CLINICCAI2023

3RD CONFERENCE ON CLINICAL TRANSLATION OF MEDICAL
IMAGE COMPUTING & COMPUTER ASSISTED INTERVENTION
8-12 October 2023 • Vancouver Convention Centre
Vancouver / Canada

Surgery and Endoscopy

Moderators: Lideke Van Der Steeg, MD and Filippo Filicori, MD

Time	Presenting Author	Title
08:15-08:25	Daichi Kitaguchi (National Cancer Center Hospital East)*	A Comparison of Recognition Performance for Key Anatomical Structures between Artificial Intelligence and Surgeons in Laparoscopic Colorectal Surgery: A Prospective Observational Study
8:25-8:35	Pieter De Backer (Orsi Academy)*	Improving Augmented Reality Surgery through first-in-human real time AI-powered instrument segmentation
8:35-8:45	Rafal D Kocielnik (California Institute of Technology)	Deep Multimodal Fusion for Classification of Surgical Feedback Components in Robot-Assisted Surgery
8:45-8:55	Nobuyoshi Takeshita (National Cancer Center Hospital East)*	Validation Study of an AI Support System for Intraoperative Recognition of Anatomical Structures in Laparoscopic/Robot-assisted Hysterectomy
8:55-9:05	Martijn Jong (Amsterdam UMC)*(Amsterdam UMC); Jacques Bergman (Amsterdam UMC)	Computer aided detection system for Barrett's neoplasia improves endoscopic detection by general endoscopists: an ex-vivo benchmarking study.
9:05-9:15	Surgery and Endoscopy Discussion (slush time)	

A Comparison of Recognition Performance for Key Anatomical Structures between Artificial Intelligence and Surgeons in Laparoscopic Colorectal Surgery: A Prospective Observational Study

Authors:

Daichi Kitaguchi[1,2], Norihito Kosugi[1], Kazuyuki Hayashi[1], Yuto Ishikawa[1], Hiro Hasegawa[1,2], Nobuyoshi Takeshita[1], Masaaki Ito[1,2]

Affiliations:

1. Department for the Promotion of Medical Device Innovation, National Cancer Center Hospital East, Chiba, Japan.
2. Department of Colorectal Surgery, National Cancer Center Hospital East, Chiba, Japan.

Presenting author:

Daichi Kitaguchi, MD, PhD

Department for the Promotion of Medical Device Innovation, National Cancer Center Hospital East

6-5-1, Kashiwanoha, Kashiwa, Chiba 277-8577, Japan

E-mail: dkitaguc@east.ncc.go.jp

Keywords:

Computer vision, Deep learning, Semantic segmentation, Intraoperative image navigation, Laparoscopic colorectal surgery

Key information:

1. **Research question:** Can automatic recognition models be used real-time in a laparoscopic colorectal surgical setting and how do they compare to surgeons?
2. **Findings:** Real-time automatic recognition models for the ureters and autonomic nerves in laparoscopic colorectal surgery were successfully developed using retrospective surgical videos. A prospective observational study determined that the models could operate in a real surgical environment and recognized key anatomical structures faster than surgeons during surgery; this result was significantly more profound in the group of inexperienced than experienced surgeons.
3. **Meaning:** The models may be able to improve surgical safety by compensating for the experience of surgeons.

MANUSCRIPT *(included in the word count)*

Introduction

In order to prevent intraoperative injuries, surgeons should always recognize key anatomical structures quickly and accurately during surgery. Since medical adverse events are largely surgical, there is a need for developing technological innovations to improve surgical safety and efficiency.

The ureter and autonomic nerves are key anatomical structures in laparoscopic colorectal surgery, and their inadvertent injury can lead to severe complications and urogenital dysfunction. To avoid such injuries, surgeons aim to recognize these anatomical structures as early and accurately as possible during surgery.

The aims of this study are to develop a real-time automatic recognition model for the ureter and autonomic nerves in laparoscopic colorectal surgery and to prospectively compare the recognition performance in a real surgical environment with surgeons.

Material and methods

This is a single-center prospective observational study.

Surgical videos recorded between January 2015 and December 2021 were used retrospectively as training and test data to develop the model. A prospective observational study conducted between September 2021 and February 2022 evaluated the model in the surgical setting.

Surgical videos of 299 and 249 cases to develop the models recognizing ureters and autonomic nerves, respectively, and 20 patients who underwent laparoscopic sigmoid colon or rectal resection during the study period to prospectively evaluate the model.

Recognition performance of the deep learning-based semantic segmentation model was compared with that of board-certified and uncertified surgeons.

We evaluated whether the developed model could automatically and accurately recognize the ureter and autonomic nerves in real-time in the operation room and compared time to recognition between surgeons and the model.

Results

In the ureter semantic segmentation task, for training, 10,711 annotated images from 252 surgical videos and 9,795 unannotated images from 36 surgical videos were used as training data and negative training data, respectively. The negative training data were used to reduce false positives. For testing, 2,266 annotated images from 47 videos and 2,350 unannotated images from nine videos were used.

In the semantic segmentation task for autonomic nerves, for training and testing, 14,577 annotated images from 194 surgical videos and 3,599 annotated images from 55 videos were used, respectively.

The Dice similarity coefficient (DSC), recall, and precision in the ureter semantic segmentation task were 0.722, 0.739, and 0.707, respectively. In addition, in the semantic segmentation task for autonomic nerves, the DSC, recall, and precision for hypogastric nerves were 0.579, 0.566, and 0.592, respectively, and those for the aortic plexuses were 0.628, 0.603, and 0.656, respectively.

In a total of 89 comparisons between all surgeons versus the model, the model could recognize targets faster than surgeons 67 out of 89 times (75%). In a comparison separately for board certification status, the model could recognize faster than board-certified and uncertified surgeons 29 out of 44 times (66%) and 38 out of 45 times (84%), respectively. A significant difference was observed between them (p = 0.043).

Discussion and Conclusion

We successfully developed a real-time automatic recognition model for the ureter and autonomic nerves in laparoscopic colorectal surgery. The developed model operated normally in a real surgical environment and could recognize target structures faster than surgeons during surgery. Notably, this result was enhanced for inexperienced surgeons, suggesting the proposed approach may be able to compensate for the skill and experience of surgeons. This study offers novel insights and may help promote research in the field of computer vision in surgery. Nevertheless, many challenges remain in terms of clinical applicability and generalizability of such systems, which should be addressed through multicenter randomized controlled trials.

Disclosures

All authors declare no financial or non-financial competing interests related to this study.

APPENDIX

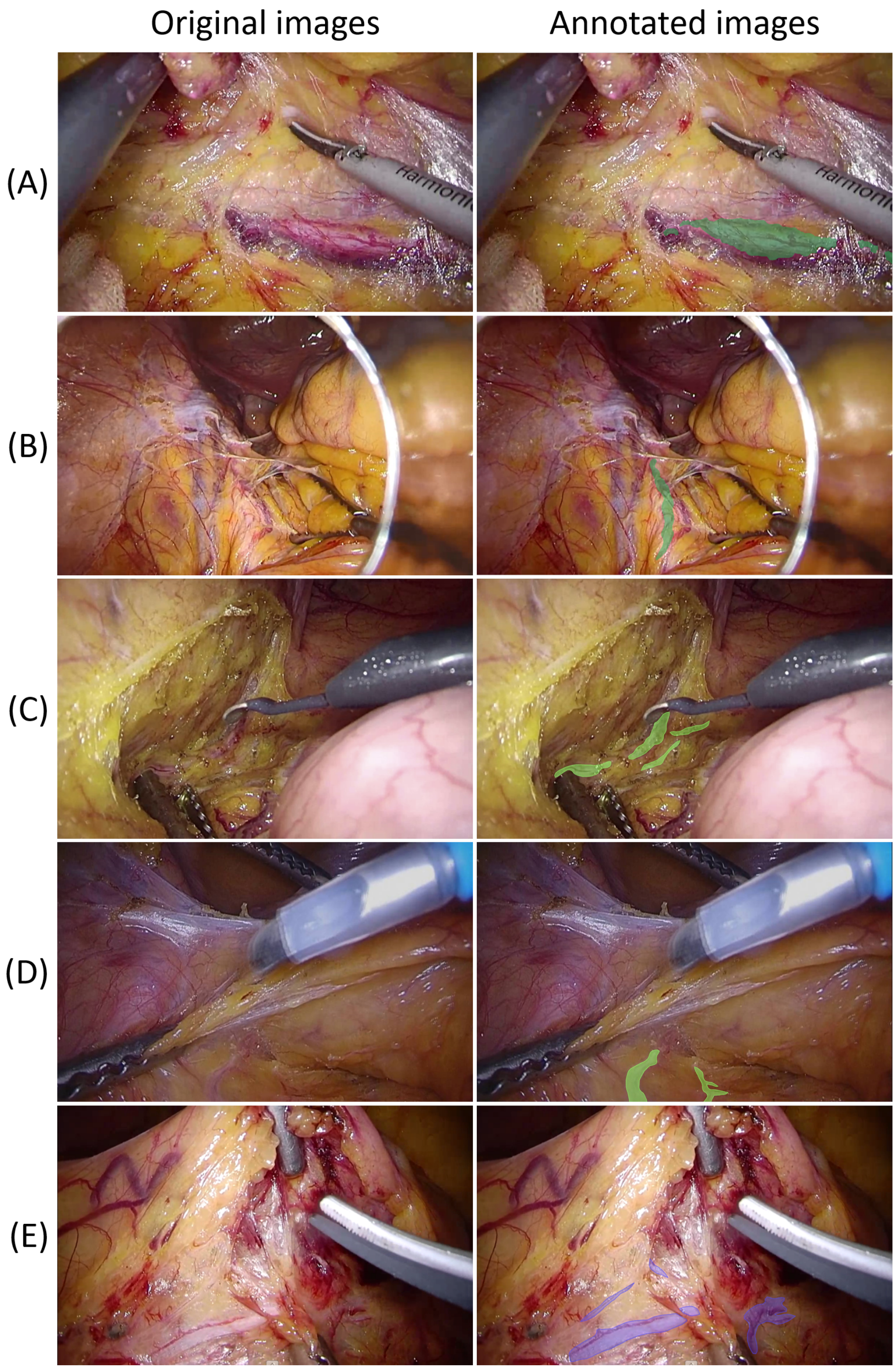


Figure 1: Reference original (left) and corresponding annotated images (right). (A) Image with ureter annotation in a medial-to-lateral mobilization approach. (B) Image with ureter annotation in a lateral-to-medial mobilization approach. (C) Image with right hypogastric nerve annotation. (D) Image with left hypogastric nerve annotation. € Image with aortic plexus annotation.

Table 1: Results of the quantitative evaluation with metrics

	DSC	Recall	Precision
Ureter	0.722	0.739	0.707
Autonomic nerves			
Hypogastric nerves	0.579	0.566	0.592
Aortic plexus	0.628	0.603	0.656

DSC: Dice similarity coefficient

Table 2: Results of the qualitative evaluation using the rubric

Questions	Ureter	Autonomic nerves
1. Would you like to use this system intraoperatively?	3.67 (± 0.985)	3.73 (± 0.594)
2. Do you think this system will help you avoid ureteral or autonomic nerve injury?	3.67 (± 0.888)	4.07 (± 0.704)
3. Do you think this system will help you select an appropriate dissection plane?	3.08 (± 0.900)	3.47 (± 0.743)
4. Do you think this system will facilitate intraoperative guidance or make it easier for the trainee to understand it?	3.75 (± 0.622)	4.20 (± 0.775)
5. Do you think this system is useful for postoperative self-review?	3.50 (± 0.905)	4.33 (± 0.816)
6. Do you think this system could improve surgical safety?	3.75 (± 0.754)	4.00 (± 0.756)

*Mean (± standard deviation)

Table 3: Results of the prospective evaluation of recognition performance

	Success	Failure	NA	Recognition rate (%)
Ureter				
Medial to lateral view	19	1	0	95.0
Lateral to medial view	17	2	1	89.5
Autonomic nerves				
Right hypogastric nerve	17	2	1	89.5
Left hypogastric nerve	15	1	4	93.8
Aortic plexus	18	1	1	94.7

NA: no appearance on the monitor

Table 4: Results of the comparison of the intraoperative recognition performance of surgeons and the developed model

	Surgeon vs AI (N = 89 comparisons)			Qualified surgeon vs AI (N = 44 comparisons)			Nonqualified surgeon vs AI (N = 45 comparisons)		
	Surgeon	AI	Draw	Surgeon	AI	Draw	Surgeon	AI	Draw
Ureter									
Medial to lateral view	5	12	2	4	4	1	1	8	1
Lateral to medial view	1	15	1	1	7	1	0	8	0
Autonomic nerves									
Right hypogastric nerve	3	14	2	1	7	1	2	7	1
Left hypogastric nerve	2	11	2	1	6	1	1	5	1
Aortic plexus	4	15	0	4	5	0	0	10	0
Surgeon win	15 (17%)			11 (25%)			4 (9%)		
AI win	67 (75%)			29 (66%)			38 (84%)		
Draw	7 (8%)			4 (9%)			3 (7%)		

AI: the developed semantic segmentation model

Improving Augmented Reality Surgery through first-in-human real time AI-powered instrument segmentation

Authors:

Pieter De Backer[1,2], Jasper Hofman[1], Ilaria Manghi[3], Jente Simoens[1], Julie Lippens[4], Tim Oosterlinck[5], Charlotte Debbaut[6], Ruben De Groote[7], Hannes Van Den Bossche[8], Charles Van Praet[2], Mathieu D'Hondt[9], Federica Ferraguti[3], Zhijin Li[10], Oliver Kutter[10], Karel Decaestecker[2,11], Alex Mottrie[1,7]

Affiliations:

1) ORSI Academy, Belgium
2) Ghent University Hospital, Urology, Belgium
3) Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Italy
4) Faculty of Medicine and Health Sciences, Ghent University, Belgium
5) Faculty of Medicine, KU Leuven, Belgium
6) IBiTech-Biommeda, Faculty of Engineering and Architecture, and CRIG, Ghent University, Belgium
7) OLV Hospital Aalst-Asse-Ninove, Urology, Belgium
8) AZ West Hospital, Urology, Belgium
9) AZ Groeninge Hospital, HPB surgery, Belgium
10) Nvidia, Santa Clara, California, USA
11) AZ Maria Middelares Hospital, Urology, Ghent, Belgium

Presenting author:

Pieter De Backer, MD
Email: pieter.de.backer@orsi.be

Keywords:

Augmented Reality, Instrument Segmentation, Robotic Surgery, On-Edge AI Deployment, Live Surgery

Key information: 100 words (OK); 1-2 sentences per item

1. Research question: Augmented Reality (AR) in robotic surgery is amongst others hampered by 3D models occluding instruments. Can deep learning instrument segmentation resolve this issue?
2. Findings: We present an instrument segmentation pipeline which segments all surgical instruments. A dedicated GPU infrastructure merges the segmentation video stream with the AR stream and feeds it back to the surgical console in real-time during 3 different robotic procedures.

