KU Leuven, Leuven, 3001, Belgium

3. School of Biomedical Engineering & Image Sciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom

4. SpineGuard S.A., Paris, France

5. ISIR Laboratory, Sorbonne University, Paris, France

6. University Spine Center Zurich, University Hospital Balgrist, University of Zurich, Zurich, 8008, Switzerland

7. Department of Radiology, Balgrist University Hospital, University of Zurich, Zurich, 8008, Switzerland

8. Core Lab ROB, Flanders Make, Leuven, Belgium

9. Department of Mechanical Engineering, BioMechanics (BMe), Smart Instrumentation, KU Leuven, 3000 Leuven, Belgium

Presenting author: Nicola Alessandro Cavalcanti, MD

Department of Research in Orthopedic Computer Science, Balgrist University Hospital, University of Zurich, Zurich, Switzerland

Email: Nicola.cavalcanti@balgrist.ch, Phone number: 0041445107379

Keywords: Robotic ultrasound, pedicle screw placement, conductivity sensing, lumbar spine, 3D reconstruction

Key information:

1. Research question: Can robotic pedicle screw trajectories be accurately, autonomously executed based on a 3D reconstructed ultrasound navigation by the FAROS robotic system?
2. Findings: In most cases, pedicle screw trajectories can be drilled with clinical acceptance. Further, ultrasound can achieve accurate results for reconstructing anatomical features to navigate said trajectories.
3. Meaning: The FAROS system shows promise for future robotic autonomous applications.

MANUSCRIPT

Introduction

Due to the complex spinal anatomy and the need for high precision, existing spinal navigation systems utilize ionizing radiation to generate high-resolution bone imaging. However, these systems result in a substantial radiation burden to patients and healthcare professionals. In the European Union project FAROS, we aimed to leverage non-visual sensing technology to improve the autonomy of robotic pedicle screw (PS) drilling while eliminating the need for ionizing radiation.[1] We report the results for final validation, including ex-vivo human and in-vivo animal experiments.

Material and methods

The FAROS robotic system integrates a robotic arm (LBR-Med-7, KUKA, Augsburg, Germany) with ultrasound (US) (ML6-15 (15MHz) linear probe, GE Logiq, GE Healthcare, Illinois, USA) for anatomy reconstruction and trajectory planning, optical breathing compensation, a robotic drill, and conductivity sensors (DSG, SpineGuard, Paris, France) for real-time breach detection of the spinal canal while drilling.[2] The FAROS robotic system is based on several steps for robotic spine surgery. Initially, fiducial markers (Clear Guide Medical, Baltimore, MD, USA) were placed on the anatomy for motion tracking using an RGBD camera (ZED-2i, Stereolabs, Paris, France). Next, a US probe was attached to a robotic arm to scan the entire spine using a pre-defined path (Figure 1).[3,4] The bone anatomy of the 2D US images was segmented with U-NET, 3D-reconstructed, and ICP-registered with preoperative CT data containing the preoperative planning information.[5] The end effectors' US probe was then replaced by a drilling tool (Figure 2). After spine exposure, the system drilled autonomously to the preoperatively specified depth. During drilling, DSG monitors conductivity changes to stop drilling if a spinal canal breach is detected.

The following validation studies were performed:

1. Ex-vivo human validation:
   Ten PS trajectories were planned on L1 to L5 vertebrae using preoperative 3D CT models. After executing the workflow, the accuracy of the drilled canals was assessed surgically and radiologically through pedicle palpation and CT image evaluation. The placements were categorized based on their proximity to pedicle walls using the Gertzbein-Robbins (GR) classification.[6]

2. In-vivo animal validation:
   The reliability of the DSG technology was tested on live animals due to the unreliable tissue conductivity signals in ex-vivo specimens. This involved using an in-vivo porcine specimen to compare the DSGs' breach detection capabilities with postoperative CT scans.

3. The accuracy of the 3D US reconstruction:
   This was evaluated in separate ex-vivo human experiments. A 200 x 100 mm skin area was scanned using an S-shaped US protocol to cover vertebrae from L1 to L5, repeated thrice. The US images were registered with a preoperative CT model using an iterative closest point algorithm to evaluate the reconstruction quality of the anatomical features.[7,8]

Results

In the ex-vivo system validation, seven (77.78%) screws were placed with GR grade A, while one screw each achieved grades B and C (Table 1, Figure 3). One screw was excluded due to soft tissue proximity and surgical approach issues. The US navigation faced challenges in the in-vivo experiments due to increased spine curvature and breathing. Time constraints led to skipping preparation steps, resulting in technical errors. The separate evaluation of the 3D US reconstruction accuracy revealed an error of 1.74 ± 0.89 mm.[3]

Discussion and Conclusion

The system demonstrated clinically acceptable outcomes in 8 of 9 pedicle drillings, underscoring its potential for future autonomous robotic applications in spine surgery. Despite the promising results, the prototype is still in its developmental phase, with drawbacks like lengthy setup and preparation times posing challenges for immediate clinical adoption. Efforts are focused on refining the preparation process and advancing toward a minimally invasive technique to fully leverage the systems' capabilities and enhance their clinical applicability.

References

1. EU Horizon 2020 Project FAROS. https://h2020faros.eu/
2. SPG manual drilling probe. https://www.spineguard.com/dynamic-surgical-guidance-technology
3. Li R, Davoodi A, Cai Y, et al. Robot-assisted ultrasound reconstruction for spine surgery: from bench-top to pre-clinical study. *International Journal of Computer Assisted Radiology and Surgery*. 2023/09/01 2023;18(9):1613-1623.
4. Li R, Davoodi A, Niu K, Cai Y, Vander Poorten E. Automatic Robotic Scanning for real-time 3D Ultrasound Reconstruction in Spine Surgery. 2022;
5. Li R, Davoodi A, Cai Y, et al. Development and evaluation of robot-assisted ultrasound navigation system for pedicle screw placement: An ex-vivo animal validation. *Int J Med Robot*. Oct 24 2023:e2590.
6. Gertzbein SD, Robbins SE. Accuracy of pedicular screw placement in vivo. *Spine (Phila Pa 1976)*. Jan 1990;15(1):11-4.
7. Li R, Davoodi A, Cai Y, et al. Robot-assisted ultrasound reconstruction for spine surgery: from bench-top to pre-clinical study. *Int J Comput Assist Radiol Surg*. May 12 2023
8. Assche KV, Li R, Davoodi A, et al. Robotic Path Re-planning for US Reconstruction of the Spine. Manuscript submitted 2024

Disclosures

F. Teyssere, J.A.P. Velásquez, and T. Chandanson are employed by SpineGuard S.A.

APPENDIX

Figure 1: The FAROS robotic system with the robotic ultrasound end effector attached.

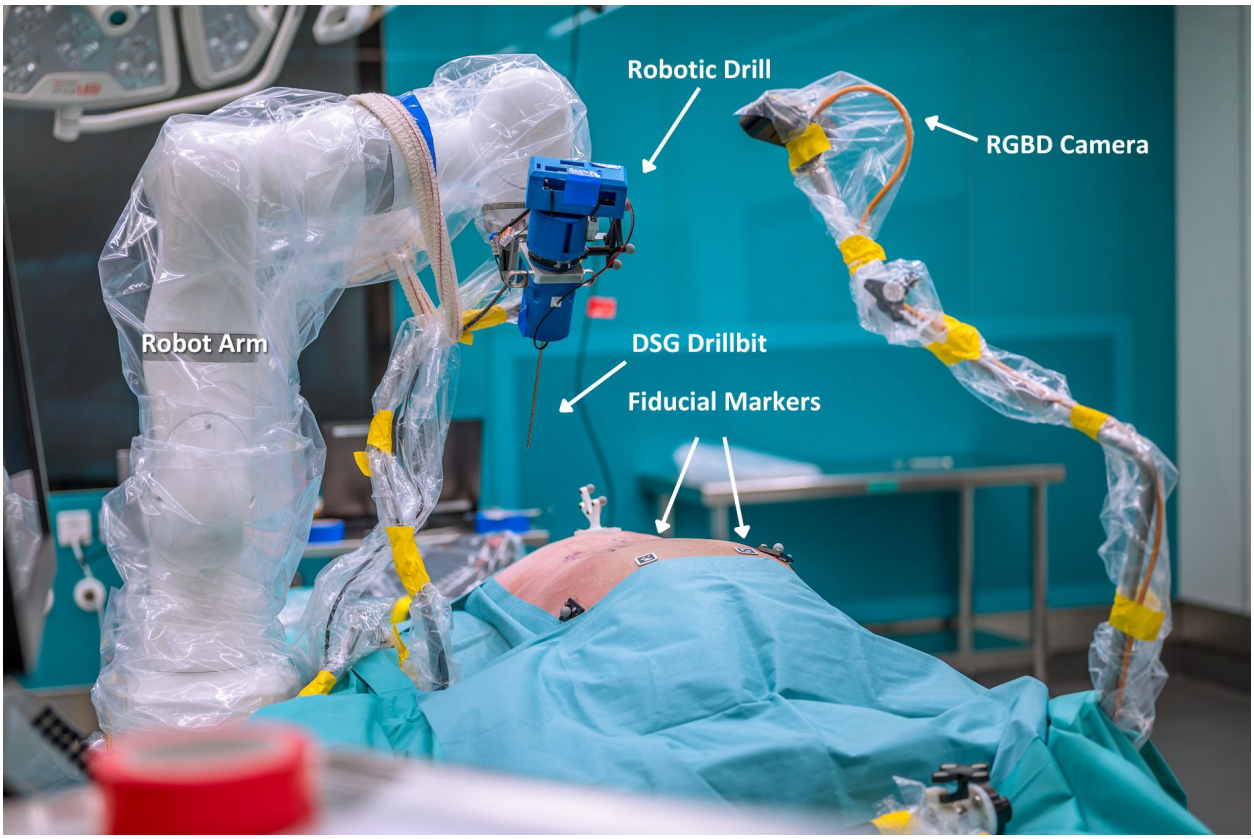


Figure 2: The FAROS robotic system with the robotic drill attached and fiducial markers placed for the optical (RGBD) camera.

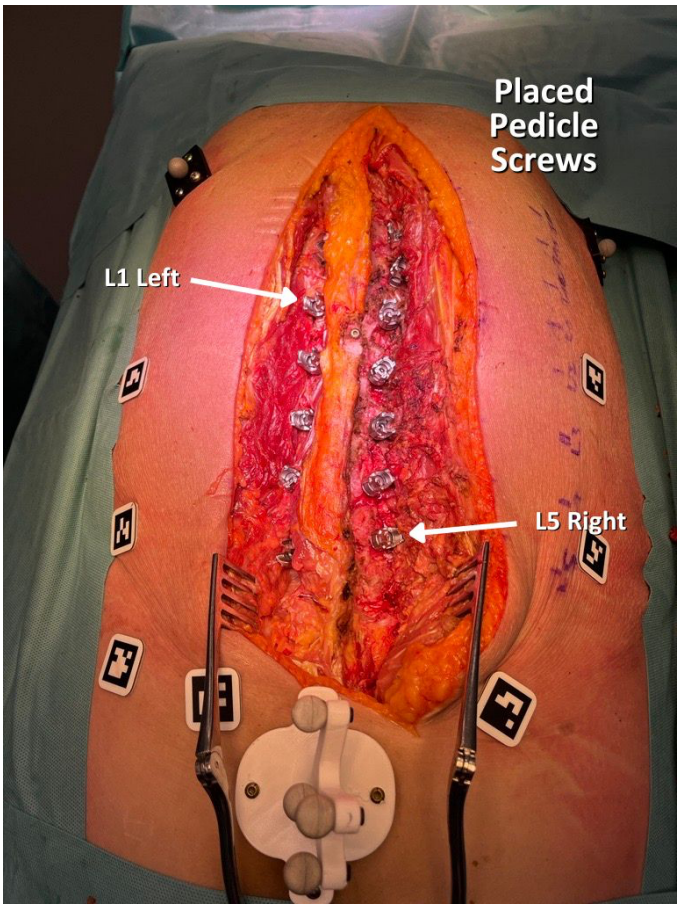


Figure 3: Placed pedicle screws after autonomous robotic drilling from L1 to L5. Thoracic screw placement was only performed for system training and not part of the validation experiment.