3. Meaning: Instrument occlusion no longer poses a significant bottleneck for safe clinical implementation of AR in renal surgery. We obtain a 13 milliseconds latency which enables real-time instrument de-occlusion during AR surgery.

MANUSCRIPT

Introduction

The integration of Augmented Reality (AR) into surgical practice faces several challenges, such as accurate registration of pre-operative data and context-awareness to display relevant information[1]. Accurate registration requires, amongst other, handling with real and virtual occlusions caused by the AR model[2]. Figure 1 depicts the use of AR for renal surgery and shows how instrument occlusion inhibits a safe AR experience.

In this work, we present the software development and hardware implementation of a robust real-time binary segmentation pipeline to de-occlude surgical instruments. We show the pipeline works efficiently during 3 robot-assisted surgeries.

Material and methods

A binary segmentation algorithm was trained, tested and validated on a dataset of 31812 images in which all non-organic items were manually delineated[3]. The dataset was sampled uniformly across 100 full length robot-assisted partial nephrectomies. After splitting on a procedural basis, 24087, 4545 and 3180 images were used for training, validation and testing respectively. A Feature Pyramid Network (FPN) architecture[4] with EfficientNetV2 encoder backbone[5] was selected as the most resource efficient combination in a separate optimization study.

The hardware solution consists of a NVIDIA Clara AGX developer kit[1] as embedded computing architecture for demanding video processing applications, with live video capture through an integrated DELTA-12G-elp-key capture card[2]. The software framework was developed in the NVIDIA Holoscan SDK.
Preoperative 3D models are manually fabricated pre-operatively using Mimics (Materialise, Leuven, Belgium), based on CT or MRI imaging.

Figure 2 displays the operating room setup. The capture card takes in the video stream through SDI, and directly offloads it to the GPU of the Clara AGX, which also has the 3D model preloaded. The user can interact with the 3D model through a keyboard and mouse for correct intra-operative alignment. The output stream is sent to an external monitor as well as to the surgical console where it is viewed in the Intuitive TilePro™ window.

We apply the setup in 3 different hospitals during a vascular stent removal, liver metastasectomy and a partial nephrectomy.

[1] NVIDIA CLARA AGX DEVELOPER KIT FOR AI-ENABLED MEDICAL DEVICES. Details: https://resources.nvidia.com/en-us-enabling-smart-hospitals-ai-ep/nvidia-clara-agx-dev?lx=KWlJE5&xs=301547

[2] DELTA-12G-ELP-KEY 11. Details: https://www.deltacast.tv/products/developerproducts/sdi-capture-cards/delta-12g-elp-key-11

Results

The trained segmentation algorithm achieves a 98,37% test set accuracy. Inference time was reduced from 40,5 to 5,1 seconds through conversion from PyTorch to TensorRT, without impacting accuracy.

The pipeline effectively addresses instrument occlusion with a latency of 13ms per frame. Qualitative surgical feedback indicated that the perceived end-to-end latency is acceptable for real-time surgical application. In case of liver surgery, 3D model alignment was more difficult due to floppiness of the liver and the extent of the liver when compared to the renal tumour or vascular stent. Surgeons found the AR overlay to be particularly useful during moments where echography was used as AR can provide an extra sense of depth. Surgical instrument segmentation now allows them to efficiently manipulate and localize with echography, aided by AR. All patients had a normal postoperative course with no adverse events.

Discussion and Conclusion

In this work, we show that AR induced instrument occlusion is a resolvable issue when combining dedicated hardware and software solutions. The segmentation algorithm is shown to transfer smoothly across 3 different domains of robot-assisted surgery in 3 hospitals. Despite being trained only on robot-assisted partial nephrectomy instrument segmentation, the algorithm seems to generalize well across robotic surgery, also on unseen instruments. Correct overlay and registration remain bothersome issues and are found to be domain specific, e.g. it is more bothersome in liver than in vascular or renal AR, due to larger liver deformations. Future work should continue in focussing on improving registration and model deformation.

References

1. Qian L, Wu JY, DiMaio SP, Navab N, Kazanzides P. A review of augmented reality in robotic-assisted surgery. *IEEE Trans Med Robot Bionics*. 2020;2(1):1-16.

2. Suzuki R, Karim A, Xia T, Hedayati H, Marquardt N. Augmented reality and robotics: A survey and taxonomy for AR-enhanced human-robot interaction and robotic interfaces. In: *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM; 2022. doi:10.1145/3491102.3517719

3. De Backer P, Eckhoff JA, Simoens J, et al. Multicentric exploration of tool annotation in robotic surgery: lessons learned when starting a surgical artificial intelligence project. *Surg Endosc*. 2022;36(11):8533-8548.

4. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. *arXiv [csCV]*. December 2016. http://arxiv.org/abs/1612.03144.

5. Tan M, Le QV. EfficientNetV2: Smaller models and faster training. *arXiv [csCV]*. April 2021. http://arxiv.org/abs/2104.00298.

Disclosures

No conflicts of interest to declare.

APPENDIX

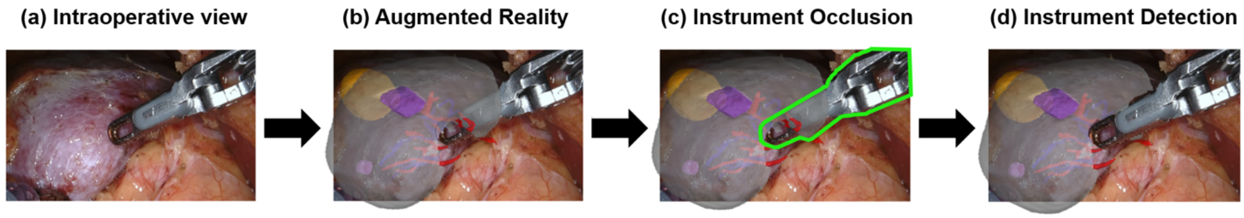


Figure 1: The problem of instrument occlusion by 3D models during AR. Figure C highlights in green the problem of instrument occlusion which inhibits fluent AR interaction. Figure D shows how it can be tackled through instrument segmentation.

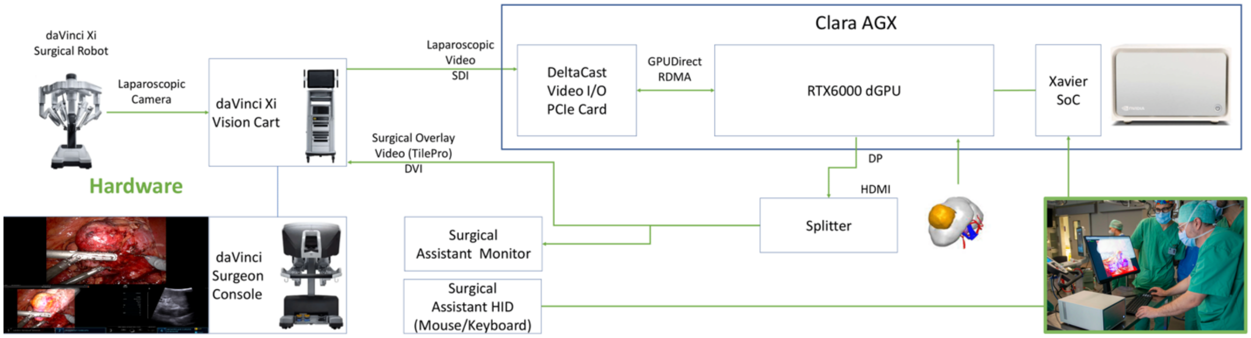

daVinci Xi Surgical Robot
Laparoscopic Camera
daVinci Xi Vision Cart
Laparoscopic Video SDI
Clara AGX
DeltaCast Video I/O PCIe Card
GPUDirect RDMA
RTX6000 dGPU
Xavier SoC
Surgical Overlay Video (TilePro) DVI
DP
HDMI
Splitter
Hardware
daVinci Surgeon Console
Surgical Assistant Monitor
Surgical Assistant HID (Mouse/Keyboard)

Figure 2: Hardware setup enabling real-time AR instrument de-occlusion. In the right bottom corner, we see a second surgeon performing manual 3D model registration on the endoscopic view. The endoscopic view is pulled into the Clara AGX's GPU directly from the endoscopic tower. The operating surgeon sees this identical screen in the left lower corner of robotic console.

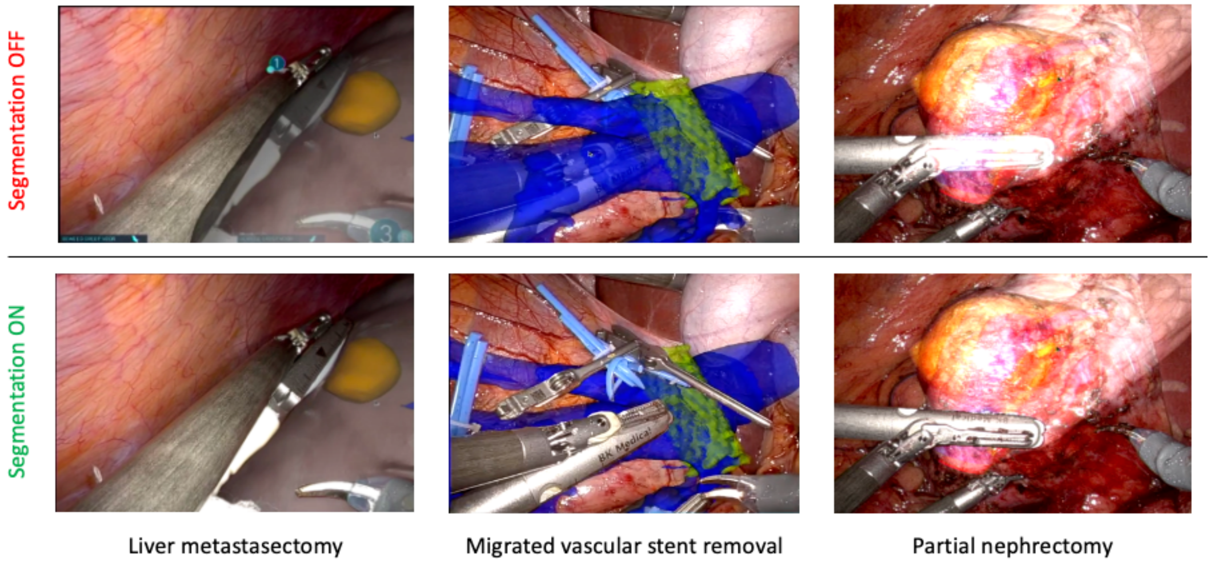


Figure 3: Toggling on/off instrument occlusion for all 3 cases during echography.

Deep Multimodal Fusion for Classification of Surgical Feedback Components in Robot-Assisted Surgery

Authors:

Andrew Hung[1], Rafal Kocielnik[2], Elyssa Wong[1], Timothy Chu[1], De-An Huang[3], Animashree Anandkumar[2]

Affiliations:

[1]University of Southern California, [2]California Institute of Technology, [3]NVIDIA

Presenting author:

Andrew J. Hung, MD

andrew.hung@med.usc.edu

University of Southern California Institute of Urology

1441 Eastlake Avenue Suite 7416

Los Angeles, California 90089

Keywords: surgical feedback quantification, multimodal fusion, robot-assisted

surgery Key information:

1. **Research question**: *Can deep-learning based multimodal fusion approach automatically classify surgical feedback components? Does a combination of video, audio, and text improve classification AUC?*
2. **Findings:** *We achieve high AUCs for classification of feedback categories ranging from 77 to 96 and that fusion improves performance by 6.8%. We learn that Staged training, that is first pretraining each modality separately and then training them jointly, is more effective than training modalities together from the start.*
3. **Meaning:** *This work offers an important first look at the feasibility of automated classification of real-word live surgical feedback based on video, audio, and text modalities. This can lead to improvements in surgical training and outcomes.*

MANUSCRIPT *(words: 599)*

Introduction

Real-time informal verbal feedback delivered by experienced surgeon to trainees during live surgery is a key component of surgeons' training process [1]. Prior work has shown that the quality of such feedback can affect intraoperative performance [3], as well as impact surgical skill acquisition [4] and trainee's autonomy [7]. Quantification and systematic analysis of properties of real-world feedback is challenging and time consuming. We aim to leverage deep multi-modal fusion model to classify surgical feedback components automatically.

Material and methods

We use a dataset of real-life feedback delivered by trainers to trainees during live robot-assisted surgery that has been introduced and rigorously annotated in [6]. Feedback is categorized into 1) Anatomic "Familiarity with anatomic structures and landmarks", 2) Procedural - "Pertains to timing and sequence of surgical steps", and 3) Technical - "Performance of discreet task with appropriate knowledge of exposure, instruments, traction, etc.", as well as into delivery categories: 4) Positive Reinforcement - "A complementary remark" and 5) Visual Aid - "Addition of visual element to direct trainee's attention or focus". The categories are non-exclusive.

We leverage multi-modal inputs composed of video, audio, and text (Fig. 1-A) in order to perform binary multi-label classification of surgical feedback into 5 components (Fig. 1-B). In our experiments we systematically vary 2 dimensions: 1) complexity of the fusion model architecture (1-C) and 2) training strategy (1-D).

We obtain individual baselines for each modality by fine-tuning models for the same number of epochs and reporting AUC on the test set. We use label-balancing for each feedback dimension obtained via random downsampling of majority class. For each experiment we perform an 80%/20% random train/test split. We perform each experiment 3 and report mean AUC as well as standard deviation. Dimension specific label balancing leads to variable dataset sizes, specifically: Anatomic (N =2208), Procedural (N =1634), Technical (N =1378), Positive Reinforcement (N =524), Visual Aid (N =606).

We extract 10 second video and audio around a human annotated feedback timestamp. This includes 5 seconds before (to capture context) and 5 seconds after (to capture delivery) the feedback onset. We downsample the video resolution to 320x250 and extract 16 randomly uniformly sampled frames.