Table 1. Gertzbein-Robbins classification for the ex-vivo pedicle drilling experiments.

Pedicle drilling trajectories	*Gertzbein-Robbins Classification*
L1 right	A
L1 left	A
L2 right	A
L2 left	A
L3 right	C
L3 left	B
L4 right	A
L4 left	A
L5 right	A
L5 left	Excluded due to soft tissue proximity/ surgical approach issues

Deep Learning Artificial Intelligence Tool for Automated Radiographic Determination of Posterior Tibial Slope in Patients with Anterior Cruciate Ligament Injuries

Ayoosh Pareek, MD[1]
Yining Lu, MD[2]
Linjun Yang, PhD[2]
Pouria Rouzrokh, MD[2]
Bardia Khosravi, MD[2]
Xuankang Pan, BS[2]
Kelechi R. Okoroha, MD[2]
Aaron J. Krych, MD[2]
Christopher L. Camp, MD[2]

[1]Hospital for Special Surgery Department of Orthopedic Surgery, New York, NY, United States

[2]Mayo Clinic Department of Orthopedic Surgery, Rochester, MN, United States

Presenting Author: Ayoosh Pareek
Email: ayooshp@gmail.com
Phone: 1+ (347) 885-5672

Key Information

Research Question: The purpose of this study is to develop a deep learning artificial intelligence technique for the automated measurement of posterior tibial slope (PTS) from standard lateral knee radiographs.

Findings: PTS corresponds to increased rate of graft failure in ACL deficient knees. Validated human measurement is inefficient and prone to interobserver variability.

Meaning: An accurate and reliable deep learning computer vision algorithm was developed to automate the measurement of PTS on lateral knee radiographs. This tool demonstrated good agreement with human annotations and will be deployed for clinical use on an institution-wide basis. This represents an effective clinical adjunct for the measurement of PTS as part of the preoperative assessment of ACL-injured patients.

Abstract:

Background:

Anterior cruciate ligament reconstruction (ACLR) generally leads to favorable overall outcomes; however, re-current ACL injury (ipsilateral or contralateral) may occur in up to 20-30% of cases[1,4] An increased posterior tibial slope (PTS) corresponds with an increased risk of graft failure following anterior cruciate ligament (ACL) reconstruction. [5] Validated methods of manual PTS measurements are subject to potential interobserver variability and can be inefficient on large datasets. Thus, the purpose of this study is to develop a deep learning artificial intelligence technique for the automated measurement of PTS from standard lateral knee radiographs.

Methods:

A deep learning U-Net model was developed on a cohort of 300 postoperative short leg lateral radiographs from ACLR patients to segment the tibial shaft, tibial joint surface, and tibial tuberosity. The model was trained via a random split following an 80:20 train-validation scheme. Masks for training images were manually segmented and the model was trained for 400 epochs. An image processing pipeline was then deployed to annotate and measure the PTS using the predicted segmentation masks by calculating angle distended by a best-fit line through the tibial shaft mask and a line across the top of the tibial joint surface. Finally, the performance of this combined pipeline was compared to human measurements performed by two study personnel using a previously validated manual technique for measuring PTS on short leg lateral radiographs on an independent test-set of 90 images. An example of this pipeline and the produced masks can be found in Figure 1.

Results:

The U-Net semantic segmentation model achieved a mean Dice similarity coefficient of 0.885 on the validation cohort. The mean difference between human-made and computer-vision measurements via a one-sided t-test was 1.92^{o} (σ = 2.81^{o}, P=0.24). Extreme disagreements between human and machine

measurements as defined by differences ≥5^{o} occurred less than 5% of the time. The model was incorporated into a web-based digital application front-end for demonstration purposes which can provide measurement of a single uploaded image in portable network graphics format in less than 5 seconds.

Discussion/Conclusion:

PTS has long been established as a crucial measurement that has significant implications for knee biomechanics, and increased slope has specifically been correlated to an increased risk of graft failure following ACLR[5]. Therefore, reliable measurements of PTS are integral to the ability to risk stratify patients both during preoperative counseling for ACL surgery as well as during postoperative followup. Presently, the gold standard measurement of PTS requires human annotation of long-leg lateral radiographs. However, these require dedicated views that are not routinely obtained and lack generalizability. While several methods exist for manual annotation on short-leg lateral radiographs of the knee[2,3,5], we developed a fully automated deep learning artificial intelligence that can rapidly process and annotate PTS in an accurate and reproducible manner.

This tool demonstrates good agreement with human annotations and will be deployed for clinical use on an institution-wide basis. This represents an effective clinical adjunct for the measurement of PTS as part of the preoperative assessment of ACL-injured patients.

Key Terms: artificial intelligence, automated, ACL, posterior tibial slope, machine learning, preoperative assessment, radiograph

Disclosures: None relevant to this study

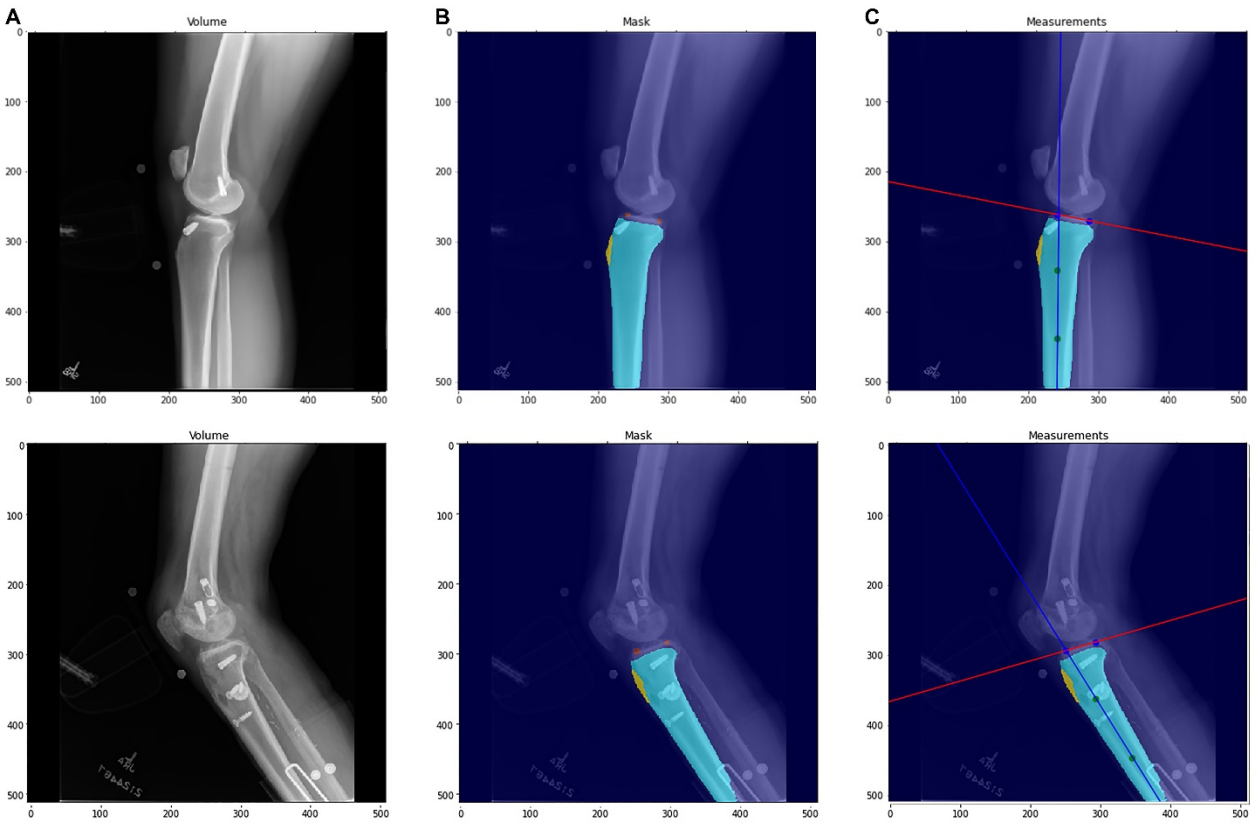


Figure 1: Overview of the pipeline for automatic measurement of posterior tibial slope on short leg lateral radiographs in patients post ACLR. (A) Original radiographic images, (B) overlay of semantic segmentation masks from the U-Net models on the original radiographs, (C) best fit lines through the joint surface and tibial anatomic axis through which the posterior tibial slope can be measured.

1. Allen MM, Pareek A, Krych AJ, et al. Are Female Soccer Players at an Increased Risk of Second Anterior Cruciate Ligament Injury Compared With Their Athletic Peers? *Am J Sports Med.* 2016;44(10):2492-2498.
2. Dean RS, DePhillipo NN, Chahla J, Larson CM, LaPrade RF. Posterior Tibial Slope Measurements Using the Anatomic Axis Are Significantly Increased Compared With Those That Use the Mechanical Axis. *Arthroscopy.* 2021;37(1):243-249.
3. Faschingbauer M, Sgroi M, Juchems M, Reichel H, Kappe T. Can the tibial slope be measured on lateral knee radiographs? *Knee Surgery, Sports Traumatology, Arthroscopy.* 2014;22(12):3163-3167.
4. Gwinner C, Janosec M, Wierer G, Wagner M, Weiler A. Graft Survivorship After Anterior Cruciate Ligament Reconstruction Based on Tibial Slope. *Am J Sports Med.* 2021;49(14):3802-3808.
5. Lin LJ, Akpinar B, Meislin RJ. Tibial Slope and Anterior Cruciate Ligament Reconstruction Outcomes. *JBJS Rev.* 2020;8(4):e0184.

Weakly Supervised Pretraining and Multi-Annotator Supervised Finetuning for Facial Wrinkle Detection

Authors:

Ik Jun Moon[1†], Junho Moon[2†], Ikbeom Jang[3]*

Affiliations:

[1]Department of Dermatology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea

[2]Department of Artificial Intelligence Semiconductor, Hanyang University, Seoul, South Korea

[3]Division of Computer Engineering, Hankuk University of Foreign Studies, Yongin, South Korea

Presenting author:

Full name: Ik Jun Moon

Affiliation: Department of Dermatology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea

Email: ikjun.moon@amc.seoul.kr

Keywords:

Deep learning, Semantic segmentation, Transfer learning, Weakly supervised learning, Facial wrinkle

Key information:

1. Research question: With the growing interest in skin diseases and skin aesthetics, the ability to predict facial wrinkles is becoming increasingly important. This study aims to evaluate whether a computational model, convolutional neural networks (CNN), can be trained for automated facial wrinkle segmentation.
2. Findings: Our study presents an effective technique for integrating data from multiple annotators and illustrates that transfer learning can enhance performance, resulting in dependable segmentation of facial wrinkles.
3. Meaning: This approach automates intricate and time-consuming tasks of wrinkle analysis with a deep learning framework. It could be used to facilitate skin treatments and diagnostics.