Results

We achieve high AUCs varying from 76.5 to 96.2 which make it feasible to apply our model to replace manual annotation (Table 1). Through ablation studies we find that the model training process is more important for fusion effectiveness (gain of 6.8%) than model architecture (gain of 2.0%). We arrive at an optimal **Staged Fusion** approach which starts with independent training of each modality and continues with training modalities jointly. This approach helps mitigate the dominance of one modality that can suppress extracting information from other modalities. We confirm our intuition that video modality is most important for classification of "Visual Aid" dimension and emotion extracted from audio is very important for "Positive Reinforcement" classification.

Discussion and Conclusion

This work is the first to explore automated classification of components of real-world informal live surgical feedback. We show that it is feasible to classify components of such feedback with high AUCs varying from 76.5 up to 96.2. Secondly, we show that this feedback is indeed inherently multi-modal, and fusion can meaningfully improve AUC by as much as 10%. Third, we show that the multi-modal fusion through staged training is more effective than the fusion model architecture itself.

The quantification of surgical feedback as an important first step towards generating or selecting the optimal feedback automatically [5]. We open opportunities for quantification of surgical feedback at scale from video and audio recordings, which can lead to improvements in surgical training and outcomes.

References

1. Agha RA, Fowler AJ, Sevdalis N. The role of non-technical skills in surgery. Annals of medicine and surgery. 2015 Dec 1;4(4):422-7.
2. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP. Comprehensive surgical coaching enhances surgical skill in the operating room. Annals of surgery. 2015 Aug 1;262(2):205-12.
3. Ma R, Lee RS, Nguyen JH, Cowan A, Haque TF, You J, Roberts SI, Cen S, Jarc A, Gill IS, Hung AJ. Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills—an unblinded pilot randomized controlled trial. The Journal of Urology. 2022 Aug;208(2):414-24.
4. Haglund MM, Cutler AB, Suarez A, Dharmapurikar R, Lad SP, McDaniel KE. The surgical autonomy program: a pilot study of social learning theory applied to competency-based neurosurgical education. Neurosurgery. 2021 Apr;88(4):E345-50.
5. Laca JA, Kocielnik R, Nguyen JH, You J, Tsang R, Wong EY, Shtulman A, Anandkumar A, Hung AJ. Using Real-time Feedback To Improve Surgical Performance on a Robotic Tissue Dissection Task. European Urology Open Science. 2022 Dec 1;46:15-21.
6. Wong, E.Y., Yang, C.H., Dalieh, I.S., Sotelo, D.C., Laca, J.A., Ma, R., Chu, T.N., Kocielnik, R., Goldenberg, M.G., Nabhani, J.A., Cen, S., Hung, A.J.: Deconstructing and quantifying live surgical feedback in the operating room. American Urological Association Annual Conference. 2023 (in print)

Disclosures

Andrew Hung is a consultant for Intuitive Surgical; not related to the presented work. Animashree Anandkumar and De-An Huang are paid employees of Nvidia. Rafal Kocielnik is partially supported by Activision-Blizzard-King sponsored research grant; not related to the presented work

APPENDIX

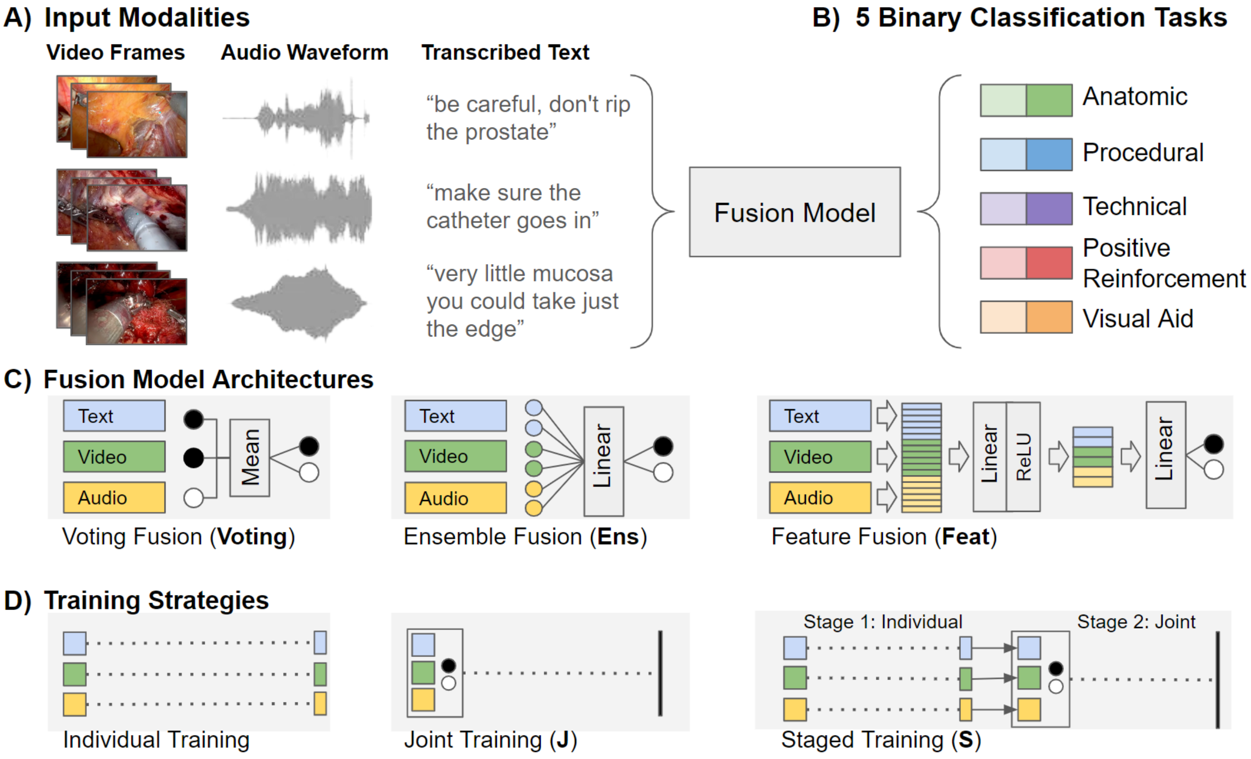
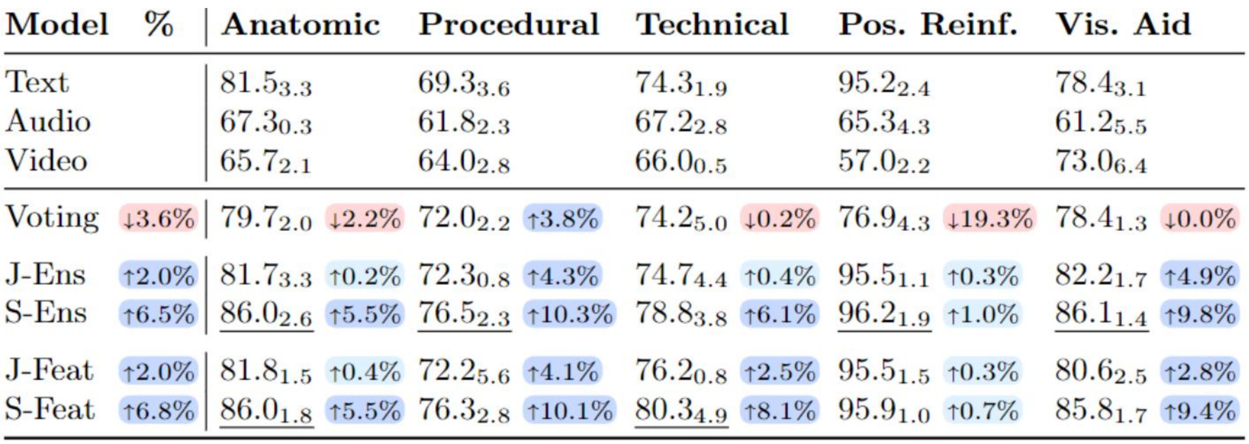


Figure 1: Overview of our multimodal inputs consisting of video, audio, and text (A) and 5 binary multi-label feedback classification outputs (B). We explore model architectures (C) as well as training strategies (D) for increasing the gain from multimodal fusion.



Model	%	Anatomic		Procedural		Technical		Pos. Reinf.		Vis. Aid	
Text		$81.5_{3.3}$		$69.3_{3.6}$		$74.3_{1.9}$		$95.2_{2.4}$		$78.4_{3.1}$	
Audio		$67.3_{0.3}$		$61.8_{2.3}$		$67.2_{2.8}$		$65.3_{4.3}$		$61.2_{5.5}$	
Video		$65.7_{2.1}$		$64.0_{2.8}$		$66.0_{0.5}$		$57.0_{2.2}$		$73.0_{6.4}$	
Voting	↓3.6%	$79.7_{2.0}$	↓2.2%	$72.0_{2.2}$	↑3.8%	$74.2_{5.0}$	↓0.2%	$76.9_{4.3}$	↓19.3%	$78.4_{1.3}$	↓0.0%
J-Ens	↑2.0%	$81.7_{3.3}$	↑0.2%	$72.3_{0.8}$	↑4.3%	$74.7_{4.4}$	↑0.4%	$95.5_{1.1}$	↑0.3%	$82.2_{1.7}$	↑4.9%
S-Ens	↑6.5%	$\underline{86.0_{2.6}}$	↑5.5%	$\underline{76.5_{2.3}}$	↑10.3%	$78.8_{3.8}$	↑6.1%	$\underline{96.2_{1.9}}$	↑1.0%	$\underline{86.1_{1.4}}$	↑9.8%
J-Feat	↑2.0%	$81.8_{1.5}$	↑0.4%	$72.2_{5.6}$	↑4.1%	$76.2_{0.8}$	↑2.5%	$95.5_{1.5}$	↑0.3%	$80.6_{2.5}$	↑2.8%
S-Feat	↑6.8%	$\underline{86.0_{1.8}}$	↑5.5%	$76.3_{2.8}$	↑10.1%	$\underline{80.3_{4.9}}$	↑8.1%	$95.9_{1.0}$	↑0.7%	$85.8_{1.7}$	↑9.4%

Table 1: Mean AUC scores for binary classification of feedback components from 3 runs with different data splits. In the subscript we report the standard deviations. The highest AUC for each component is underscored. ↓ represents loss, ↑ represents small gain withing 1.0% and ↑ represents larger gain. The models are trained on the dataset using human-transcribed text. In the top 3 rows we report performance of individual models on each modality. The Voting represents the simple baseline majority vote fusion. The subsequent Ensemble Fusion (Ens) and Feature Fusion (Feat) models represent

progressively more complex fusion architectures. Each architecture variant is trained either Jointly (J) or in a Staged fashion (S) for 20 epochs. We can see that Staging always improves AUC, while model complexity has practically no impact.

Validation Study of an AI Support System for Intraoperative Recognition of Anatomical Structures in Laparoscopic/Robot-assisted Hysterectomy

Authors:

Nobuyoshi Takeshita[1,4)], Shin Takenaka[1,2)], Yusuke Hirose[3)], Makoto Nakabayashi[3)], Ryo Koike[3)], Yuichi Harada[3)], Koji Matsumoto[3)], Mitsumasa Honma[4)], Hiroki Matsuzaki[4)], Hiroshi Tanabe[2)], Masaaki Ito[1,4)]

Affiliations:

1) Department for the Promotion of Medical Device Innovation, National Cancer Center Hospital East

2) Department of Gynecology, National Cancer Center Hospital East

3) Department of Obstetrics and Gynecology, Showa University Hospital

4) Jmees Inc.

Presenting author:

Nobuyoshi Takeshita

Department for the Promotion of Medical Device Innovation, National Cancer Center Hospital East

Jmees Inc.

ntakeshi@east.ncc.go.jp

Keywords:

Deep learning, Semantic segmentation, Hysterectomy

Key information:

1. Research question:
   Can an AI system be used to support surgeons' recognition of anatomical structures in total laparoscopic/robot-assisted hysterectomy?
2. Findings:
   The ability to recognize the ureter/bladder with and without the use of AI was assessed. As a result, a significant improvement in sensitivity was observed with the use of AI, regardless of whether the physician was a specialist or non-specialist in gynecology.
3. Meaning:
   These results suggest the system can be used to improve surgeons' recognition of anatomical structures during total laparoscopic/robot-assisted hysterectomy and it might lead to prevention of intraoperative organ injuries.

MANUSCRIPT *(included in the word count)*

Introduction

In laparoscopic or robot-assisted total hysterectomy for conditions such as uterine fibroids or uterine cancer, intraoperative injuries of the ureter and bladder are considered the most notable incidental complication (frequency: 0.5-2.2%)[1-3). Injuries of the ureter or bladder can occur due to misidentification or inadequate confirmation during the dissection and suturing around the uterus. There have also been reports that the risk of ureteral injury increases 4.4 times when the operation is performed by surgeons with experience of less than 30 cases[3). We have developed an artificial intelligence (AI) model to recognize the ureter and bladder intraoperatively. We conducted the standalone performance testing and the support performance testing of the AI model to see if it could support surgeons in recognizing the ureter and bladder.

Material and methods

For the construction of the AI model, intraoperative endoscopic images from laparoscopic or robot-assisted total hysterectomy were used. For the AI model of ureter recognition support, annotated training data from 41 institutions, 409 cases, and 13,934 images were used. For the AI model of bladder recognition support, annotated training data from 38 institutions, 220 cases, and 4,940 images were used.

For the standalone performance testing, 300 short videos with and without the ureter and 240 short videos with and without the bladder are prepared and evaluated the accuracy of the AI model in detecting them.

For the support performance testing, 150 short videos with and without the ureter and 120 short videos with and without the bladder are prepared. Factors that could affect the recognition ability of the AI model, such as severe adhesion of Douglas, large myomas, and a history of cesarean section, were included in the short videos as difficult cases. We conducted a recognition test for eight gynecological specialists and eight non-specialists, where they had to determine the presence and position of the ureter and bladder in videos without AI model support, and then conducted same recognition tests using videos overlaid with AI model inference results. We verified whether the accuracy of anatomical recognition by physicians changed with or without the support of the AI model.

Results

In the standalone performance testing, the sensitivity was 51.3% for the ureter and 80.0% for the bladder, with specificity of 92.7% and 94.2% respectively. In the support performance testing, there were significant increases in sensitivity under AI assistance, with the ureter increasing from 43.5% to 58.1%, and the bladder from 54.2% to 70.0%. Especially among non-specialists, significant increases were observed, with the ureter's sensitivity improving from 35.3% to 54.9% and the bladder's sensitivity from 46.6% to 68.4%.

Discussion and Conclusion

From the results of the two performance testing, it appears that our AI model can support physicians in recognizing the ureter and bladder. These results suggest the system can be used to improve surgeons' recognition of anatomical structures during total laparoscopic/robot-assisted hysterectomy and it might lead to prevention of intraoperative organ injuries.

References

1) Aarts JW, et al. Surgical approach to hysterectomy for benign gynaecological disease. Cochrane Database Syst Rev. 2015. PMID: 26264829

2) J Mäkinen et al. Morbidity of 10 110 hysterectomies by type of approach.Hum Reprod. 2001 Jul;16(7):1473-8.

3) A Kiran et al. The risk of ureteric injury associated with hysterectomy: a 10-year retrospective cohort study. BJOG 2016 Jun;123(7):1184-91.

Disclosures

Drs Takeshita, Ito and Mr Matsuzaki are cofounders of Jmees Inc.. Drs Matsumoto and Tanabe received grants from Jmees Inc..

APPENDIX

Figures and pictures should be added in this section.

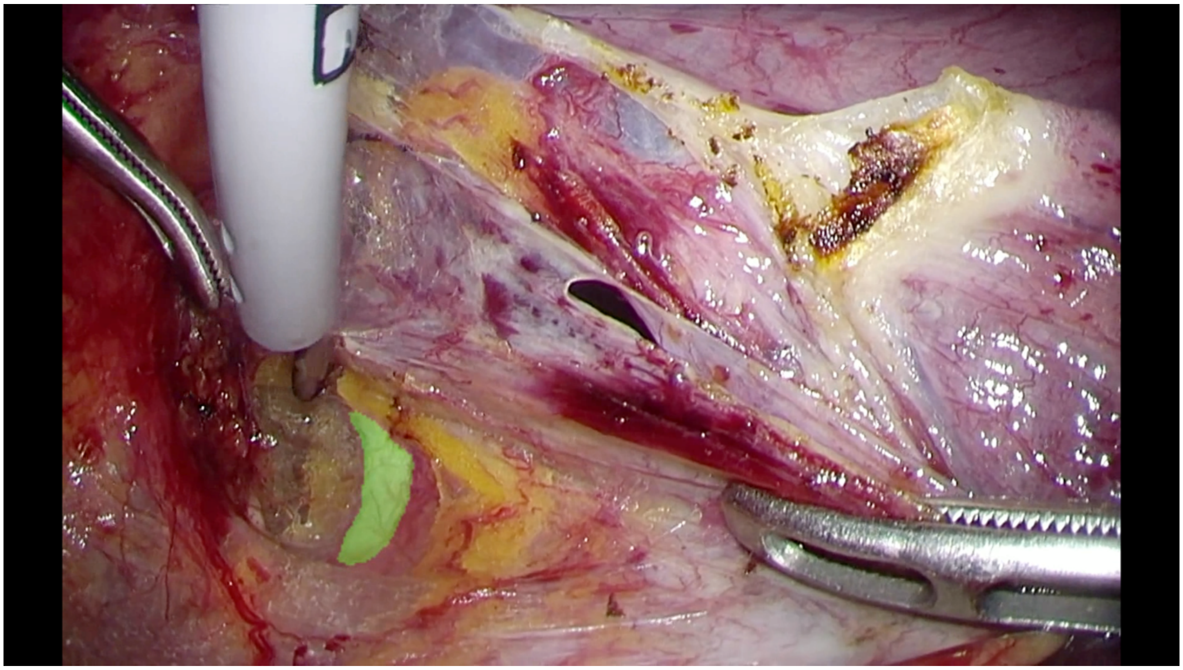
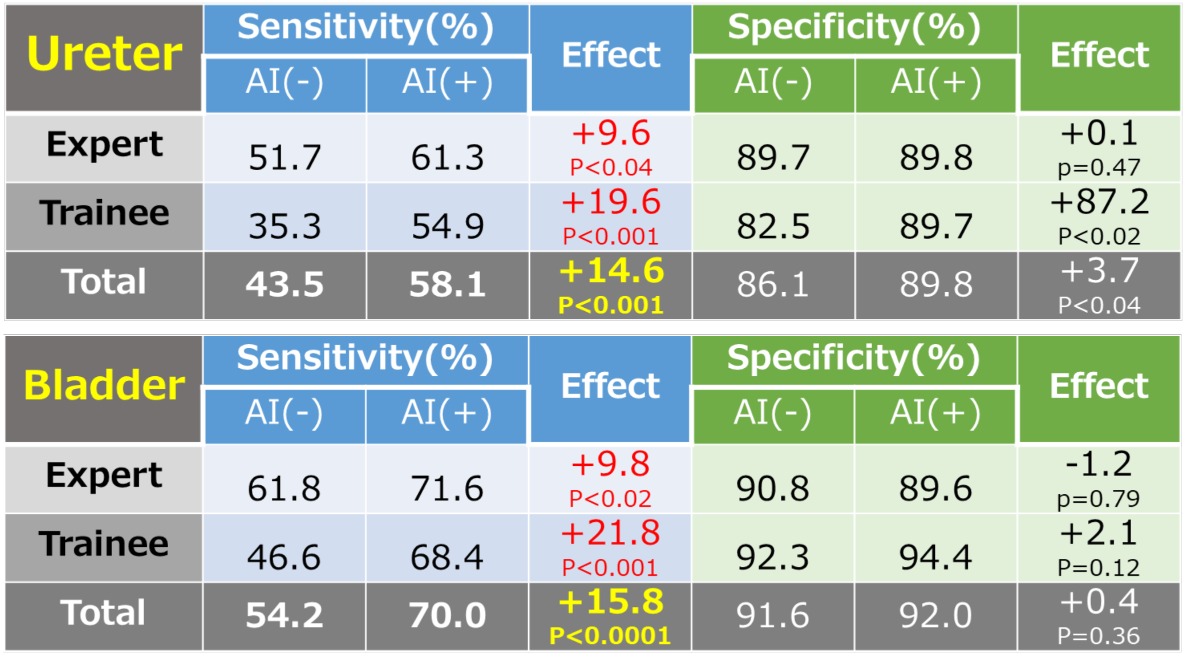


Figure 1: AI system identifies and highlights ureter to alert and assist surgeon's recognition.

Tables should be added here.

Table title: Positive effects for the recognition of anatomical structures by AI support



Ureter	Sensitivity(%)		Effect	Specificity(%)		Effect
	AI(-)	AI(+)		AI(-)	AI(+)	
Expert	51.7	61.3	+9.6 P<0.04	89.7	89.8	+0.1 p=0.47
Trainee	35.3	54.9	+19.6 P<0.001	82.5	89.7	+87.2 P<0.02
Total	43.5	58.1	+14.6 P<0.001	86.1	89.8	+3.7 P<0.04

Bladder	Sensitivity(%)		Effect	Specificity(%)		Effect
	AI(-)	AI(+)		AI(-)	AI(+)	
Expert	61.8	71.6	+9.8 P<0.02	90.8	89.6	-1.2 p=0.79
Trainee	46.6	68.4	+21.8 P<0.001	92.3	94.4	+2.1 P=0.12
Total	54.2	70.0	+15.8 P<0.0001	91.6	92.0	+0.4 P=0.36

Computer aided detection system for Barrett's neoplasia improves endoscopic detection by general endoscopists: an ex-vivo benchmarking study.

Authors:

M.R. Jong[1], K.N. Fockens[1], J.B. Jukema[1], T.G.W. Boers[2], C.H.J. Kusters[2], J.A. van der Putten[2], R.E. Pouw[1], L.C. Duits[1], N.S.M. Montazeri[3], B.L.A.M. Weusten[4,5], L. Alvarez Herrero[5], M.H.M.G. Houben[6], W.B. Nagengast[7], J. Westerhof[7], A. Alkhalaf[8], R.C. Mallant-Hent[9], P. Scholten[10], K. Ragunath[11], S. Seewald[12], P. Elbe[13,14], F. Baldaque-Silva[13,15], M. Barret[16], J. Ortiz Fernández-Sordo[17], G. Moral Villarejo[17], O. Pech[18], T. Beyna[19], F. van der Sommen[2], P.H. de With[2], A.J. de Groof[1], J.J. Bergman[1] on behalf of the BONS-AI consortium.

Affiliations:

1. Department of Gastroenterology and Hepatology of the Amsterdam University Medical Centers, Amsterdam, The Netherlands
2. Department of electrical engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
3. Biostatistics Unit, Department of Gastroenterology and Hepatology, Amsterdam University Medical Center, location Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
4. Department of gastroenterology and hepatology, UMC Utrecht, University of Utrecht, Utrecht, the Netherlands
5. Department of gastroenterology and hepatology, Sint Antonius hospital, Nieuwegein, the Netherlands
6. Department of gastroenterology and hepatology, HagaZiekenhuis Den Haag, Den Haag, the Netherlands
7. Department of gastroenterology and hepatology, UMC Groningen, University of Groningen, Groningen, the Netherlands
8. Department of gastroenterology and hepatology, Isala Hospital Zwolle, Zwolle, the Netherlands
9. Department of gastroenterology and hepatology, Flevoziekenhuis Almere, Almere, the Netherlands
10. Department of gastroenterology and hepatology, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands
11. Department of gastroenterology and hepatology, Royal Perth Hospital, Curtin University, Perth, Australia
12. Department of gastroenterology and hepatology, Hirslanden Klinik, Zurich, Switzerland
13. Department of digestive diseases, Karolinska University Hospital, Stockholm, Sweden
14. Division of Surgery, Department of Clinical Science, Intervention and Technology – CLINTEC, Karolinska Institutet, Stockholm, Sweden
15. Center for Advanced Endoscopy Carlos Moreira da Silva, Gastroenterology Department, Pedro Hispano Hospital, ULSM Matosinhos, Portugal
16. Department of gastroenterology and hepatology, Cochin hospital Paris, Paris, France
17. Department of gastroenterology and hepatology, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom
18. Department of gastroenterology and hepatology, St. John of God Hospital, Regensburg, Germany
19. Department of gastroenterology and hepatology, Evangalisches Krankenhaus Düsseldorf, Düsseldorf, Germany

Presenting author:

- *Martijn Roderick Jong*
- *Department of Gastroenterology and Hepatology of the Amsterdam University Medical Centers, Amsterdam, the Netherlands*
- *M.jong3@amsterdamumc.nl*

Keywords:

Barrett's Esophagus, Computer-aided detection, Endoscopy, Gastroenterology

Key information:

1. Research question: *to develop, test and benchmark a computer aided detection (CADe) system for Barrett's neoplasia*
2. Findings: *we developed a CADe system on the largest data set to date; performance is superior to general endoscopists and on par with experts; detection rate of general endoscopists significantly increases when they receive CADe assistance.*
3. Meaning: *using CADe as an assistive tool has the potential to increase the detection rate of endoscopists towards the level of experts*

MANUSCRIPT *(included in the word count)*

Introduction

Esophageal cancer is the 6th largest contributor to cancer-related deaths globally. Adenocarcinoma, a main subtype, has a fast rising incidence in Western cultures [1]. Barrett's esophagus (BE) is a well-known precursor for esophageal adenocarcinoma [2]. BE patients are therefore subject to regular endoscopic surveillance to detect neoplasia at an early stage. Detecting early BE neoplasia may be challenging for endoscopists as neoplastic lesions often have a subtle endoscopic appearance. Studies suggest that early lesions are missed on a regular basis [3]. Computer aided detection (CADe) systems may help to overcome this challenge.

The goal of this study was to develop, validate and benchmark a CADe system for early BE neoplasia.

Material and methods

First, the CADe system was pretrained with ImageNet followed by domain-specific pretraining with GastroNet, an in-house data set comprising over 5 million unlabeled endoscopic images from the gastrointestinal tract. The system was then trained with a large, heterogeneous data set of 14,146 white light images of 2,506 BE patients originating from 14 hospitals. All imagery had pathology confirmation by means of biopsy or resection specimen. Neoplasia segmentation was performed by 14 BE expert endoscopists. Total model size was 5.2 MB with an expected inference speed on an embedded FPGA of 20 frames per second and was developed to enable direct implementation onto current endoscopy platforms.

For external validation, the system was evaluated on two independent test sets. The "all-comers test set" comprised 119 consecutive cases (409 images, 251 videos) collected during a two-month interval, thereby representing daily clinical practice. The "benchmarking test set" comprised 175 cases (400 images, 188 videos) and was artificially enriched with challenging cases of subtle neoplasia (Figure 1). This test set was evaluated by 112 endoscopists from six countries. First without CADe assistance, and after a six-week wash-out period, with CADe assistance. Additionally, 28 external and internationally renowned BE expert endoscopists reviewed this test set.

Results

The CADe system detected virtually all neoplasia in the all-comers test set with an acceptable amount of false positives. In the benchmarking test set, CADe system was superior to endoscopists in detecting neoplasia and non-inferior to BE experts. With CADe assistance, neoplasia detection of endoscopists significantly increased without compromising specificity. Numerical data is described in Table 1.

Discussion and Conclusion

This study describes the rigorous development and evaluation of a CADe system for BE neoplasia using white light endoscopy on both still images and video. CADe outperformed endoscopists and, when used as an assistive tool, CADe significantly improves the detection rate of general endoscopists towards the level of experts. CADe detected nearly all neoplasia in a test set representing daily clinical practice.

The study comprises two limitations. First, the system is developed and evaluated on high-quality data from expert centers. To improve robustness and generalizability to community-based hospitals, more data should be collected within this domain. Second, this study was performed in an in-silico setting with no live AI-endoscopist interaction. We aim to perform a clinical pilot study soon to further investigate live performance.

References

1. Morgan E, Soerjomataram I, Rumgay H, et al. The Global Landscape of Esophageal Squamous Cell Carcinoma and Esophageal Adenocarcinoma Incidence and Mortality in 2020 and Projections to 2040: New Estimates From GLOBOCAN 2020. *Gastroenterology*. 2022;163(3):649-658.e2. doi:10.1053/j.gastro.2022.05.054
2. Sharma P. Barrett Esophagus: A Review. JAMA. 2022;328(7):663-671. doi:10.1001/jama.2022.13298
3. Schölvinck DW, van der Meulen K, Bergman JJGHM, Weusten BLAM. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. *Endoscopy*. 2017;49(2):113-120. doi:10.1055/s-0042-118312

Disclosures

This study received financial and logistical support from Olympus (Tokyo, Japan).