†: equal contribution

*: corresponding author

MANUSCRIPT

Introduction

With the growing interest in skin diseases and skin aesthetics, the ability to predict facial wrinkles is becoming increasingly important. Facial wrinkles are significant indicators of aging[1] and can be useful in skin conditions assessment[2], skin care, and early diagnosis of skin diseases[3]. We propose a deep learning-based approach to automatically segment facial wrinkles.

Analyzing extensive collections of images can be exceedingly resource-intensive if each facial wrinkle must be individually assessed. Moreover, the subjectivity inherent in manual segmentation processes can diminish the reliability of research findings and pose a substantial issue.

To address this issue, we effectively combine wrinkle data labeled by multiple annotators to minimize inter-rater variability and utilize these image-label pairs for training our model. Additionally, we enhanced performance using transfer learning that leverages the knowledge from a pretrained model. Unlike traditional methods, our approach utilizes knowledge from a pretrained model through transfer learning for downstream tasks, allowing it to achieve high performance even with limited labeled data. Furthermore, this method significantly reduces the time and cost involved in manually labeling wrinkles, offering substantial advantages over manual methods.

Material and methods

We utilized the FFHQ(Flickr-Faces-HQ)[4] dataset, which consists of 70,000 high-quality face images captured under various conditions. The images are 1024x1024 in size, and we used them without any downsampling or preprocessing. For pretraining, we randomly selected 25,000 images from this dataset. To generate the ground truth for weakly supervised pretraining, we extracted texture maps[5] from the face images and masked out non-facial areas to produce the final texture masks for ground truth (Figure 1-a). For finetuning, 500 face images were randomly selected. The ground truth for supervised finetuning consisted of manually annotated wrinkle masks. The wrinkle annotation process involved three annotators, all highly experienced in image processing and analysis. Recognizing the subjective nature of wrinkle identification, a consistent standard was established prior to the commencement of labeling. During the labeling, specific emphasis was placed on annotating wrinkles in key facial areas including the forehead, crow's feet, and nasolabial folds. To mitigate inter-rater variability and enhance the reliability of the ground truth, a majority voting mechanism was implemented, where only pixels labeled by at least two groups were retained (Figure 1-b). We allocated 80% of the dataset for training, 10% for validation, and 10% for testing.

Figure 2 shows the entire training pipeline. Initially, we trained the segmentation network in a weakly supervised manner using texture masks and subsequently finetuned it in a supervised manner using manually labeled masks. The training was performed using U-Net[6] architecture. During the weakly supervised pretraining stage, the model learns to output texture masks from face images, receiving images as input and producing texture masks as output (Figure 2-a). Then, in the supervised finetuning stage, the model learned to classify each pixel of a face image as either a wrinkle or not, using images and texture maps as inputs and producing two output classes that generate logit values for background and wrinkles, respectively (Figure 2-b).

In the weakly supervised pretraining stage, the model was trained for 300 epochs. Subsequently, supervised finetuning was performed for 150 epochs using manually annotated data. To evaluate the effectiveness of transfer learning in scenarios with limited label data, the supervised finetuning was performed using varying proportions of the whole training dataset: 100%, 50%, 25% and 5%. We used the Jaccard similarity index (JSI) as the evaluation metric, which is defined as follows:

$$JSI = \frac{|A \cap B|}{|A \cup B|}$$

where A is the predicted segmentation, and B is the actual label. This is a suitable metric for evaluating wrinkle segmentation performance as it measures the overlap between the predicted segmentation and the actual label.

Results

Table 1 displays JSI of our method compared to training exclusively with manual data (no pretraining). Our method demonstrates improved performance by 1.92%, 0.35%, 1.06%, and 8.53% for datasets comprising 100%, 50%, 25%, and 5% of the data, respectively. Additionally, our method demonstrates a significant performance improvement over methods pretrained with self-supervised learning techniques (Table 2). This highlights the efficacy of our proposed method in enhancing the accuracy of wrinkle segmentation, particularly in scenarios with sparse data availability. Figure 3 provides a visual performance of our methodology.

Discussion and Conclusion

We propose a reliable ground truth generation strategy and an efficient transfer learning approach. During the weakly supervised pretraining stage, the model learns to highlight skin features such as edges and textures. In the supervised finetuning stage, the model develops the capability to accurately discern wrinkle regions from edges and textures. Through this effective learning strategy, we can enhance the efficiency of wrinkle segmentation even with limited data, thereby achieving reliable wrinkle segmentation outcomes. However, despite majority voting, the subjectivity in wrinkle annotation remains a challenge. We plan to involve dermatologists in the wrinkle annotation process and research reliable label combination methods, such as label smoothing, to improve the reliability of manual wrinkle masks.

References

1. Aznar-Casanova J, Torro-Alves N, Fukusima S. How much older do you get when a wrinkle appears on your face? modifying age estimates by number of wrinkles. Neuropsychology, development, and cognition.Section B, Aging, neuropsychology and cognition. 2010;17(4):406-421.

2. Warren R, Gartstein V, Kligman AM, Montagna W, Allendorf RA, Ridder GM. Age, sunlight, and facial skin: A histologic and quantitative study. J Am Acad Dermatol. 1991;25(5):751-760.

3. Kim K, Choi Y-H, Hwang E. Wrinkle feature-based skin age estimation scheme. In: IEEE Int. Conf. Multimedia and Expo. IEEE; 2009. pp. 1222–1225.

4. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. arXiv.org. 2019.

5. Yoon H, Lee J, Yoo S. Semi-automatic labeling and training strategy for deep learning-based facial wrinkle detection. The Institute of Electrical and Electronics Engineers, Inc.(IEEE) Conference Proceedings. 2022.

6. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. arXiv.org. 2015.

7. Yu C, Wang J, Chao P, Gao C, Yu G, Sang N. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. arXiv.org. 2018.

Disclosures

There is no disclosure to report.

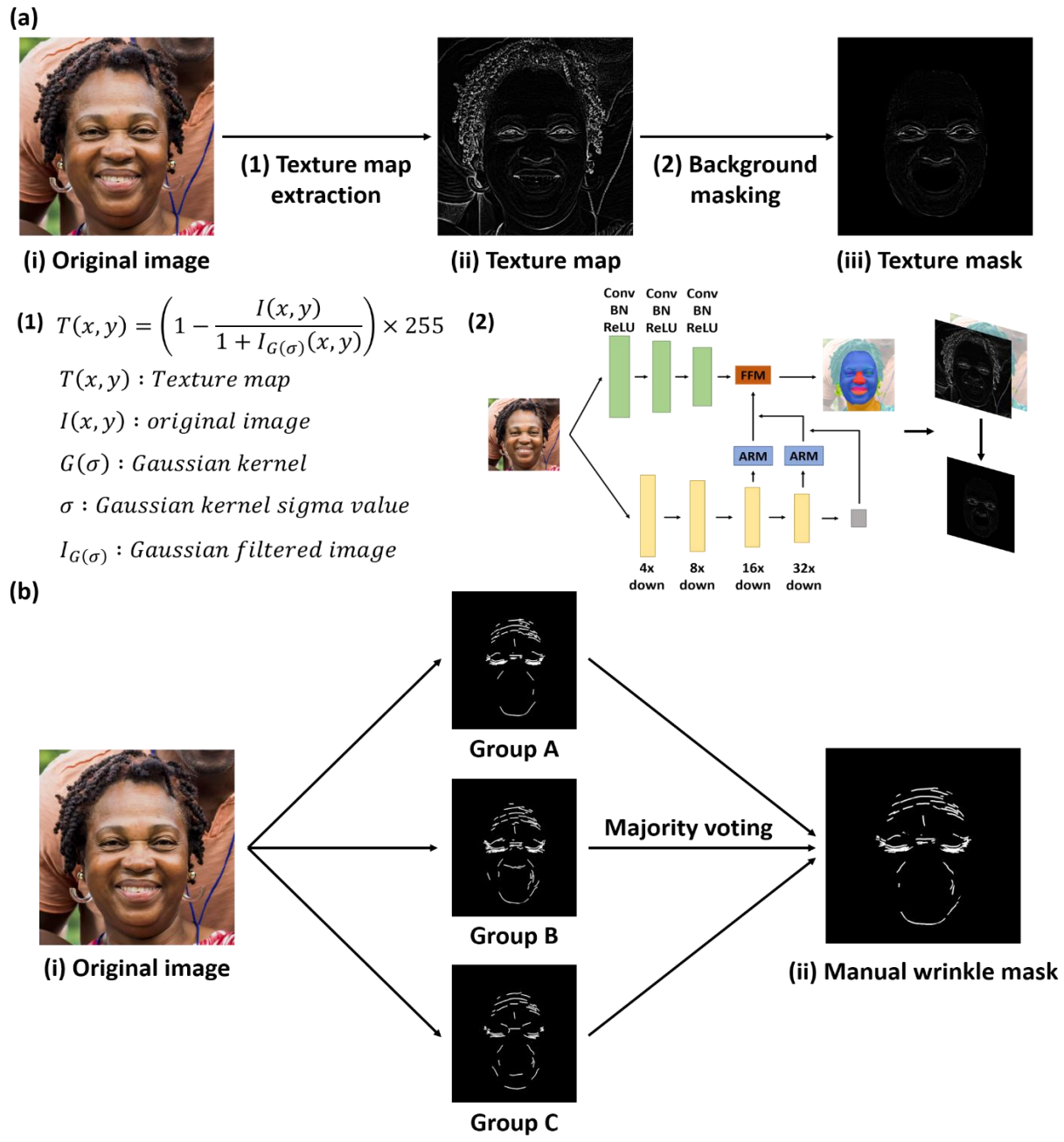


Figure 1: Ground truth wrinkle generation pipeline. (a) denotes the generation process of ground truth used during the weakly supervised pretraining stage. A texture map is extracted from the face image using a filter based on a Gaussian kernel. Since these texture maps contain numerous false positives from backgrounds, non-facial areas are subsequently masked using a face-parsing deep learning model based on the BiSeNet[7] architecture to produce the final texture mask. (b) denotes the ground truth generation process for the supervised finetuning stage. Three annotators each annotate the face images, and through majority voting, these annotations are combined to create the final manual wrinkle mask.