APPENDIX

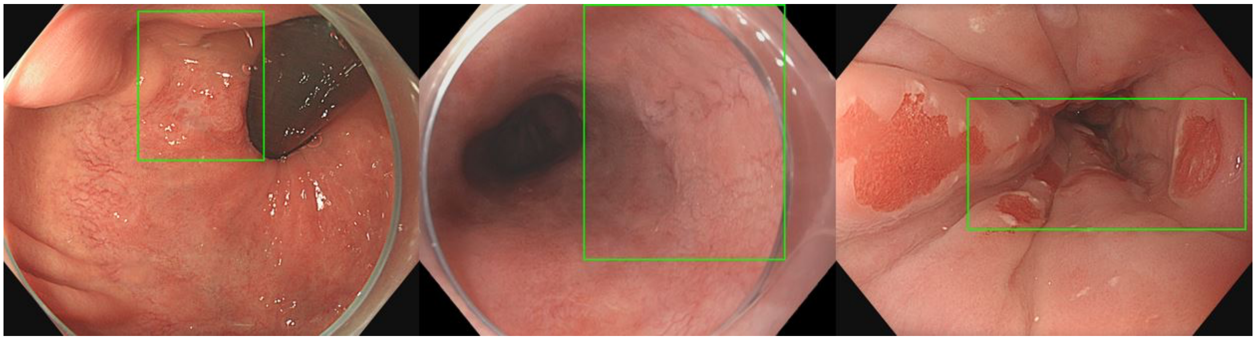


Figure 1. Example cases of Barrett's neoplasia in the benchmarking test set with corresponding CADe detections.

Table 1. Results of the CADe system and endoscopists on all test sets.

Test set	Scored by	Classification		Localization	
		Sensitivity	Specificity	Score	Method
All-comers image test set	CADe	95%	70%	100%	Bounding box
All-comers video test set	CADe	97%	85%	NA	NA
Benchmarking image test set	CADe	90%	80%	100%	Bounding box
	General endoscopists	74%	89%	92%	Biopsy mark
	General endoscopists with CADe	88%	90%	92%	Biopsy mark
	Expert endoscopists	87%	86%	94%	Biopsy mark
Benchmarking Video test set	CADe	91%	82%	NA	NA
	General endoscopists	67%	96%	100%	Biopsy mark
	General endoscopists with CADe	79%	94%	96%	Biopsy mark
	Expert endoscopists	86%	90%	96%	Biopsy mark

Note. NA = Not applicable

Surgery beyond Endoscopy

Moderators: Sandrine de Ribaupierre, MD and (pending)

Time	Presenting Author	Title
09:15-09:25	Bin Zheng (Surgical Simulation Research Lab, Dept. of Surgery, University of Alberta)	Surgical Team: Measuring the Shared Cognition and Performance
09:25-09:35	Andrew Wood (Cleveland Clinic Foundation)*	Sarcopenia and Hypoalbuminia are associated with decreased overall survival after Nephrectomy and IVC Thrombectomy for renal cell carcinoma
09:35-09:45	Jennifer Eckhoff (Surgical Artificial Intelligence and Innovation Laboratory, Massachusetts General Hospital)*	The SAGES Critical View of Safety Challenge – Infrastructure of a Biomedical Data Challenge from the Perspective of a Clinical Society
09:45-09:55	Mahdi Ebnali (MGB/Harvard Medical School)*	Using Deep Learning to Assess Teamwork During Cardiac Surgery
9:55-10:05	Surgery beyond Endoscopy Discussion (slush time)	

Surgical Team: Measuring the Shared Cognition and Performance

Authors: Bin Zheng, MD PhD[1]; Xianta Jiang, PhD[2]; M. Stella Atkins, PhD [3]; Roman Bednarik, PhD [4]

Affiliations: [1] Department of Surgery, University of Alberta; [2] Computing Science, Memorial University of Newfoundland; [3] Computing Science, Simon Fraser University; [4] School of Computing, University of Eastern Finland

Presenting author: Bin Zheng, Associate Professor, Department of Surgery, University of Alberta (bin.zheng@ualberta.ca)

Keywords: Surgical Team, Human Factors, Shared Cognition, Performance

Key information:

1. **Research questions:** Can we effectively measure the level of shared cognitions among members of a surgical team? Is there a correlation between increased shared cognition and improved team performance?
2. **Findings:** We found that a correlation between shared cognition towards surgical team tasks, measured by the similarity of answers to a list of surgical-related questions, and improved team performance are existed. Furthermore, leading surgical teams exhibited greater synchronization in eye movements (measured by the dual eye-tracking) and brain activities.
3. **Meaning:** Team cognition serves as the foundation for team performance. With the introduction of new tracking technologies, we expect to have more behavioural evidence available to assess team cognition and its impact on performance.

MANUSCRIPT *(word count: 760)*

Introduction

The capacity of an individual human operator to process information in complex surgical tasks is limited [1]. Therefore, collaboration among healthcare providers, including surgeons, nurses, and anesthesiologists from different specialties, is crucial in the operating room. Their collective expertise and teamwork are essential for achieving successful surgical outcomes. To optimize results, surgical team members must assess the situation and patient condition, effectively communicate important information, manage available resources, and synchronize their actions toward a common goal [1, 2]. This mutual understanding is often referred to as shared cognition.

However, it remains uncertain whether we can accurately measure the level of shared cognition among members of a surgical team and whether increasing shared cognition is correlated with improved team performance. In this study, we designed a surgical team task within a simulated environment for surgical residents, OR nurses, and anesthetists. We assessed shared knowledge regarding the surgical goal, patient condition, operating approach, and strategies for managing surgical crises using a set of multiple-choice questions (MCQs).

We hypothesized that an increase in shared cognition would correlate with improved team performance on the task, as measured by task completion time, objective assessment scores, and visual scanning patterns observed among team members.

Material and methods

This series of research was conducted at the Surgical Simulation Research Lab at the University of Alberta. A team consisting of six surgical residents, four OR nurses, and four anesthesia residents was assembled to perform a laparoscopic gastrectomy case within a simulated environment. During the operation, two unexpected events—a small artery bleeding and bowel perforation—were introduced to assess the response of the surgical team members to these situations.

Prior to the simulation training, each participant completed a questionnaire comprising 12 multiple-choice questions (MCQs) that covered various aspects such as patient condition, pre-surgical preparation, surgical goals, operating approach and steps, strategies for managing potential surgical crises, and post-operative management. The total score was adjusted to 100 points. Additionally, video footage was used to count the number of anticipatory movements performed by the nurse during the surgical procedure. The ability to perform anticipatory movements is influenced by a nurse's experience in participating in surgical cases [3-5].

The surgical performances of each three-member team were recorded through video and audio for the purpose of measuring operation time and evaluating their performance using the Observational Teamwork Assessment for Surgery (OTAS)[6]. The total score was reported in 100 points.

Two team members, specifically the surgeon and nurse, were required to wear head-mounted eye-trackers to record their eye scanning movements during the operation. The trajectories of their eye movements were analyzed using gaze overlap[7] and cross-recurrence analysis (CRA) [8]. By incorporating both spatial and temporal features of the eye scanpath into its calculations, CRA provides a more accurate mathematical outcome for assessing the similarity among different scanpaths, thereby enhancing our ability to describe shared cognition among team members.

Results

A total of 15 surgical teams were formed to complete 15 laparoscopic partial gastrectomy cases. On average, their MCQs scores were 86 ± 21 (maximum 100 points). The average procedure time was 37 ± 16 minutes, and average OTAS was 92 ± 33.

Correlation between MCQs scores and procedure time was weak but significant (r =-0.11; P < 0.05); MCQs scores and OTAS score was not significant correlated (r =0.06; P = 0.183).

On average, OR nurse performed 9 ± 6 anticipatory movements, including delivering grasper, scissors, and sutures to surgeons without needs for a verbal command from surgeons. Correlation between number of anticipatory movement and MSQs scores (r =0.24; P < 0.01) was moderate, and procedure time was weak but significant (r =-0.10; P < 0.05).

On average, surgeons and nurses achieved 34 % time during the procedure gazing on the same surgical areas. The gaze overlap correlated weakly with MSQs scores (r =0.12; P < 0.05). Cross Recurrence analysis yielded an improved outcome, the good performance teams (n=4) achieved 51 % gaze overlap, their gaze trajectories were more synchronized with a small phase delay (0.41 s). In contrast, the poor performance teams (n=3) only achieved 31% gaze overlap; their gaze trajectories were less synchronized with a larger phase delay (1.74 s).

Discussion and Conclusion

Team cognition forms the essential foundation for team performance. When team members possess a clear understanding of the team's goals, each other's roles within the team, and effective strategies for collaboration, they can develop a higher level of collective knowledge. This shared understanding allows team members to anticipate each other's actions and work together efficiently. In addition to traditional observational methods, advancements in tracking technologies, such as eye-tracking, provide a valuable way to measure team cognition, generating more substantial behavioral evidences to assess team performance.

References:

[1] Cannon-Bowers JA, Salas E, Converse SA. Cognitive psychology and team training: Training shared mental models and complex systems. Hum Factors Socie Bullet 1990;33:1-4.

[2] Salas E, Dickinson TL, Converse SA, Tannenbaum SI. Toward an understanding of team performance and training. Teams: Their training and performance. Westport, CT, US: Ablex Publishing, 1992. pp. 3-29.

[3] Zheng B, Swanström LL, MacKenzie CL. A laboratory study on anticipatory movement in laparoscopic surgery: a behavioral indicator for team collaboration. Surg Endosc 2007;21:935-40.

[4] Zheng B, Taylor MD, Swanström LL. An observational study of surgery-related activities between nurses and surgeons during laparoscopic surgery. Am J Surg 2009;197:497-502.

[5] Afkari H, Bednarik R, Mäkelä S, Eivazi S. Mechanisms for maintaining situation awareness in the micro-neurosurgical operating room. Int J Hum-Comput Int 2016;95:1-14.

[6] Undre S, Sevdalis N, Healey AN, et al. Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. World J Surg 2007;31:1373-81.

[7] Khan RS, Tien G, Atkins MS, et al. Analysis of eye gaze: do novice surgeons look at the same location as expert surgeons during a laparoscopic operation? Surg Endosc 2012;26:3536-40.

[8] Hajari N, Cheng I, Zheng B, Basu A. Determining team cognition from delay analysis using cross recurrence plot. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016. pp. 3482-5.

Disclosures

Authors do not have any potential conflict of interest handled during the work

Sarcopenia and Hypoalbuminia are associated with decreased overall survival after Nephrectomy and IVC Thrombectomy for renal cell carcinoma

Authors:

Andrew M Wood MD[1], Nour Abdallah MD[1], Sohan Shah BS[1], Crystal An BA[1], Saeid Mirzai DO[2], Ian Persits DO[2], Po-Hao Chen MD[3], Christopher Weight MD[1]

Affiliations:

1. *Glickman Urological and Kidney Institute, Cleveland Clinic Foundation.*
2. *Department of Internal Medicine, Cleveland Clinic Foundation.*
3. *Imaging Institute, Cleveland Clinic Foundation.*

Presenting author:

Andrew M Wood MD, Glickman Urological and Kidney Institute, Cleveland Clinic Foundation.
Email: wooda10@ccf.org.

Keywords:

Renal cell carcinoma, Sarcopenia, Automated imaging segmentation, CT scan, Preoperative Risk Stratification

Key information:

1. Research question: Does sarcopenia as measured by fully automated AI-based software examination of preoperative CT scans predict overall survival following radical nephrectomy and IVC thrombectomy.
2. Findings: Sarcopenia as measured by a fully automated segmentation system, in combination with other measures of nutrition and cytoreductive intent, is an independent predictor of overall survival following IVC thrombectomy.
3. Meaning: Sarcopenia, hypoalbuminia, and cytoreductive intent are associated with inferior OS after Nephrectomy and IVC thrombectomy. These variables should be utilized in preoperative risk stratification.

MANUSCRIPT *(included in the word count)*

Introduction

Sarcopenia has recently been shown to be an important predictor of outcomes in cancer patients. However, despite the significant morbidity associated with the procedure, there exists no study examining the impact of body composition and sarcopenia on survival following nephrectomy and inferior vena cava (IVC) thrombectomy. We aimed to assess associations of sarcopenia, muscle density, and albumin levels with overall survival (OS) after nephrectomy and IVC thrombectomy for renal cell carcinoma.

Material and methods

A total of 179 patients undergoing nephrectomy with IVC thrombectomy from 2006 to 2022 had sufficient clinical data and available digitized preoperative CT scan of the abdomen. Fully automated multi-slice measurements of skeletal muscle volume and density were made using the Data Analysis Facilitation Suite (Voronoi Health Analytics, Vancouver, Canada) at the mid L3 level. Skeletal muscle index (SMI) was calculated with the skeletal muscle area (cm2) normalized for height (m2), and skeletal muscle density (SMD) was calculated from average Hounsfield units. OS was estimated with the Kaplan-Meier method. Associations between body composition, preoperative albumin, preoperative measures of inflammation, and relevant clinical variables and OS were assessed with univariable and multivariate Cox analyses.

Results

103 of the 179 patients (56.4%) were sarcopenic. 56 of 179 surgeries (31.3%) were cytoreductive in nature, and 82 (45.8%) involved level 3 or 4 IVC thrombus. Sarcopenia was associated with increased age (p<0.001), lower preop albumin (p=0.012), and lower SMD (p=0.006). The median OS was 22.7 and 62.0 months for sarcopenic and nonsarcopenic patients, respectively (P = .027). On univariate analysis, cytoreductive intent (p<0.001), elevated neutrophil lymphocyte ratio (p=0.047), and hypoalbuminemia (p=0.015) were all also associated with OS. On multivariate cox analysis, sarcopenia, hypoalbuminemia, and cytoreductive intent were all independently associated with OS. The worst OS was observed in sarcopenic patients with hypoalbuminemia undergoing cytoreductive surgery (median OS 7.9 months) vs the best in non-sarcopenic patients with normal albumin and no metastatic disease (median OS 67.4 months).

Discussion and Conclusion

Sarcopenia, hypoalbuminia, and cytoreductive intent are associated with inferior OS after Nephrectomy and IVC thrombectomy. These variables should be utilized in preoperative risk stratification, and additional consideration for preoperative systemic therapy should be given to those at highest risk of poor survival. Limitations of this study include the retrospective nature which introduces the possibility of selection bias, and the single center experience which may limit external validity.