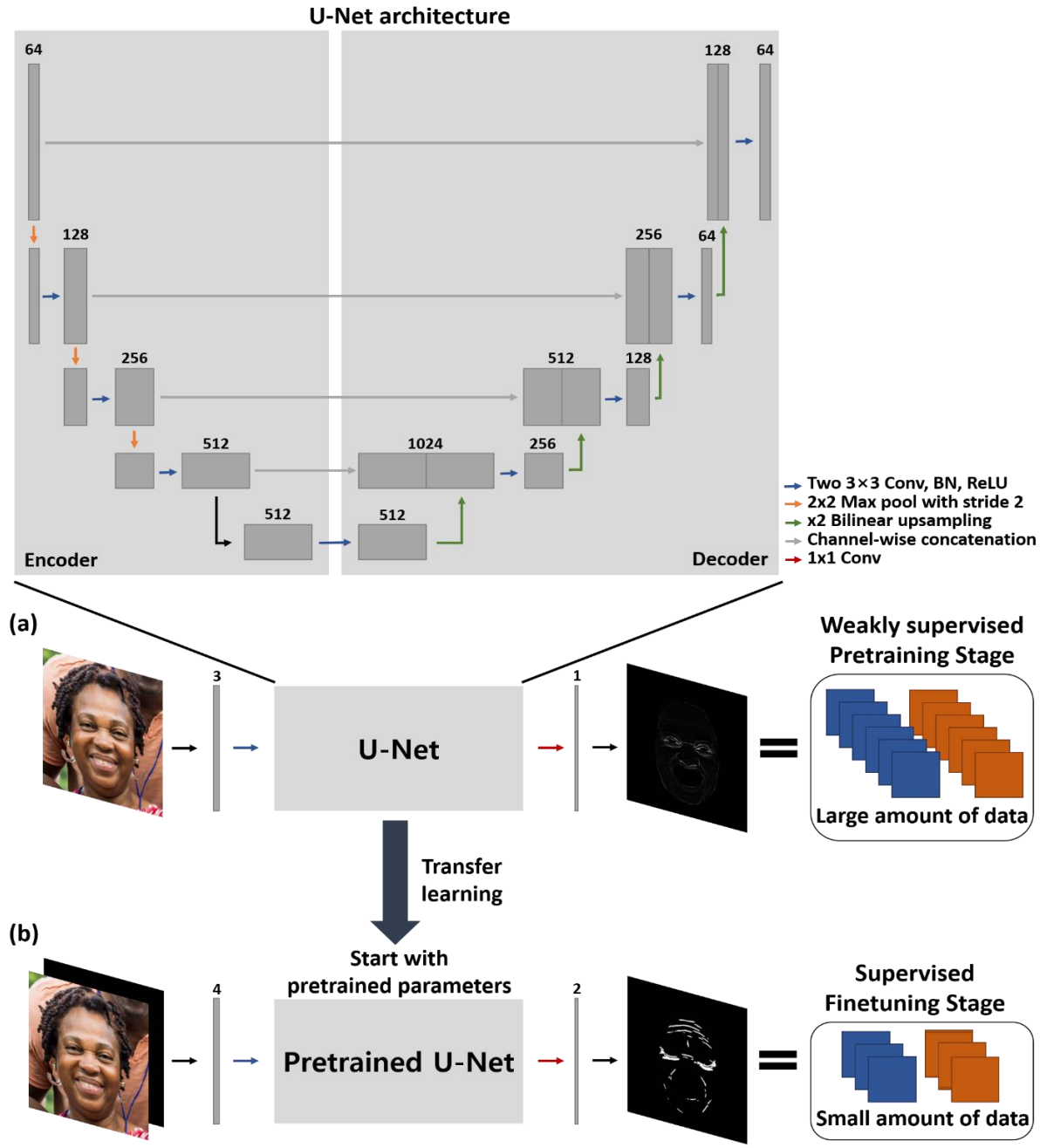


Figure 2: Weakly supervised pretraining and supervised finetuning with multi-annotator labels for facial wrinkle segmentation. (a) denotes the weakly supervised pretraining stage, where the segmentation network learns to extract texture masks from RGB face images. The model receives 3-channel RGB face images as input and produces 1-channel texture masks as output. (b) denotes the supervised finetuning stage, during which the segmentation network learns to extract facial wrinkles from RGB face images and texture masks. The model receives 3-channel RGB face images and 1-channel texture masks as input and produces 2-channel output classes, each generating logit values for background and wrinkles, respectively. By finetuning the weights of the model, which was trained with generic face texture extraction capabilities, we specifically enhanced its ability to detect facial wrinkles using manual data.

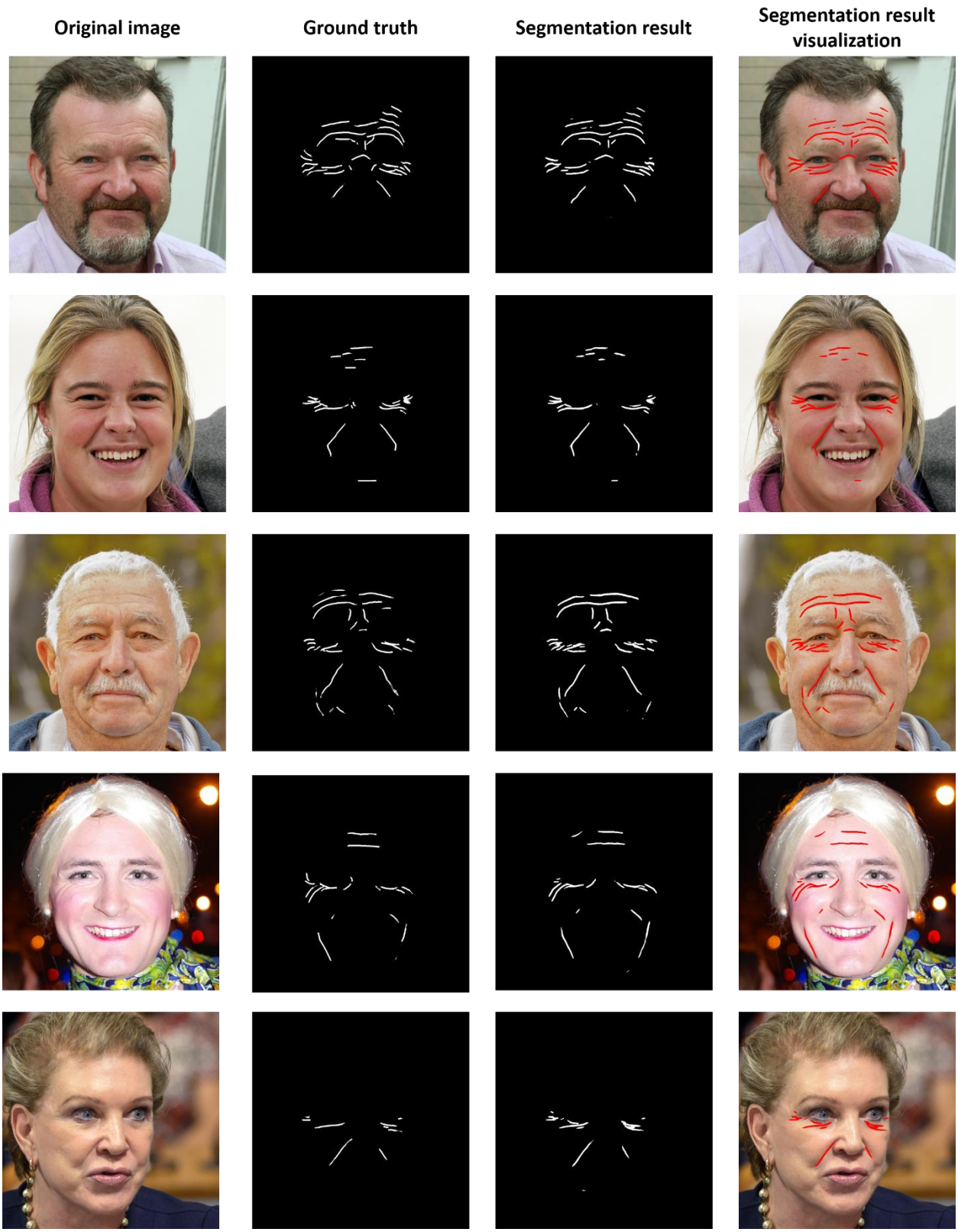


Figure 3: Test image samples and their predicted wrinkles compared with ground truth annotation.

Table 1: Segmentation performance comparison between a model trained exclusively with manually labeled wrinkle data (No pretraining) and our proposed method, which utilizes weakly supervised pretraining on filtered data followed by transfer learning. In the supervised finetuning stage, each method was trained on 100%, 50%, 25%, and 5% of the training dataset.

Method	*100% (400)*	*50% (200)*	*25% (100)*	*5% (20)*	n_{params}
No pretraining	0.4319	0.4120	0.3594	0.2608	17.263M
Ours	**0.4511**	**0.4155**	**0.3700**	**0.3461**	17.264M

(#): number of datasets

Table 2: Model performance by different pretraining strategies. Jaccard similarity index illustrates the performance of models pre-trained using various self-supervised learning techniques compared to our proposed method. Our approach significantly outperforms other methods, demonstrating that our pretraining strategy is highly suitable for downstream tasks in wrinkle detection.

Method	*Reconstruction*	*Deblur*	*Denoise*	*Super-Resolution*	*Ours*
Jaccard similarity index	0.4334	0.4393	0.4364	0.4384	**0.4511**

Preeclampsia screening by minimally trained operators on ultrasound using Artificial Intelligence, a paired diagnostic pilot study

Authors: Saad Slimani MD[1,2], Maryam El Boujamaaoui[2], Leila Noureddine MD[1,2], Mourad Gridach[3], Fatim-Ezzahra Ouamer MD[1], Ahmed Elmoctar MD[1], Amine Kassimi[2], Salaheddine Hounka[2], Amine Lamrissi MD[1], Jalal Mohamed MD[1], El Houssine Bouyakhf[2]

Affiliations:

1: Ibn Rochd University Hospital, Casablanca, Morocco

2: Deepecho, Rabat, Morocco

3: University of Oxford, UK

Presenting author:

Saad Slimani MD, Ibn Rochd University Hospital & Deepecho | saad.slimani@deepecho.io | +212661054053 (WhatsApp and personal number)

Keywords:

Preeclampsia – Ultrasound – Deep Learning – Artificial Intelligence

Key information:

1. **Research question/hypothesis:** *Hypertensive disorders of pregnancy (HDP) is one of the largest causes of maternal mortality in the world accounting for 14% of pregnancy related deaths. However it is notoriously difficult to diagnose especially in the cases of early and late onset HDP. Artificial intelligence models may have the potential to predict HDP from ultrasound blind sweeps video acquired by minimally trained sonographers, thus democratizing access to effective screening of HDP in the Global South.*
2. **Findings:** *109 patients were recruited for this study in Ibn Rochd University Hospitals of Casablanca's Maternity Department, Morocco. The results showed that using the UniFormer architecture, we were able to achieve a sensitivity, specificity and f1 score of 0.76, 0.71 and 0.75, respectively.*
3. **Meaning:** *On a small cohort of patients using low-cost handheld devices from four manufacturers we were able to achieve encouraging results. To the best of our knowledge this is the first study to explore the use of AI to screen for preeclampsia using ultrasound imaging data only. It results could lead to a novel cost-efficient, accurate, and easy-to-use screening tool for preeclampsia in the Global South and to a better understanding of imaging biomarkers of preeclampsia.*

MANUSCRIPT

Introduction

Hypertensive disorders of pregnancy (HDP) and their severe manifestations like preeclampsia and eclampsia lead to higher maternal and fetal mortality rates. HDP are responsible for 14% of maternal deaths worldwide[1]. The need for new screening tools for HDP arises from the current lack of reliable tests to predict term preeclampsia, which accounts for the majority of disease burden and severe maternal and perinatal morbidity[2]. Recent studies some of which have been led by Google have shown that untrained operators with the help of Artificial Intelligence (AI) can gather valuable information about the pregnancy status with expert-level sensitivity and specificity on gestational age, fetal malpresentation, placenta localization using an easy to learn "Blind Sweeps" approach[3,4]. If that is possible then we hypothesized that would be the same to detect HDP and preeclampsia using the same approach.

Material and methods

Study participants were recruited for this study in Ibn Rochd University Hospitals of Casablanca's Maternity Department, Morocco from October the 23rd 2023 to April the 25th 2024. The protocol consisted of seven B-mode (no Doppler) ultrasound sweeps based on external anatomical landmarks performed by radiology residents with less than one year of experience in fetal ultrasound, using five different handheld portable ultrasound devices (Clarius HC3, Butterfly IQ3, Butterfly IQ+, Vitalscan, Giantman). Ground truth on the HDP status of a patient were taken from medical records including the following criteria: Blood Pressure > 14/9, accompanied by proteinuria after 20 weeks of pregnancy, or evidence of other maternal organ dysfunction: acute kidney injury, liver involvement, neurological complications, hematological complications, uteroplacental dysfunction (such as fetal growth restriction, abnormal umbilical artery Doppler waveform analysis, or stillbirth. The data was split into training (80%) validation (10%), and testing (10%) at the participant level. We explored Recurrent Neural Network(RNN) and Convolutional Neural Network (CNN) architectures as well as a transformer.

Results

109 patients were recruited with an average age of 30 years and 40% (44) of them lived with HDP, note that this high proportion is due to the sampling site, and the fact that study participants were hospitalized. 87 videos were used for training, 11 for validation and 11 for testing. The transformer we used, UniFormer[5] (figure 1) significantly outperformed the RNN and CNN approaches reaching a sensitivity, specificity and f1 score of, 0.76, 0.71 and 0.75, respectively versus 0.62, 0.36 and 0.61 (see table 1).

Discussion and Conclusion

To the best of our knowledge, this is the first study that has explored the use of a "Blind Sweep" approach combine to explore HDP and preeclampsia detection. Our results are encouraging and show that there is untapped information even in Blind Sweeps about the HDP status of a patient, paving the way towards new, cost-efficient and accurate screening total for HDP. B-mode ultrasound are not currently used to screen for preeclampsia because there is no recognized feature for preeclampsia[6]. This study shows that AI approaches can extract new imaging biomarkers that could also help understand the imaging manifestation of preeclampsia. One limitation of our study is its small sample size and monocentric nature, however this is an exploratory study that achieved its goal by showing that AI models are able to detect HDP even from relatively poor quality videos taken by minimally trained operators.

References

1. WHO | Trends in maternal mortality: 1990 to 2010. WHO. Accessed February 14, 2021. http://www.who.int/reproductivehealth/publications/monitoring/9789241503631/en/

2. MacDonald TM, Walker SP, Hannan NJ, Tong S, Kaitu'u-Lino TJ. Clinical tools and biomarkers to predict preeclampsia. *EBioMedicine*. 2021;75:103780. doi:10.1016/j.ebiom.2021.103780

3. Pokaprakarn T, Prieto JC, Price JT, et al. AI Estimation of Gestational Age from Blind Ultrasound Sweeps in Low-Resource Settings. *NEJM Evidence*. 0(0):EVIDoa2100058. doi:10.1056/EVIDoa2100058

4. Gomes RG, Vwalika B, Lee C, et al. A mobile-optimized artificial intelligence system for gestational age and fetal malpresentation assessment. *Commun Med*. 2022;2(1):1-9. doi:10.1038/s43856-022-00194-5

5. Li K, Wang Y, Gao P, et al. UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning. Published online February 8, 2022. doi:10.48550/arXiv.2201.04676

6. Gestational Hypertension and Preeclampsia. Accessed April 29, 2024. https://www.acog.org/clinical/clinical-guidance/practice-bulletin/articles/2020/06/gestational-hypertension-and-preeclampsia

Disclosures

Saad Slimani, and El Houssine Bouyakhf are shareholders and employees of Deepecho Inc. the company that funded the study

Maryam El Boujamaaoui, Leila Noureddine, Salaheddine, Hounka and Amine Kassimi are employee of Deepecho Inc. the company that funded the study

Mourad Gridach, Fatim-ezzahra, Ahmed El Moctar received payment for their work at Deepecho Inc. the company that funded the study

The rest of the authors have nothing to disclose.

APPENDIX

Figures :

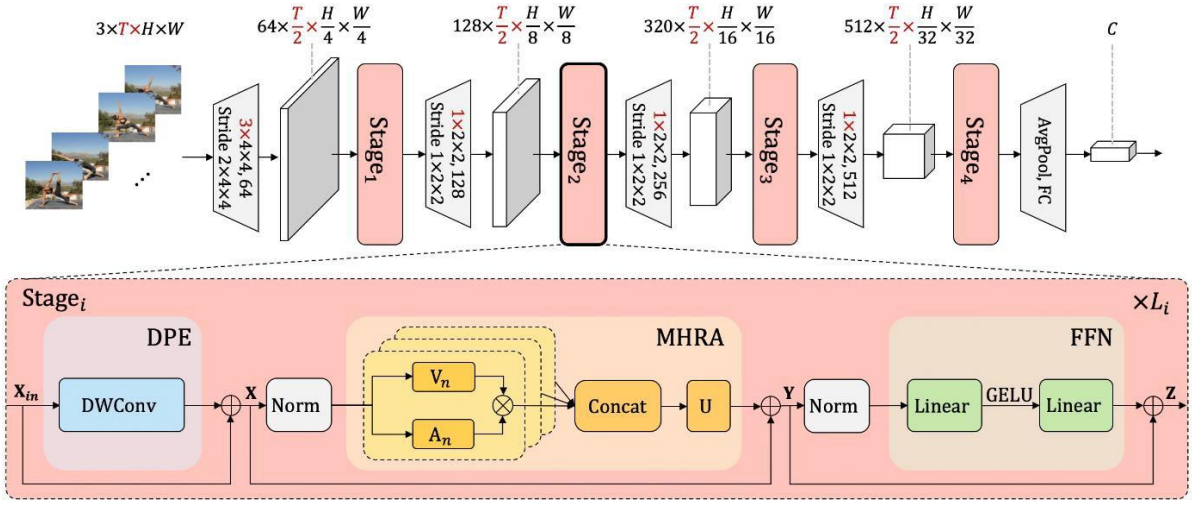


Figure 1 : Unified transFormer architecture [1]

Tables

Table 1: Performance Evaluation of Uniformer vs. CNN/RNN Models

Models	*F1-score*	*Specificity*	*Sensitivity*
CNN/RNN	0.61	0.36	0.62
Uniformer	**0.75**	**0.71**	**0.76**

A Randomized Controlled Trial of Artificial Intelligence to Assist in Screening Fetal Ultrasound Scans

Authors:

Thomas G Day[1,2], Jacqueline Matthew[1], Samuel F Budd[1], Alfonso Farruggia[1], Lorenzo Venturini[1], Robert Wright[1], Bernhard Kainz[3,4], John M Simpson[1,2], Jo V Hajnal[1], and Reza Razavi[1,2]

Affiliations:

1. School of Biomedical Engineering and Imaging Sciences, King's College London, UK
2. Department of Congenital Heart Disease, Evelina London Children's Healthcare, Guy's and St Thomas' NHS Foundation Trust, London, UK
3. Department of Computing, Imperial College London, London, London, UK
4. Image Data Exploration and Analysis Lab, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Erlangen, Germany

Presenting author:

Dr Thomas Day
School of Biomedical Engineering and Imaging Sciences, King's College London, UK
Department of Congenital Heart Disease, Evelina London Children's Healthcare, Guy's and St Thomas' NHS Foundation Trust, London, UK
Thomas.day@kcl.ac.uk

+447944326254

Keywords:

Artificial Intelligence, fetal medicine, ultrasound, congenital heart disease, screening, obstetrics

Key information:

1. Research question: *how does AI assistance change the performance of the screening fetal anomaly ultrasound scan?*
2. Findings: *AI assistance did not alter sensitivity or specificity in detecting fetal malformation, but resulted in significant time savings and reduced sonographer cognitive load. Automatically measured biometrics were highly repeatable and reproducible.*
3. Meaning: *AI assistance may be of value in improving the efficiency of fetal ultrasound scans, without a reduction in diagnostic performance.*

MANUSCRIPT

Introduction

Congenital malformations are the most common cause of infant mortality in high-income countries, but are commonly missed by current ultrasound-based screening programmes.[1,2] Artificial intelligence (AI) has been proposed as a means of improving the performance of these scans. Most work performed to date has been based on retrospective analysis of ultrasound images, with no prospective randomised controlled trials performed, other than a small pilot study by our group that included only healthy fetuses.[3–5]

Material and methods

An AI tool was developed with models designed to automatically save 13 standard image planes, and perform fetal biometry. The models were trained and tested using a prospectively acquired dataset of 7309 full videos of routine fetal ultrasound scans. The models were packaged into a clinically usable tool that operated in real time during the scan, with feedback to the sonographer via a tablet.

The tool was assessed using a randomized, single center, open label trial in a large teaching hospital, comparing AI-assisted vs standard manual scans. Pregnant participants with fetal congenital heart disease (CHD, selected as an example lesion as it the most common and most commonly missed group of malformations) and with healthy fetuses were recruited and scanned with both methods. Screening sonographers were recruited from regional hospitals and were randomized to scan with the AI tool or in the standard fashion, blinded to fetal CHD status.

The main outcome measures were the diagnostic performance in detecting fetal malformation, scan duration, repeatability and reproducibility of biometrics, and sonographer cognitive load.

Results

78 pregnant participants (26 with fetal CHD, 1 with an incidental finding of an extracardiac malformation) and 58 sonographers were recruited. The sensitivity and specificity of the AI-assisted scan in detecting fetal malformation was 88.9% and 98.0% respectively, with the standard scan achieving 81.5% and 92.2% (not significant, table 1). AI-assisted scans were significantly shorter than standard scans (median 11.4 min vs 19.7 min, $p < 0.001$, figure 1), with no change in reporting time. Sonographer cognitive load was significantly lower in the AI-assisted group (median NASA TLX score 35.2 vs 46.5, $p < 0.001$, Paas scale 5 vs 6, $p = 0.004$, figure 2). For all biometrics, the AI repeatability and reproducibility was superior to manual.

Discussion and Conclusion

This is the first randomised controlled trial of AI in obstetric ultrasound that incudes fetuses with known disease. The AI tool did not include any disease detection models, but rather altered sonographer workflow by automating mundane and repetitive tasks such as image saving and fetal biometry. By changing the workflow in this way, a significant time saving was generated (as well as a reduction in sonographer cognitive load). We did not see any change in diagnostic performance, which is reassuring as automation bias leading to a drop in performance has been a concern. Automatic biometry appears to be highly repeatable and reproducible.

The main limitation of this trial was that it was not performed on unselected participants. The cohort was enriched with abnormal fetuses, and the sonographers were self-selected so may not be representative of the screening workforce. Further trials are planned to address both points, but as fetal malformation is relatively rare, very large sample sizes of several thousand participants will be required in order to be adequately powered.

Another limitation is that we did not incorporate any disease detection models in the clinical tool. This is an active area of research, and further iterations of the tool are planned to include this. Nevertheless, this is one of the relatively few randomized prospective controlled trials of AI in medicine and raises the exciting prospect of future human-AI collaboration in this field.