APPENDIX

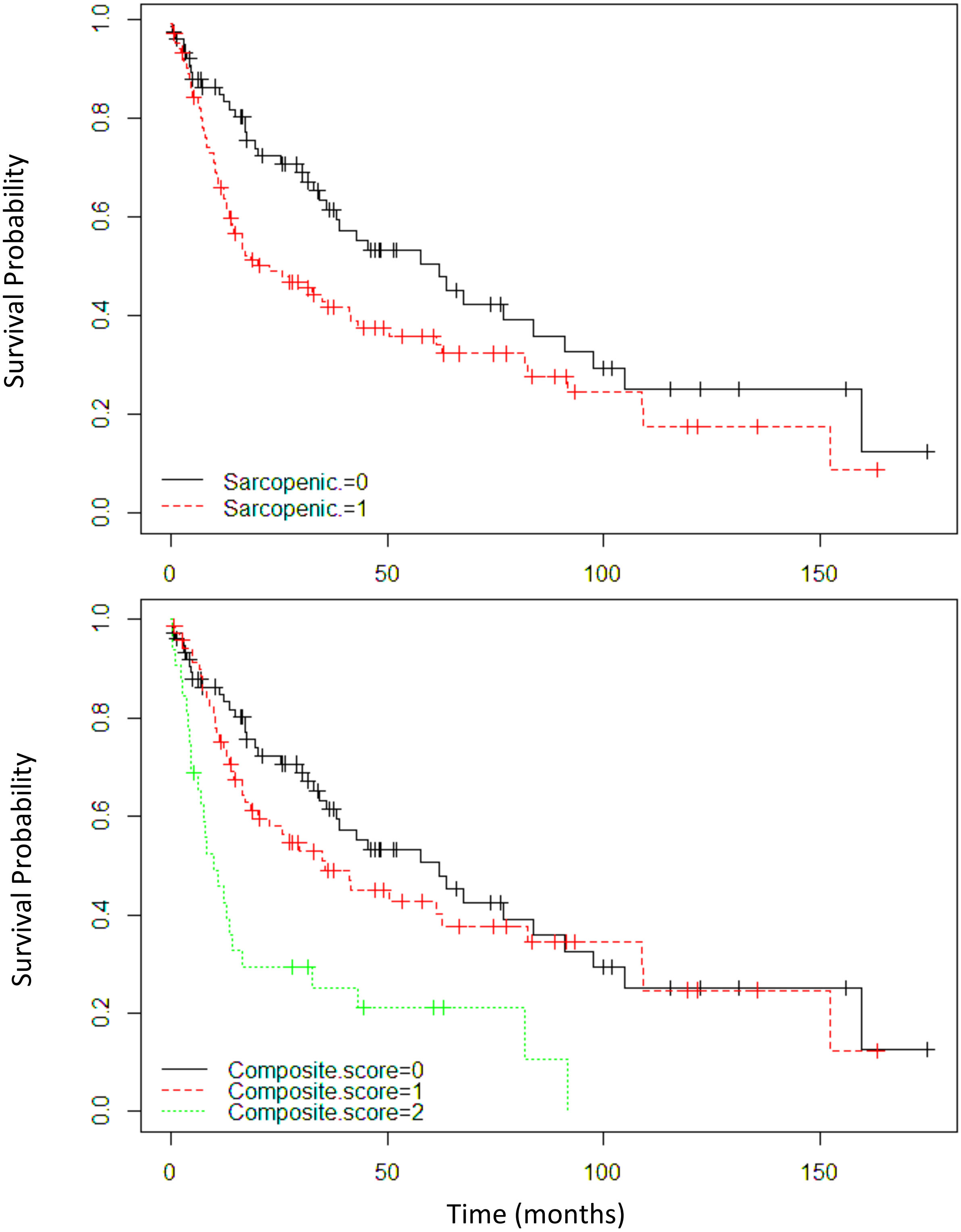


Figure 1: Kaplan Meier Curves for Overall Survival in IVC Thrombectomy Patients: Stratified by Sarcopenia and Hypoalbuminemia

The SAGES Critical View of Safety Challenge – Infrastructure of a Biomedical Data Challenge from the Perspective of a Clinical Society

Authors:

Eckhoff JA*[1, 2], Li X[3], Ban Y*[1], Alapatt D[4, 7], Mazallier JP[4, 7], Mascagni P*[4, 5], Lyu Z[6] , Choksi S[8], Filicori F[8], Rosman G[1, 5], Hashimoto DA[9] , Li Q[5], Padoy N[4, 7], Meireles OR[1]

Affiliations:

1 - Surgical Artificial Intelligence and Innovation Laboratory, Department of Surgery, Massachusetts General Hospital, 15 Parkman Street, WAC339, Boston, MA 02114, USA

2 - Department of General, Visceral, Tumor and Transplant Surgery, University Hospital Cologne, Kerpenerstrasse 62, 50937 Cologne, Germany

3 - Center For Advanced Medical Computing And Analysis (CAMCA), Harvard University, Boston, USA

4 - IHU-Strasbourg, Institute of Image-Guided Surgery, Strasbourg, France

5 - Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

6 - Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA

7 - ICube, University of Strasbourg, CNRS, France

8 - Intraoperative Performance Analytics Laboratory (IPAL), Department of General Surgery, Northwell Health, Lenox Hill Hospital, New York, NY, USA

9 - Hospital of the University of Pennsylvania, Department of Surgery, Penn Computer Assisted Surgery and Outcomes (PCASO) Laboratory

Presenting author:

Jennifer A. Eckhoff, MD

jeckhoff@mgh.harvard.edu

Keywords:

Data Challenge, Surgery, Risk Mitigation, Society of American Gastrointestinal Surgeons, Interdisciplinary Collaboration

Key information:

1. Research question: *What are critical insights from organizing a biomedical data challenge from the perspective of a surgical society?*
2. Findings: *We present the preliminary insights gained from the SAGES CVS Challenge. We elaborate on consensus-based findings of three advisory committees consisting of experts in the field of surgical data science on the individual stages of the challenge.*
3. Meaning: *The SAGES CVS Challenge provides a standardized infrastructure for future surgical data challenges by fostering the collection of a global surgical dataset reflective of the real-world diversity and leveraging interdisciplinary collaboration to achieve robust, scalable, and reliable AI for risk mitigation in surgery.*

MANUSCRIPT *(included in the word count)*

Introduction

Data challenges have led to considerable breakthroughs in Artificial Intelligence (AI). They are, therefore, a common methodology among computer scientists in acquiring large-scale datasets to explore various computational solutions for distinct real-world issues. Among the most famous examples is the ImageNet Challenge[1], which spurred the development of deep learning algorithms, particularly convolutional neural networks. In surgery, significant examples include the MICCAI Endoscopic Vision Challenge (EndoVis)[2], focused on instrument segmentation, and video understanding in endoscopy, and the Robotic Vision Challenge (ROBVIS)[3], which addresses the classification of tissue and anatomic structures and surgical scene understanding. These challenges target important spatiotemporal features and interdependencies related to surgical success. However, they are predominantly organized and therefore influenced by computer scientists.

In contrast, the Critical View of Safety (CVS) Challenge, administrated by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), incorporates the perspectives of computer scientists, surgeons, and industry representatives and presents a diverse, globally acquired surgical video dataset. The challenge targets a universally recognized surgical safety measure in laparoscopic cholecystectomy (minimally invasive Gallbladder removal) – the Critical View of Safety (CVS)[4–6]. The procedure is widely recognized as the benchmark for surgical AI, and the CVS presents an ideal use case for computer vision analysis and exploration by the global data science community.

Material and methods

From its inception in 2021, the CVS Challenge Organizers assembled three advisory committees (AC), consisting of clinical and technical experts in surgical data science and industry representatives, for consensus-based decision-making. The AC's expertise was deployed in three focus areas of the challenge: (A) Data Structure and Acquisition, (B) Data Use and Annotation, and (C) Governance and Execution of Surgical Data Challenges. A partnership with Surgical Safe Technologies (SST) was established to develop a video acquisition portal for safe and reliable deidentification and uploading of surgical video data. Two summits were conducted in 2022 and 2023 to compose detailed pipelines for each focus area. A three-round Delphi Consensus led to the composition of an annotation guideline and curriculum, and an additional Delphi is currently being conducted to compose challenge metrics and competition rules.

Results

We present the preliminary results of this global initiative and elaborate on key findings from conducting a biomedical data challenge from the perspective of a surgical society. To date, we acquired 413 videos from 44 countries through the SAGES Video Acquisition Portal with an additional data donation of 150 videos from other countries through Dropbox Business. We report on the composition of an 'Annotation Guideline' and the successful implementation of an 'Annotation Curriculum' for standardized training of annotators based on proficiency-based progression. Moreover we present the AC's perspective on evaluation approaches of challenge results and thoughts on trade-offs between compute and evaluation rigorousness.

Discussion and Conclusion

The SAGES CVS Challenge provides a comprehensive guideline for conducting a surgical data challenge. By addressing the stages from video recording in the operating room, HIPAA/GDPR compliant data storage and distribution, surgical video annotation to dataset release, and challenge organization, we aim to establish a standardized infrastructure, and evaluation criteria for future surgical challenges. The Video Data Acquisition Portal ensures compliance with worldwide varying legal and ethical regulations to facilitate global contributions to increase data diversity. The Annotation pipeline promises robust, reproducible, and clinically relevant annotations, scalable to other use cases. And finally, leveraging interdisciplinary collaboration and the expertise of clinicians, computer scientists, and industry alike will allow us to conduct a competition resulting in generalizable and universally applicable AI. Ultimately this interdisciplinary initiative aims to foster the development of AI for real-time deployment in the operating room to enhance surgical safety.

References

1. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015;115(3):211-252.

2. Bodenstedt S, Allan M, Agustinos A, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv [csCV]. May 2018. http://arxiv.org/abs/1805.02475.

3. Hall D, Talbot B, Bista SR, et al. The Robotic Vision Scene Understanding Challenge. arXiv [csRO]. September 2020. http://arxiv.org/abs/2009.05246.

4. Mascagni P, Alapatt D, Garcia A, et al. Surgical data science for safe cholecystectomy: a protocol for segmentation of hepatocystic anatomy and assessment of the critical view of safety. arXiv [eessIV]. June 2021. http://arxiv.org/abs/2106.10916.

5. Namazi B, Iyengar N, Hasan S, et al. AI for Automated Detection of the Establishment of Critical View of Safety in Laparoscopic Cholecystectomy Videos. J Am Coll Surg. 2020;231(4):e48.

6. Gupta V, Jain G. Safe laparoscopic cholecystectomy: Adoption of universal culture of safety in cholecystectomy. World J Gastrointest Surg. 2019;11(2):62-84.

Disclosures

The Critical View of Safety Challenge is conducted by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). This data challenge is sponsored and supported by an industry advisory committee of representatives from Surgical Safe Technologies Inc., Intuitive Surgical Inc., Medtronic Plc., and Olympus (SST Inc.).

USING DEEP LEARNING TO ASSESS TEAMWORK DURING CARDIAC SURGERY

Authors:

Mahdi Ebnali Harari (1,2)
Marco Zenati (2,3)
Vaibhav Unhelkar (4)
Steven Yule (5)
Roger Dias (1,2)

Affiliations:

1. *Harvard Medical School, Boston, MA, USA*
2. *Mass General Brigham, Boston, MA, USA*
3. *VA Boston Healthcare System, West Roxbury, MA, USA*
4. *Department of Computer Science, Rice University, Houston, TX, USA*
5. *Department of Clinical Surgery, University of Edinburgh, Scotland*

Presenting author:

*Mahdi Ebnali Harari, Harvard Medical School (*mebnali-heidari@bwh.harvard.edu

Keywords:

Non-technical skills, teamwork, deep learning, cardiac surgery

Key information:

1. Research question: This study aimed to assess the feasibility of a deep learning (DL) approach to objectively evaluate critical non-technical skills – namely, teamwork performance – during real-life cardiac surgery.
2. Findings: *The results demonstrated that teams with higher teamwork ratings presented lower median team displacement extracted from video analysis and lower PSD across most frequency bands compared to low performance teams, indicating less irregularity of body motion patterns at the team level during the separation from the bypass phase of cardiac surgery.*
3. Meaning: *This study shows the feasibility of using deep learning to analyze teamwork performance based on OR video recordings. These findings along with future studies may help to establish more standardized and objective methods for evaluating teamwork and other non-technical skills in the OR.*

MANUSCRIPT *(included in the word count)*

Introduction

The successful delivery of surgical care in high-risk and complex environments, such as cardiac surgery, depends on effective teamwork among clinicians. The significance of these skills as part of non-technical skills (NTS) has been widely recognized in surgical settings [1-2]. The evaluation of teamwork in the OR has traditionally relied on post-hoc subjective and qualitative methods, such as observational rating scales and self-assessment tools. Real-life skill assessment usually involves the direct observation of surgical trainees or retrospective analysis of operation videos, with skills rated by experts based on predefined criteria. Although these methods provide a realistic assessment and can be blinded, they are constrained by limited reproducibility and rater availability [3]. Furthermore, subjective methods suffer from significant limitations, such as observer recall bias, subjective perception of performance, and lack of specificity. The unstructured and descriptive approaches to teamwork assessment may also fail to capture the complexities of team dynamics in the OR. These limitations emphasize the importance of developing more objective and quantitative methods for evaluating teamwork in the OR [4-6]. Recent studies have explored various quantitative methods for evaluating surgical team performance, such as motion-capturing suits and gloves, gesture detection using specialized RGBD cameras like Microsoft Kinect, body optical markers, wearable device sensors, and instrument tracking. Despite the potential benefits of these technologies, their implementation in OR environments remain challenging due to concerns surrounding privacy and patient safety [1]. Moreover, several hurdles must be overcome to integrate these technologies into clinical practice, including the need for specialized hardware and software, the cost and complexity of data collection and analysis, and potential disruptions to the clinical workflow [7]. Recent advancements in machine learning have enabled the application of DL methods in computer vision (CV) for the assessment of surgical skills[4,8]. Specifically, automated surgical skill assessment in robotic interventions has been a subject of interest due to the accessibility of kinematic data and video recordings from the console[9]. However, there is a significant gap in the current research, which mainly focuses on evaluating technical surgical skills, leaving the potential of DL methods in assessing non-technical surgical skills, such as teamwork, largely unexplored. This study aimed to assess the feasibility of a DL approach to objectively evaluate two critical non-technical skills – namely, teamwork performance – during real-life cardiac surgery.