References

1. Ely D, Driscoll A. *Infant Mortality in the United States, 2020*. Vol 71. Hyattsville, MD; 2020. doi:https://dx.doi.org/10.15620/cdc:120700.

2. *National Congenital Heart Disease Audit, Summary Report*. The National Institute for Cardiovascular Outcomes Research; 2021.

3. Matthew J, Skelton E, Day TG, et al. Exploring a new paradigm for the fetal anomaly ultrasound scan: Artificial intelligence in real time. *Prenat Diagn*. 2022;42(1):49-59. doi:10.1002/pd.6059

4. Baumgartner CF, Kamnitsas K, Matthew J, et al. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Trans Med Imaging*. 2017;36(11):2204-2215. doi:10.1109/TMI.2017.2712367

5. Baumgartner CF, Kamnitsas K, Matthew J, Smith S, Kainz B, Rueckert D. Real-Time Standard Scan Plane Detection and Localisation in Fetal Ultrasound Using Fully Convolutional Neural Networks. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, eds. *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2016*. Cham: Springer International Publishing; 2016:203-211.

Disclosures

TD, JM, SB, AF, LV, RW, BK, JH, and RR are founders and shareholders of Fraiya Ltd, a university spinout company that is commercializing some of this work.

APPENDIX

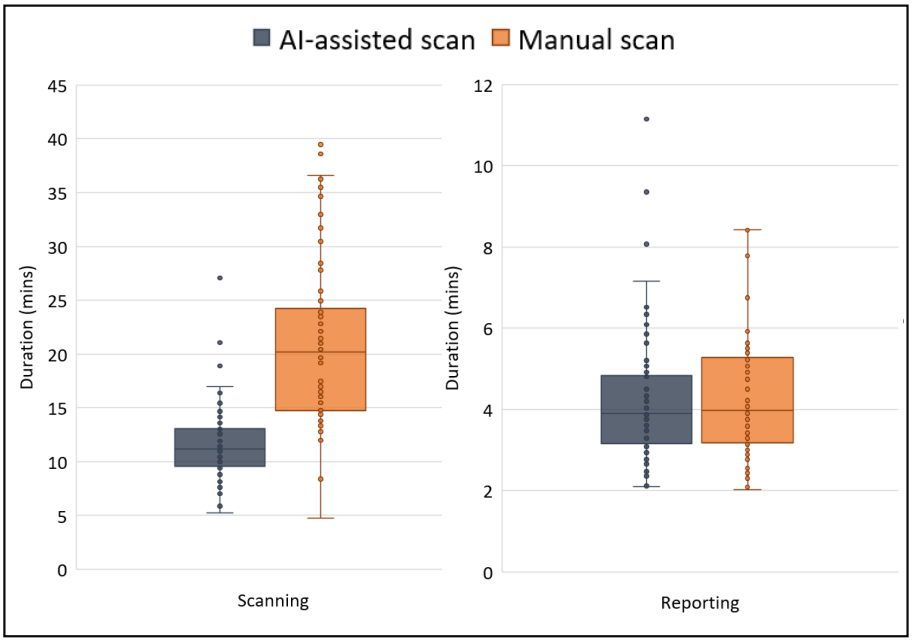


Figure 1: *duration of scanning and reporting by both methods.*

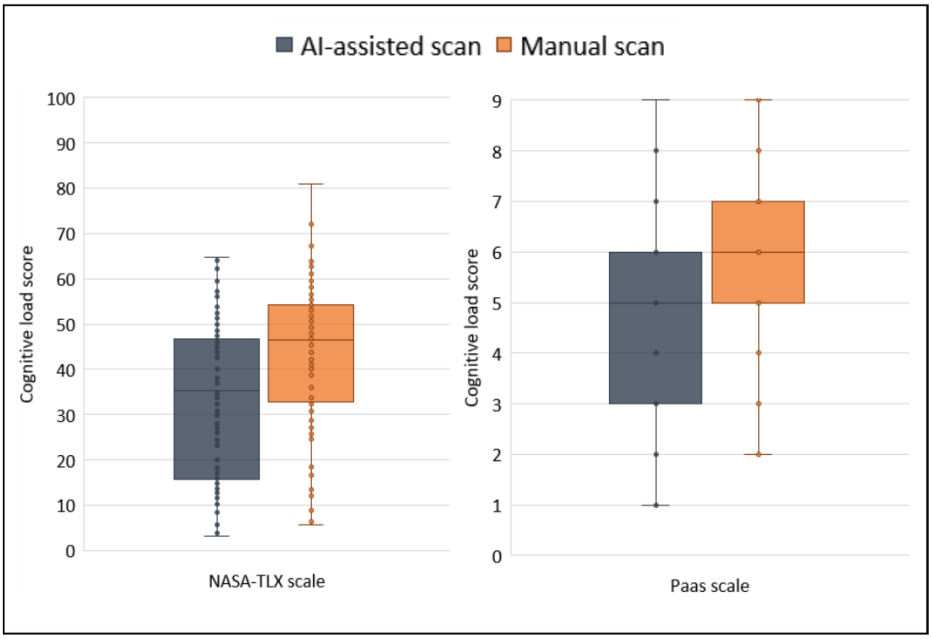


Figure 2: *cognitive load of the sonographers compared between the two scanning methods.*

Table 1: *diagnostic performance of the two methods in detecting all fetal structural malformations (affected group n=27, unaffected group n=51).*

	AI-assisted scan	**Manual scan**	**P value***
True positive (n)	24	22	-
False positive (n)	1	4	-
True negative (n)	50	47	-
False negative (n)	3	5	-
Sensitivity (95% CI)	88.9% (70.8-97.6%)	81.5% (61.9-93.7%)	0.480
Specificity (95% CI)	98.0% (89.6-100%)	92.2% (81.1-97.8%)	0.180

*McNemar's test for paired proportions

EyeLiner: Longitudinal retinal fundus image registration through clinically guided keypoint detection

Authors:

Stephen M. McNamara, OD[1], Yoga Advaith Veturi, MSc*[1], Scott Kinder, MS[1], Christopher W. Clark, MS[1], Upasana Thakuria, MS[1], Benjamin Bearce, MS[1], Malik Y. Kahook, MD[1], Naresh Mandava, MD[1], Niranjan Manoharan, MD[1], Praveer Singh, PhD[1], Jayashree Kalpathy-Cramer, PhD[1]*

** Co-first Authors*

Affiliations:

[1]Ophthalmology, University of Colorado School of Medicine, Aurora, CO

Presenting author:

Stephen M. McNamara, Ophthalmology, University of Colorado School of Medicine

Email: steve.mcnamara@cuanschutz.edu, Phone: +1(815)2285509

Keywords:

Age-related macular degeneration; Image registration; Deep Learning; Change detection

Key information:

1. Research question: Can artificial intelligence be used to align fundus imaging acquired at different timepoints from the same patient?
2. Findings: We evaluated an open-source pipeline "EyeLiner", for registering, or aligning, longitudinal fundus imaging on our internal dataset. EyeLiner shows improvements in performance over the current state-of-the-art techniques for fundus image registration.
3. Meaning: Our pipeline effectively compensates for image variation induced during image capture, while preserving the biological variations associated with the target disease. This allows clinicians to better visualize pathological changes across different timepoints.

MANUSCRIPT

Introduction

Detecting changes in longitudinal fundus imaging is key to monitoring disease progression in chronic ophthalmic diseases, such as geographic atrophy (GA) due to age-related macular degeneration (AMD)[1-5]. With the recent FDA approval of therapies to slow GA progression, detecting lesion growth becomes of even higher importance. Yet, clinicians still typically assess changes in disease status by either independently reviewing or manually juxtaposing longitudinally acquired fundus imaging. This makes manual image evaluation variable and subjective, potentially impacting clinical decision-making. Our pipeline, EyeLiner, obtained state-of-the-art results on the Fundus Image REgistration (FIRE) dataset[6]. In

this study, we evaluate our deep learning (DL) pipeline, "EyeLiner," for registering, or aligning, two-dimensional fundus imaging in the context of GA.

Material and methods

EyeLiner registers a "moving" image to a "fixed" image using an anatomically guided keypoint matching pipeline (Figure 1). Our pipeline uses out-of-the-box DL algorithms to segment the blood vessels from the fixed and moving images. Subsequently, the vessel segmentations are provided as input to the SuperPoint and LightGlue algorithms to output corresponding keypoint sets for the fixed and moving images[7,8]. Finally, the spatial transformation between keypoint sets is computed using a thin-plate spline algorithm, which is used to resample the final aligned image. We evaluate EyeLiner on 50 color fundus photo (CFP) pairs and 176 fundus autofluorescence (FAF) pairs from an internal dataset of age-related macular degeneration patients from the Colorado Ophthalmology Research Information System (CORIS)[9]. We qualitatively evaluated EyeLiner using flicker animations, checkerboards and subtraction maps where the fixed and registered images were alternated to determine continuity in the blood vessels. Quantitatively, we compute the mean distance (MD) between clinician-annotated keypoints on the fixed and the registered moving images for CFPs. Furthermore, defining a successful registration that is below a distance threshold, we plot the success rate over varying thresholds. This gives a monotonic curve from which we compute the area under curve (AUC) statistic to summarize model performance.

Results

We found EyeLiner effectively aligns longitudinal image pairs from FIRE and our CORIS dataset, as qualitatively evaluated through registration checkerboards (Figures 2 and 3). We observe that the continuity in static blood vessels was preserved, indicating a good-quality registration. Furthermore, subtraction maps were generated for FAFs, with hypointense regions indicating areas of change between fixed and moving images. Quantitatively, we found that the MD decreased for this model after alignment from 33.77 to 1.34 pixels for the CFPs from the CORIS dataset. Finally, we obtained an AUC of 0.97, which beat the existing state-of-the-art model SuperRetina (AUC=0.82).

Discussion and Conclusion

Our pipeline demonstrates excellent alignment of CFPs in comparison to the previous state-of-the-art method in patients with AMD. EyeLiner finds better correspondences due to the deep learning based keypoint matching pipeline, achieving superior MD and AUC values compared to the SuperRetina pipeline, which relies on a brute-force matching strategy. Furthermore, we qualitatively show clinical applicability to GA progression tracking by using subtraction maps on fundus autofluorescence (FAF) imaging (Figure 3). This information will better assist clinicians in determining whether GA is stable or might benefit from treatment for slowing GA progression. However, our study is not without limitations. While retinal blood vessels are considered anatomically stable, further research is required on the impact of disease-specific lesions, and their progression, on altering vessel morphology and appearance in imaging over time. Additionally, our pipeline uses vessel segmentation to obtain clinically relevant keypoints. We hope to further develop end-to-end vessel keypoint detectors to bypass the need for vessel segmentation. Nevertheless, we envision that EyeLiner will facilitate better visualization of disease change over time, enabling clinicians to make more informed decisions.

References

1. Heijl A, Bengtsson B. Diagnosis of early glaucoma with flicker comparisons of serial disc photographs. *Investigative Ophthalmology & Visual Science*. 1989;30(11):2376-2384.
2. O'Toole L, Coleman K, Franco O Penya H, Coleman J, Keegan DJ. Exploring potential for automatic change alert for diagnosis of optic nerve change head changes in a diabetic population. *Investigative Ophthalmology & Visual Science*. 2023;64(8):232-232.
3. Ramsey DJ, Sunness JS, Malviya P, Applegate C, Hager GD, Handa JT. Automated Image Alignment and Segmentation to follow Progression of Geographic Atrophy in Age-related Macular Degeneration. *RETINA*. 2014;34(7):1296-1307. doi:10.1097/iae.0000000000000069
4. Hussain MA, Govindaiah A, Souied E, Smith RT, Bhuiyan A. Automated tracking and change detection for Age-related Macular Degeneration Progression using retinal fundus imaging. IEEE; 2018:394-398.
5. Sabaté-Cequier A, Dumskyj M, Usher D, et al. Accurate registration of paired macular and nasal digital retinal images : a potential aid to diabetic retinopathy screening.
6. Veturi A MS, Kinder S, Clark C, Bearce B, Mandava N, Kahook MY, Singh P, Kalpathy-Cramer, J. EyeLiner: Longitudinal retinal fundus image alignment through clinically guided keypoint detection. presented at: Association for Research in Vision and Ophthalmology Annual Meeting; May 6th, 2024 2024; Seattle, WA.
7. Lindenberger P, Sarlin P, Pollefeys M. LightGlue: Local Feature Matching at Light Speed. arXiv 2023. *arXiv preprint arXiv:230613643*.
8. DeTone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description. 2018:224-236.
9. Hernandez-Matas C, Zabulis X, Triantafyllou A, Anyfanti P, Douma S, Argyros AA. FIRE: fundus image registration dataset. *Modeling and Artificial Intelligence in Ophthalmology*. 2017;1(4):16-28.

Disclosures

We acknowledge potential conflicts of interest among the authors. Yoga Advaith Veturi, Scott Kinder, Christopher Clark, Benjamin Bearce, Upasana Thakuria, and Praveer Singh declare no conflicts of interest. Stephen McNamara reports a relationship with Evolution Optiks (consulting). Naresh Mandava reports relationships with Soma Logic (consulting, research), ONL Therapeutics (consulting), Alcon (patent), 2C Tech (patent, equity, co-founder), and Aurea Medical (equity, co-founder). Malik Kahook discloses relationships with New World Medical (Patent Royalties, Consultant), Alcon (Patent Royalties), and SpyGlass Pharma (Owner, Consultant, Patent Royalties). Niranjan Manoharan declares relationships with Iveric Bio (Research grant support) and Genentech (Research grant support and Consultant). Jayashree Kalpathy-Cramer reports relationships with Genentech (Financial support to institution), GE Healthcare (Financial support to institution), and Siloam Vision (consultant).

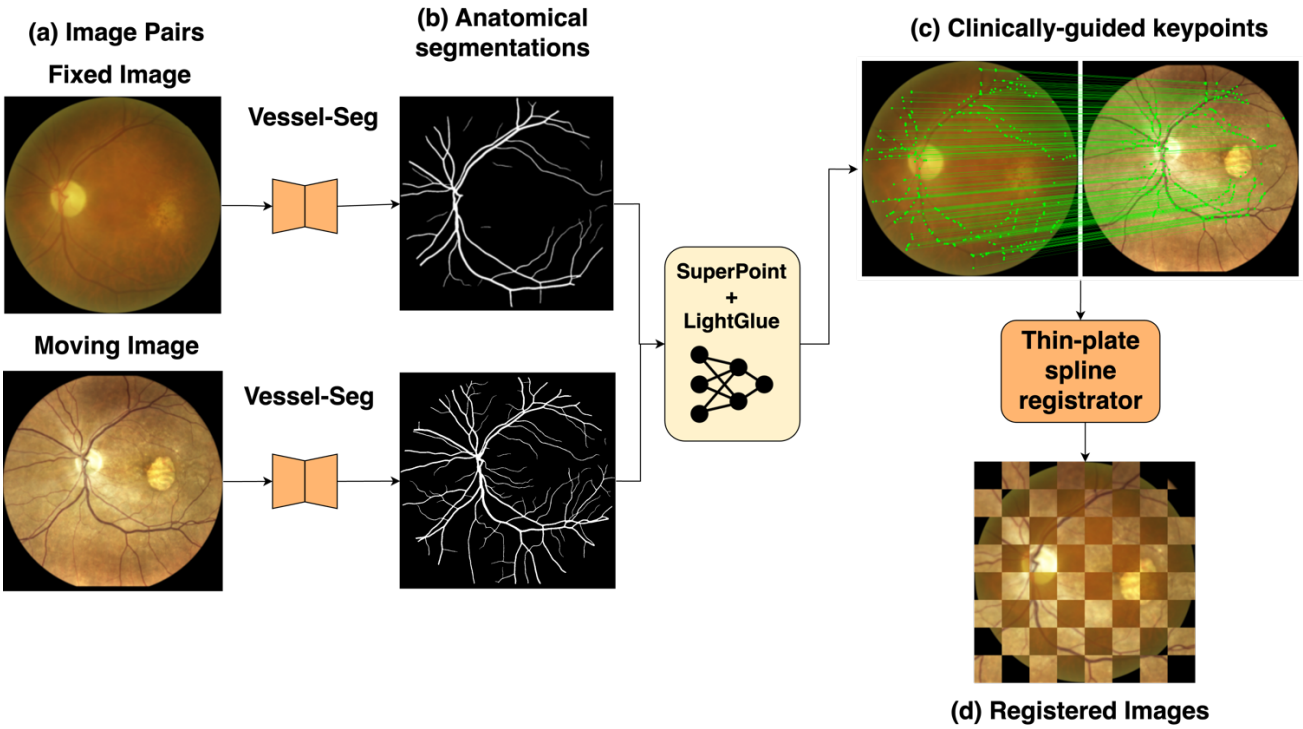


Figure 1: EyeLiner pipeline for registration of color fundus images

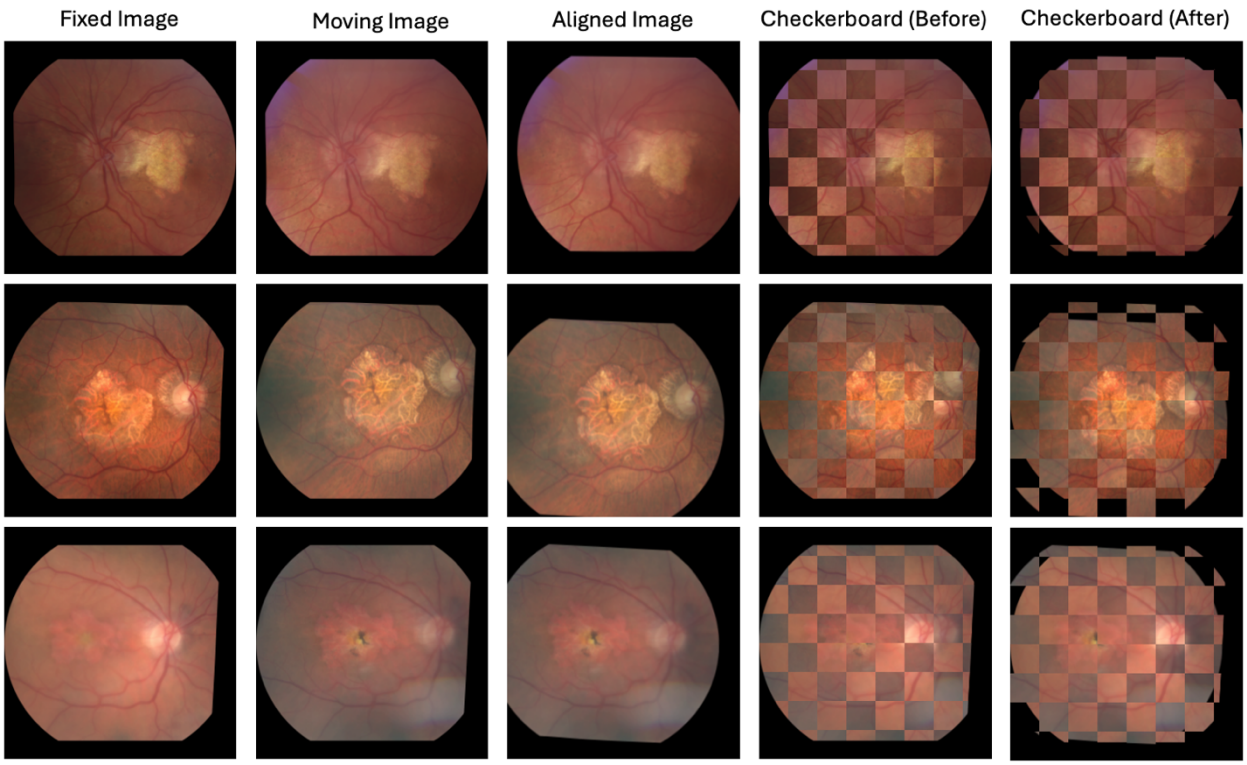


Figure 2: Qualitative evaluation of EyeLiner pipeline for three cases from our internal CORIS dataset. Checkerboards present continuities in the blood vessels after image alignment (right-most column).

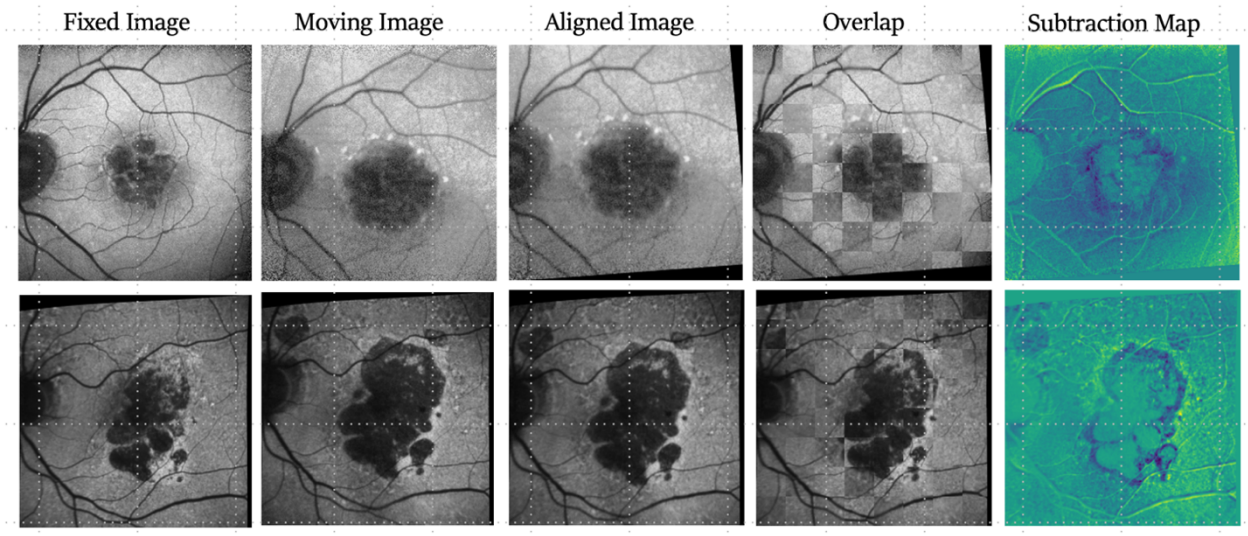


Figure 3: Utility of EyeLiner for change detection. Subtraction map between fixed and registered images shows region of hypointensity around GA lesion, demonstrating GA growth.