Material and methods

Participants: A cardiovascular OR team comprises 4 subteams: (1) a cardiac surgical team consisting of an attending surgeon, one or more residents and fellows, and a surgical physician assistant; (2) anesthesiology team consisting of an attending anesthesiologist, one or more residents/fellows, and a nurse anesthetist; (3) a perfusion team consisting of a lead and an assistant perfusionist; and (4) a nursing team consisting of scrubbed nurses and circulators. We recorded audio and video data from the entire team using three lapel microphones (surgeon, anesthesiologist, and perfusionist) and two cameras (narrow and wide field of view). This research was approved by the Institutional Review office. Informed consent was obtained from all participants, which included patients and all OR staff involved with the procedures. Data were collected during 30 cardiac surgery procedures.

NOTSS: Three trained raters used the NOTSS tool [10] to evaluate teamwork from 30 cardiac surgery video recordings. The NOTSS tool has four categories (situation awareness, decision-making, teamwork & communication, and leadership). Surgical teams were rated during separation from cardiopulmonary bypass, a critical phase of cardiac surgery using a 1-4 Likert scale. The rating for teamwork category was used to categorize teams into two groups: teams with ratings within the first tertile were classified as "Low Performance," while teams within the third tertile were classified as "High Performance". Table 1 presents an overview of the four categories of NOTTS and the corresponding behavioral elements within each category. The current study focused on the teamwork category, which encompasses exchange of information, establishment of shared understanding, and coordination of activities among team members.

DL Method: A previous study proposed a methodology based on the OpenPose library for analyzing dynamic changes in OR teams using RGB camera-recorded video data of OR staff positions during procedures [4,8] (Fig.1). In this study, we used this methodology to extract body pose estimations from each video at 30 frames per second, identifying the position of keypoints (17 keypoints, Figure 2) in the OR staff's body. OpenPose is an open-source software library released under the Apache 2.0 license and provides real-time multi-person keypoint detection and tracking for body, hands, and facial pose estimation from video streams or images. This library is based on DL and specifically uses a convolutional neural network (CNN) architecture, with the capability of running on both CPU and GPU platforms. Although OpenPose is not specifically designed for the OR, this approach was deemed appropriate for analysis in the OR due to its ability to accurately estimate body pose in complex and dynamic environments. However, it is important to note that some features important to the OR context may be missing from the approach, such as the identification of individuals by name or role. Despite these limitations, our focus was on team dynamics and collective motions of individuals, rather than individual motion. Therefore, the use of this approach allowed us to analyze and extract the team's movements and dynamics from the video data, which was essential to achieving our research objectives.

Team Motion Metrics: Average team displacement: The x and y coordinates of the neck keypoint (Figure 2) were used to calculate the Euclidean distance between the neck and a reference point (x = 0, y = 0). The average displacement per frame in pixels was then calculated across all team members to capture the entire team motion and was subsequently averaged over 1-second epochs.

Entropy: For each second, the team displacement was classified into one of 3 states (S1, S2, S3) based on which tertile that value was in the entire motion data distribution. The distribution of these states over time was quantified by calculating Shannon's entropy (H) in bits, using a 30-second sliding window updated each second. Restricted symbol expression represents low entropy, which means there is a higher level of organization in the team motion. Power Spectral Density (PSD): We also calculated the PSD of team displacement using Welch's method to analyze the frequency of team motion data. This method provides information about the frequency content of the signal and can be used to identify patterns or characteristics in the data that may not be detectable in the time domain. We chose a sampling frequency of 1 Hz, a segment length of 1000 samples, and a frequency range of 0-1 Hz to capture the various frequencies of team displacement. Previous studies reported that compared to novice, expert clinicians show smoother and less variable motion patterns and have a lower PSD and more concentrated frequency distribution [10].

Statistical Analysis: The rating data for teamwork from NOTSS was tested for normality using the Kolmogorov-Smirnov test and found to be non-normal. Therefore, the data was summarized as median (1st-3rd interquartile). For the comparison of average team displacement, entropy, and PSD, t-tests were used as these data were found to be normally distributed.

Results

NOTSS: A total of 22 cardiac surgery teams (first and third terciles) were analyzed. The median teamwork score was 3.5 (3.3-3.7) in the teams with the “High Performance” group (N=11) and 3.0 (2.8-3.1) in the teams with the “Low Performance” group (N = 11). Team displacement: The median displacement of the “High Performance” group had a median displacement of 98.61 pixels (30.25-183.32) and the “Low Performance” group was 100.82 pixels (19.06-190.16) (Figure 3). The statistical analysis showed a significant difference between the two groups (p=0.012, t= 2.44). Entropy: The median of entropy for the "High Performance" group was 0.89 bits (0.63 - 1.03), while the median of entropy for the "Low Performance" group was also 0.9 bits (0.68 - 1.04). No statistically significant difference was found in entropy between the two groups (p = 0.305). PSD: Based on the Welch's t-test on the PSD values of team motion data grouped by "High Performance" and "Low Performance", we found that there was a statistically significant difference in the mean PSD values between the two groups (p = 0.03, t = - 2.22. The mean PSD for the "Low Performance" group (39.217 dB/Hz) was higher than that of the "High Performance" group (38.577 dB/Hz) in the frequency range of 0.1-1 Hz, indicating higher energy in almost all frequency ranges for the "Low Performance" group (Figure 4).

Discussion and Conclusion

The objective of this study was to investigate the feasibility of a DL method based on the OpenPose library for the objective assessment of teamwork, which is critical non-technical skills in cardiac surgery. Our results demonstrate the feasibility of using motion metrics of surgical teams as a proxy for assessing their teamwork. Specifically, we found that the median team displacement, which reflects the overall motion of the surgical team, was significantly different between the "High Performance" and "Low Performance" groups. The "High Performance" group had a slightly smaller median displacement, indicating that they were more efficient and coordinated in their movements compared to the "Low Performance" group. In addition, we investigated the use of entropy and PSD as metrics for assessing the complexity and stochasticity of team motion. Our findings did not reveal a significant difference in entropy between the "High Performance" and "Low Performance" groups. This is in contrast to previous studies that have used metrics such as team motion to measure situational awareness at the team level during cardiac surgery [4]. The entropy of physiological metrics was also found sensitive enough to detect variations in team cognitive load and uncertainty. It is possible that our sample size was not large enough to detect a significant difference in entropy. Future studies with larger sample sizes and in different contexts could further explore the utility of entropy as a metric for assessing surgical teamwork. We observed a significant difference in PSD between the two groups. The "Low Performance" group had a higher mean PSD in the frequency range of 0.1-1 Hz, indicating higher energy in almost all frequency ranges compared to the "High Performance" group. This finding suggests that the "Low Performance" group exhibited more irregular and unpredictable team motion, which is indicative of low performance teamwork. Similarly, [10] reported that PSD can be used to differentiate between skilled and unskilled surgeons in laparoscopic tasks based on the frequency components of their motion data. This study found that skilled surgeons had lower energy (PSD) in the highfrequency components of their motion data, compared to unskilled surgeons, indicating more controlled movements.

Limitations and Future Directions: While our study provides promising results regarding the feasibility of using motion metrics and DL methods for objective assessment of surgical teamwork, there are several limitations to be considered. One limitation is the small sample size, which may have affected the ability to detect significant differences in motion metrics. Another limitation is the use of only one type of surgical procedure, which limits the generalizability of the findings to other types of surgeries. Additionally, our study only considered team motion metrics and did not take into account other potential factors that may impact teamwork, such as task complexity, team composition, and team familiarity. Future studies with larger sample sizes and diverse surgical procedures, along with the integration of additional factors, are necessary to further explore the utility of DL methods for objective assessment of surgical teamwork. While the NOTSS framework includes both teamwork and communication skills, our approach focuses primarily on analyzing teamwork through computer vision-based motion analysis. However, a limitation of our approach is that it may not capture all aspects of non-technical skills, including communication skills that rely on verbal cues. To address this limitation, future studies could explore the integration of additional data sources, such as physiological or speech data, to supplement the analysis of communication skills. Future studies need to explore the integration of team motion metrics data with physiological data, such as heart rate variability, to create more advanced computer-based cognitive systems that aim to augment the cognitive capabilities of surgical teams. Furthermore, by integrating various sources of data including individual and team motion metrics, psychophysiological parameters, instrument tracking metrics, and action recognition variables, it is conceivable to develop automated and intelligent systems that can continuously monitor the surgical workflow and evaluate technical and nontechnical surgical skills. This multi-modal approach could enable real-time analysis of the surgical team's performance and facilitate the provision of corrective feedback, ultimately leading to improved surgical outcomes and enhanced patient safety. Moreover, future studies also need to consider integrating patient outcomes data and preoperative patient data into automated assessment of teamwork, as this can provide a more holistic understanding of the factors that impact surgical outcomes and patient safety. Such integration can also help identify areas for improvement in team performance and communication, leading to enhanced patient care and better surgical outcomes.

Conclusion: The present pilot study demonstrates the potential of using DL algorithms to analyze teamwork performance based on standard surgical video recordings. The results provide preliminary evidence for the feasibility of using motion metrics to evaluate nontechnical skills, with implications for the development of standardized and objective evaluation methods for assessing these skills in the surgical environment. Furthermore, future research may build upon our findings to further refine these methods and establish their efficacy for improving performance assessment and quality improvement initiatives in the field of cardiac surgery.

References

1. Cha JS, Yu D. Objective Measures of Surgeon Non-Technical Skills in Surgery: A Scoping Review. Hum. Factors 2022;64:42–73.

2. Lauren R. Kennedy-Metz, Roger D. Dias, Heather Conboy, Maria Arshanskiy, Christopher Rioux, Sandra Park, Mahdi Ebnali, Annette Phillips, Marco A. Zenati. Surgical Workflow Distractions and Team Non-Technical Skills during the Separation from Bypass Phase of Cardiac Surgery. 2023.

3. Nicolaides M, Cardillo L, Theodoulou I, Hanrahan J, Tsoulfas G, Athanasiou T, et al. Developing a novel framework for non-technical skills learning strategies for undergraduates: A systematic review. Ann Med Surg (Lond) 2018;36:29–40. 4. Dias RD, Kennedy-Metz LR, Yule SJ, Gombolay M, Zenati MA. Assessing Team Situational Awareness in the Operating Room via Computer Vision. In: 2022 IEEE Conference 9 on Cognitive and Computational Aspects of Situation Management (CogSIMA). 2022. page 94–6.

5. Ebnali M, Kennedy-Metz LR, Conboy HM, Clarke LA, Osterweil LJ, Avrunin G, et al. A Coding Framework for Usability Evaluation of Digital Health Technologies. In: HumanComputer Interaction. Theoretical Approaches and Design Methods. Springer International Publishing; 2022. page 185–96.

6. Seo S, Kennedy-Metz LR, Zenati MA, Shah JA, Dias RD, Unhelkar VV. Towards an AI Coach to Infer Team Mental Model Alignment in Healthcare. IEEE Conf Cogn Comput Asp Situat Manag 2021;2021:39–44.

7. Dias RD, Ngo-Howard MC, Boskovski MT, Zenati MA, Yule SJ. Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. Br. J. Surg. 2018;105:491–501.

8. Mahdi Ebnali, Marco A. Zenati, Roger D. Dias. SURGICAL TEAM PERFORMANCE ANALYSIS USING COMPUTER VISION: A METHODOLOGY and USE CASE STUDY. Human Factors and Ergonomics in Healthcare; 2023.

9. Nagyné Elek R, Haidegger T. Robot-Assisted Minimally Invasive Surgical Skill Assessment—Manual and Automated Platforms. ACTA POLYTECHNICA HUNGARICA 2019;16:141–69.

10. Zecca M, Cavallo F, Saito M, Endo N, Mizoguchi Y, Sinigaglia S, et al. Using the Waseda Bioinstrumentation System WB-1R to analyze Surgeon's performance during laparoscopy - towards the development of a global performance index - [Internet]. 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems2007;Available from: http://dx.doi.org/10.1109/iros.2007.4399579

Disclosures

This work was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health [R01HL126896, R01HL157457]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank the clinical staff from the Veterans Affairs (VA) Boston Healthcare System for their valuable insights and collaboration in this study.

APPENDIX

Figures.

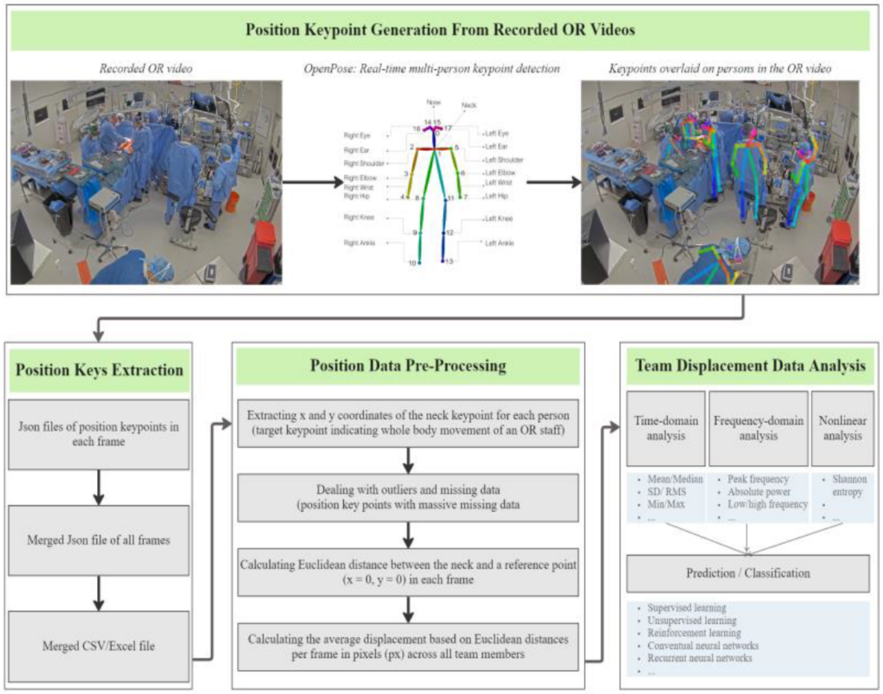

Position Keypoint Generation From Recorded OR Videos
Recorded OR video
OpenPose: Real-time multi-person keypoint detection
Keypoints overlaid on persons in the OR video
Nose
Neck
Right Eye
Right Ear
Right Shoulder
Right Elbow
Right Wrist
Right Hip
Right Knee
Right Ankle
Left Eye
Left Ear
Left Shoulder
Left Elbow
Left Wrist
Left Hip
Left Knee
Left Ankle
14 15 16 17 0 1 2 3 4 5 6 7 8 9 10 11 12 13
Position Keys Extraction
Json files of position keypoints in each frame
Merged Json file of all frames
Merged CSV/Excel file
Position Data Pre-Processing
Extracting x and y coordinates of the neck keypoint for each person (target keypoint indicating whole body movement of an OR staff)
Dealing with outliers and missing data (position key points with massive missing data
Calculating Euclidean distance between the neck and a reference point (x = 0, y = 0) in each frame
Calculating the average displacement based on Euclidean distances per frame in pixels (px) across all team members
Team Displacement Data Analysis
Time-domain analysis
Frequency-domain analysis
Nonlinear analysis
Mean/Median
SD/ RMS
Min/Max
...
Peak frequency
Absolute power
Low/high frequency
...
Shannon entropy
...
Prediction / Classification
Supervised learning
Unsupervised learning
Reinforcement learning
Conventual neural networks
Recurrent neural networks
...