Automated deep learning-based quantification of intermediate age-related macular degeneration features for the prediction of geographic atrophy and neovascularization

Authors:

Haras Mhmud, MD; Eric Thee, MD; Bart Liefers, PhD; Daniel Luttikhuizen, PhD; Caroline Klaver, MD MSc PhD

Affiliations:

[1] Department of Ophthalmology, Erasmus University Medical Center, Rotterdam, the Netherlands
[2] Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands
[3] Department of Ophthalmology, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, the Netherlands
[4] Institute of Molecular and Clinical Ophthalmology, University of Basel, Switzerland

Presenting author:

1. Haras Mhmud, MD.
2. Radboud University Medical Center, Nijmegen, the Netherlands
3. Haras.Mhmud@radboudumc.nl
4. 0031 6 87 12 70 19

Keywords:

Deep learning, ophthalmology, image analysis, clinical prediction modelling

Key information:

1. **Research question**: *Can deep learning-based automated analysis of color fundus images accurately predict of conversion to late AMD?*
2. **Findings**: *Deep learning-based automated analysis of color fundus images is a rapid and accurate method for prediction conversion to late AMD.*
3. **Meaning**: *Automated analysis of color fundus images for predicting late-stage age-related macular degeneration can be used in screening, epidemiological studies, clinical management and patient selection for clinical trials.*

Word count of manuscript: 450

MANUSCRIPT *(included in the word count)*

Introduction

Late-stage age-related macular degeneration (AMD) is the most common cause of blindness in high income countries, caused by either geographic atrophy or neovascularization. The prevalence of late AMD is expected to increase from 400,000 to 700,000 in Europe by the year 2050[1]. Intermediate AMD is a precursor stage with drusen and pigmentary changes visible on color fundus images (CFIs) that can be used for prognostic modelling. Deep learning (DL)-based quantification of these biomarkers to enable stratification patients into high-risk groups who are likely to convert to the blinding late-stage disease[2]. We developed a fully automated DL-based pipeline to rapidly and objectively stratify high-risk individuals.

Material and methods

A total of 37,420 CFIs of 17,605 eyes from 8890 participants partaking in the population-based Rotterdam Study were available. We developed a fully automated retinal image analysis pipeline that encompasses essential steps such as image selection, quality assessment, image preprocessing, automated segmentation of anatomic structures and lesions and the synthesis of quantified biomarkers.

Our DL segmentation algorithm was trained to automatically segment drusen, reticular pseudodrusen, hypo- and hyperpigmentation on CFIs and quantify the total affected macular retinal area in mm^2. The CFIs originate from a multitude of different imaging devices used over the last 35 years of the Rotterdam Study. Conversion to geographic atrophy and/or neovascular AMD within 7 years was predicted using random forest classifier based on the automatically extracted features as well as age and sex using 10-fold cross-validation. A separate model was fitted to human-graded labels of the same eyes for comparisons. Area under the receiver operating curves (AUCs) were constructed for geographic atrophy and neovascular AMD.

Results

173 eyes converted to GA and 172 to neovascular AMD within a time period of 7 years. The random forest classifier based on automated intermediate AMD area quantifications reached a similar to higher AUC for conversion to GA (0.95 (95%CI 0.93-0.97) versus 0.95 [95%CI 0.93-0.97]) and for conversion to neovascular AMD 0.93 (95%CI 0.91-0.94) versus 0.92 [95%CI 0.90-0.94]) when compared to human graded labels.

Discussion and Conclusion

Our automated retinal image analysis pipeline for prediction of conversion to late AMD can be used as a rapid and objective method for clinical prediction modelling in AMD patients at risk of GA or neovascularization. Potential use includes population screening, epidemiological studies, clinical management and patient selection for clinical trials. Implementation in the clinical setting in particular would alleviate the burden on clinical ophthalmologists.

A limitation of our tool is that, while being highly accurate, it is nevertheless based on biomarkers on CFIs only. In the future, we envisage incorporating multi-modal data such as from OCT and genetic information to more accurate predictions.

References

1. Li JQ, Welchowski T, Schmid M, Mauschitz MM, Holz FG, Finger RP. Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. Br J Ophthalmol. 2020;104(8):1077-1084. doi:10.1136/bjophthalmol-2019-314422

2. Liefers B, Colijn JM, González-Gonzalo C, et al. A Deep Learning Model for Segmentation of Geographic Atrophy to Study Its Long-Term Natural History. Ophthalmology. 2020;127(8):1086-1096. doi:10.1016/j.ophtha.2020.02.009

Disclosures

none

APPENDIX

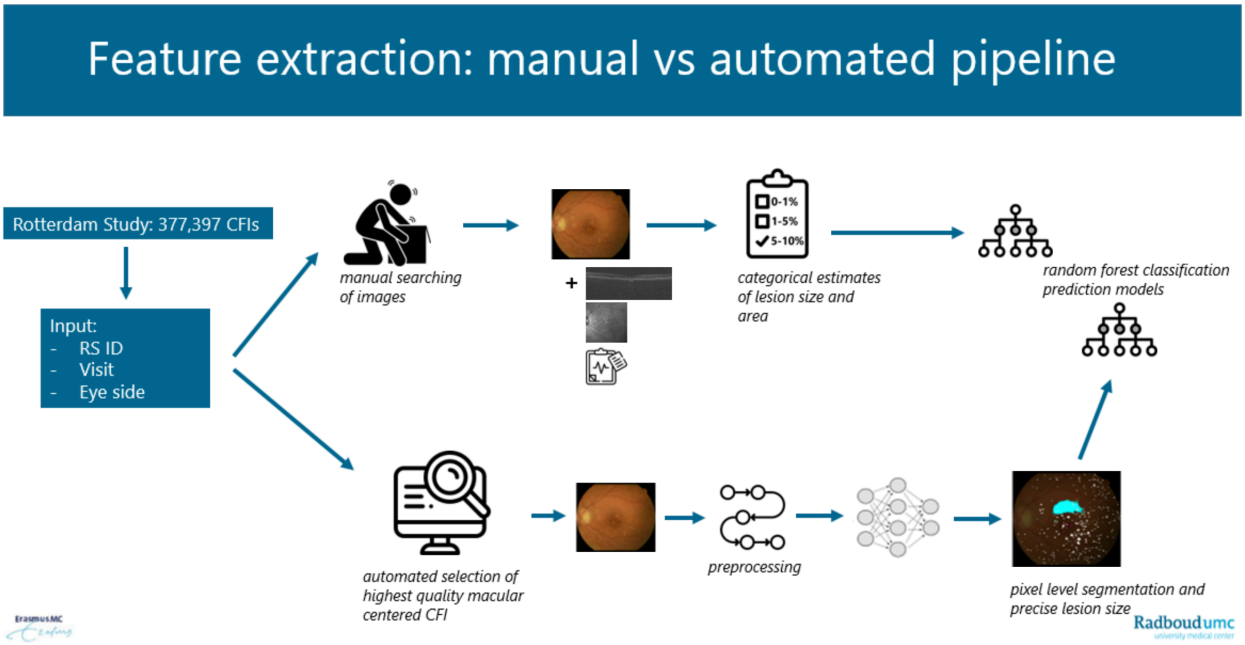


Figure 1. An infographic demonstrating (above) the human-grader process of image analysis with manual searching for images, interpretation of multimodal imaging and clinical data and then coarse quantification of lesion size in categorical estimates and (below) our fully automated pipeline encompassing image selection, quality assessment, image preprocessing, automated segmentation of anatomic structures and lesions and the synthesis of quantified biomarkers for use in a random forest classifier

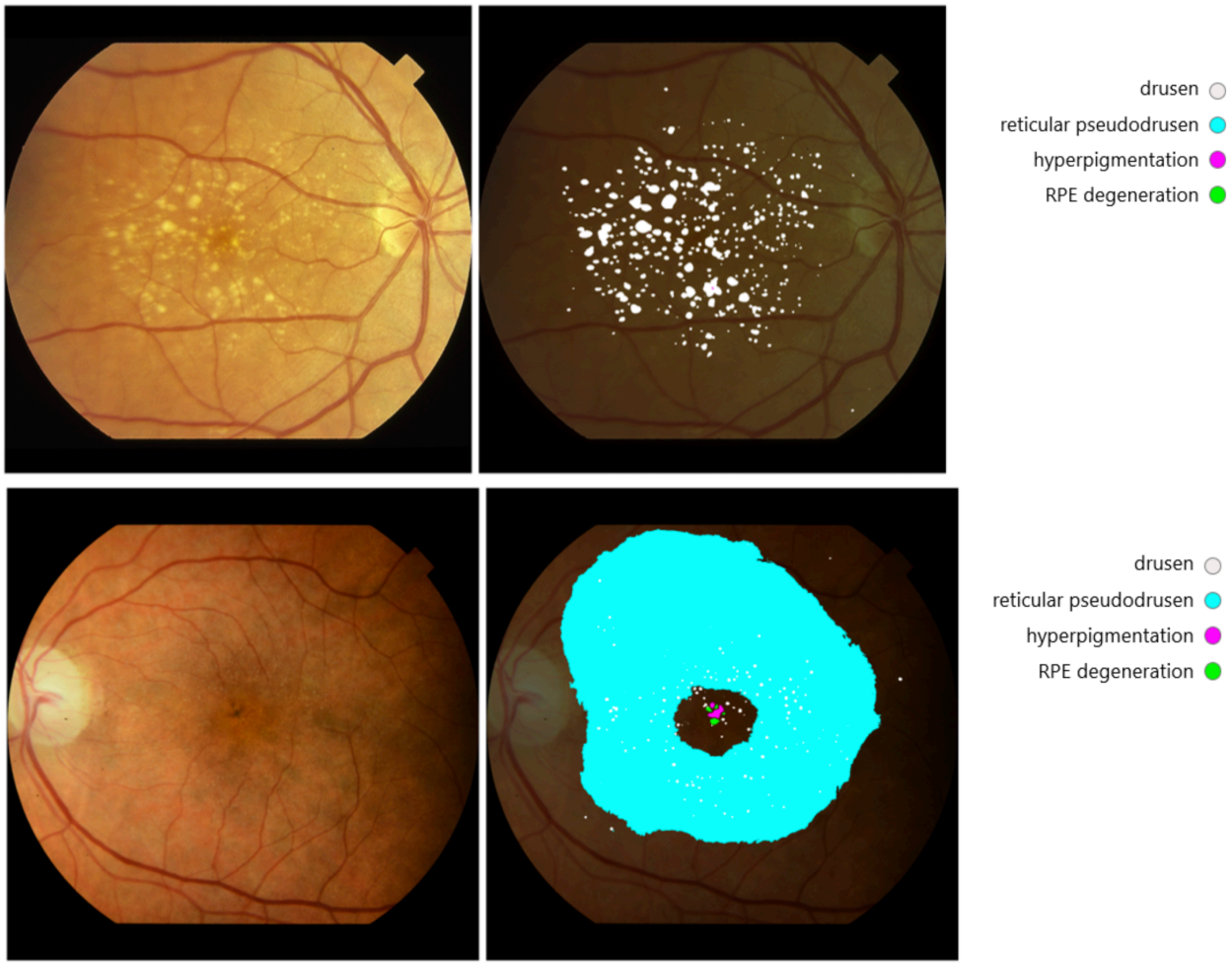


Figure 2a and b: (above) segmentation model output demonstrating automated segmentations of drusen and (below) demonstrating automated segmentations of drusen, reticular pseudodrusen, hyperpigmentation and RPE degeneration