Figure 1: A DL-based methodology for surgical team performance analysis [4,8]

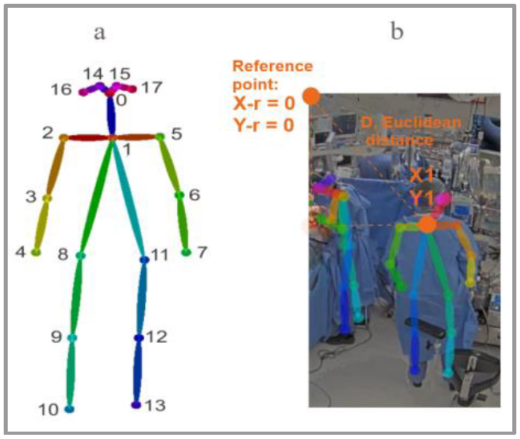


Figure 2. (a) Keypoints that can be detected by the OpenPose algorithm, (b) example of how Euclidean distance is used to measure the displacement of each individual of OR staff.

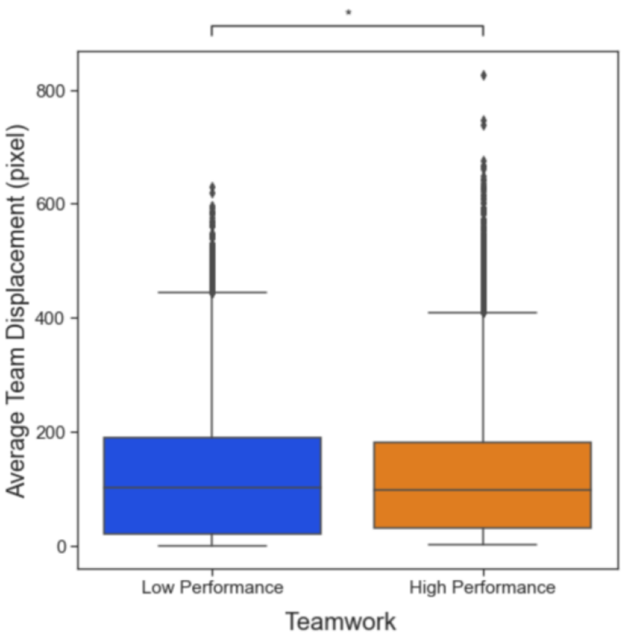


Figure 3. Team displacement between “High Performance” and “Low Performance” groups (* denotes for p <0.05)

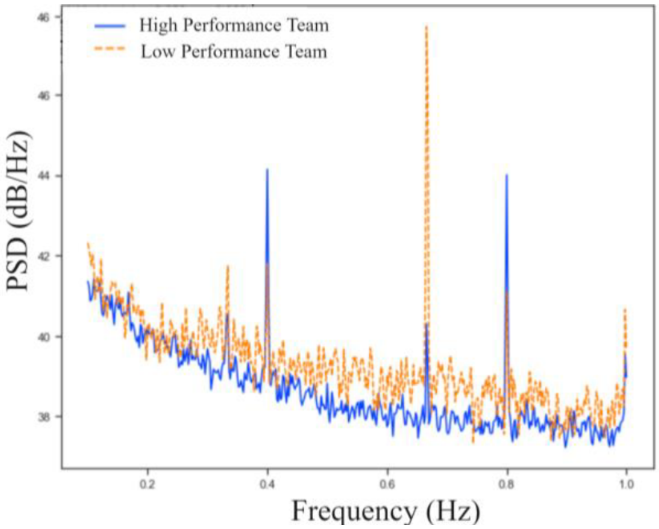
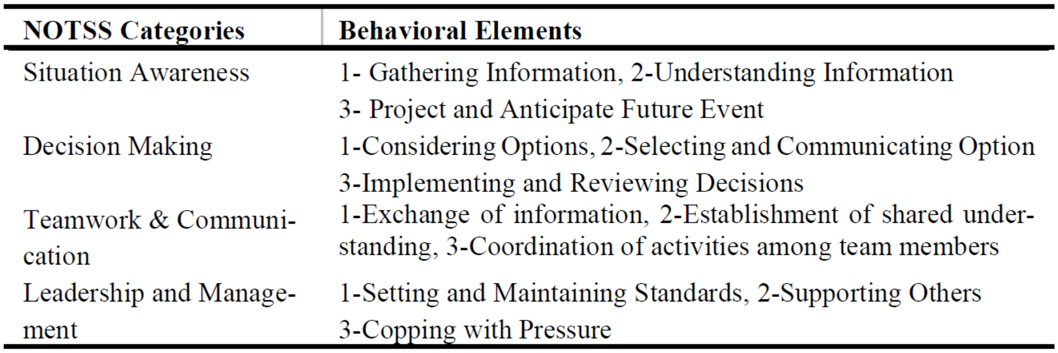


Figure 4. PSD of the 'High Performance' and 'Low Performance' groups during cardiac surgery

Tables.

Categories and Behavioral Elements of Non-Technical Surgical Skills (NOTTS)



NOTSS Categories	Behavioral Elements
Situation Awareness	1- Gathering Information, 2-Understanding Information 3- Project and Anticipate Future Event
Decision Making	1-Considering Options, 2-Selecting and Communicating Option 3-Implementing and Reviewing Decisions
Teamwork & Communication	1-Exchange of information, 2-Establishment of shared understanding, 3-Coordination of activities among team members
Leadership and Management	1-Setting and Maintaining Standards, 2-Supporting Others 3-Copping with Pressure

Subspecialty Medicine and Pathology

Moderators: Philip Edgcumbe, MD PhD and (pending)

Time	Presenting Author	Title
14:00-14:10	Christina Luong (Vancouver General Hospital)*	Validation of machine learning models for estimation of left ventricular ejection fraction on point-of-care ultrasound: Insights on features that impact performance
14:10-14:20	Mitchel Molenaar (Amsterdam UMC)*	Deep learning-based segmentation of coronary arteries in x-ray coronary angiography
14:20-14:30	Sandrine de Ribaupierre (Western University, Canada)*	Robot-Assisted SEEG Electrode Placement for Epilepsy in Pediatric Patients: Workflow Comparison between Frame-Based and Frameless approaches
14:30-14:40	Jacob Jaremko (University of Alberta)*	Feasibility of Ultrasound Screening for Hip Dysplasia in Primary Care Clinics Using AI
14:40-14:50	Willa Yim (IMCB, A*STAR)*	H&E 2.0: deep learning-enabled identification of tumor-specific CD39+CD8+ T cells in marker-free images for predicting immunotherapy response
14:50-15:00	Mai Chan Lau (BII A*STAR)*	HE2.0 web server: an image database supports interactive visualization towards AI-empowered pathology training
15:00-15:10	Discussion/Slush time	

Validation of machine learning models for estimation of left ventricular ejection fraction on point-of-care ultrasound: Insights on features that impact performance

Authors:

Christina L Luong MD, MHSc [a]*, Mohammad H. Jafari MSc, Phd[b]*, Delaram Behnami MASc, PhD[b]*, Yaksh R Shah BSc[c]*, Lynn Straatman MD[a], Nathan Van Woudenberg MASc[b], Leah Christoff NP[a], Nancy Gwadry NP[a], Nathaniel Hawkins MBChB, MD, MPH[a], Eric C. Sayre PhD[d], Darwin Yeung MD[a], Michael Tsang MD[a], Ken Gin MD[a], John Jue MD[a], Parvathy Nair MD, MHPE[a], Purang Abolmaesumi MSc, PhD [b†], and Teresa Tsang MD [a†]

* Joint first authors,
† Joint senior authors

Affiliations:

[a] University of British Columbia, Division of Cardiology, Vancouver, BC, Canada
[b] University of British Columbia, Department of Electrical and Computer Engineering, Vancouver, BC, Canada
[c] University of British Columbia, Faculty of Pharmaceutical Sciences, Vancouver, BC, Canada
[d] Arthritis Research Canada, Vancouver, BC, Canada

Presenting author:

Dr. Christina L. Luong
University of British Columbia; Diamond Health Care Centre 9th Floor Cardiology
2775 Laurel Street, Vancouver, British Columbia; V5Z 1M9
email: christina.luong@ubc.ca

Keywords:

Point-of-care ultrasound, Echocardiography, Heart failure

Key information:

1. Research question:
    a. What is the performance of a machine learning (ML) model developed using comprehensive echocardiogram when applied to point-of-care ultrasound (POCUS) compared with expert interpretation and echo reported left ventricular ejection fraction (LVEF)?
2. Findings:
    a. Of 1257 videos, from 138 subjects, the ML model generated LVEF predictions on 341 videos. We observed a good intraclass correlation (ICC) between the ML model prediction and the reference standards (ICC = 0.77-0.84). Despite good overall correlation, lower ICC was found during atrial fibrillation (ICC 0.60) versus sinus rhythm (ICC 0.83).
3. Meaning:
    a. ML models trained and tested on echocardiogram data for LVEF can be successfully applied to POCUS.

INTRODUCTION:

Machine learning (ML) models can accurately estimate LVEF from echocardiography[1-7], but few studies have validated performance on clinician-driven cardiac point-of-care ultrasound (POCUS)[8]. Cardiac POCUS is a powerful tool that can aid in diagnosing heart disease and guide treatments at the bedside. However, point-of-care ultrasound can be a difficult modality to master, resulting in suboptimal image quality that hampers ML model performance. This study aims to show the feasibility and reliability of ML LVEF estimation on clinician-driven POCUS.

MATERIAL AND METHODS:

We recruited participants from a Heart Failure clinic at an academic referral hospital between February 2021 and June 2022. The study included clinician scanners with variable scanning experience (7 physicians and 2 nurse practitioners). Eligible participants were 18 years or older with an echocardiogram within 3 months. Scanners obtained target views independently. The acquired clips were analyzed offline by the ML model for LVEF estimation and compared with reference standards. The videos were processed for ML model analysis by cropping with an in-house algorithm, downsizing to 128x128 pixels with 30 sampled frames, and rescaling pixel intensities. Successful LV segmentation by the ML model for 30 consecutive frames was required for LVEF estimation; videos of insufficient quality were excluded.

The reference standards for this study were level III echocardiographer interpretation of images and derived LVEF (calculated linear interpolation of LVEF from the subject's formal echo reports). The level III echocardiographer gold standard was established in two ways: (1) LVEF per randomized video file and (2) overall participant LVEF after viewing all videos. ML model performance was evaluated with the intraclass correlation coefficient (ICC) between the ML model LVEF and reference standards. Subgroup analysis was performed for BMI, rhythm, and scanner type.

The ML model in this study was previously developed and validated using 2,920 apical echo cines from 2,127 patients. It is based on U-Net architecture, predicting LV segmentation mask and two landmarks heatmaps (LV apex and mitral valve) to estimate LVEF by Simpson's method of disc. The model analyzes echo cine frame by frame for the entire cardiac cycle, and its architecture and performance on echo data were previously described by Jafari et al[9].

RESULTS:

There were 138 participants scanned, yielding 1257 videos. Participant characteristics are summarized in Table 1 and POCUS data by rater is shown in Tables 2 and 3. Of 1257 POCUS videos, 341 were sufficient for ML model estimation of LVEF. The ICC for ML model and level III echocardiographer LVEF was 0.772 [0.501,1.000] for visual estimates and 0.778 [0.578,1.000] when segmentation was feasible on randomized single videos. There was an increase in ICC when the echocardiographer was able to view all videos for a participant: visual LVEF ICC 0.794 [0.173, 1.000] and 0.843 [0.310, 1.000] with segmentation (Table 4). There was good agreement between the ML model and the derived reported LVEF, ICC 0.798 [0.143, 1.000]. These ICC values indicate a good level of inter-rater agreement between the ML model and several reference standards Figure 1.

The correlations between the ML model and the BMI groups ≥ 30 and < 30 were similar with good inter-rater reliability. Gender impacted ML model LVEF estimation, with higher ICC values for females compared to males. Analysis of AF/AFL and non-AF/AFL rhythms showed variability in ICC values, with better performance for non-AF/AFL rhythms (Table 5).

DISCUSSION AND CONCLUSIONS:

We demonstrate that our ML model exhibits a strong correlation with expert-estimated and echocardiogram-reported LVEF on cardiac POCUS (ICC = 0.77 to 0.84), although performance varied by comorbid conditions. Our findings highlight the potential of ML-augmented cardiac ultrasound while also shedding light on situations where performance may be compromised.

REFERENCES:

1. Asch FM, Poilvert N, Abraham T, et al. Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert. *Circulation: Cardiovascular Imaging*. 2019-09-01 2019;12(9)doi:10.1161/circimaging.119.009303
2. Jafari MH, Woudenberg NV, Luong C, Abolmaesumi P, Tsang T. Deep Bayesian Image Segmentation For A More Robust Ejection Fraction Estimation. IEEE; 2021:
3. Kazemi Esfeh MM, Luong C, Behnami D, Tsang T, Abolmaesumi P. A Deep Bayesian Video Analysis Framework: Towards a More Robust Estimation of Ejection Fraction. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing; 2020:582-590.
4. Behnami D, Luong C, Vaseli H, et al. Automatic cine-based detection of patients at high risk of heart failure with reduced ejection fraction in echocardiograms. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. 2019:1-7. doi:10.1080/21681163.2019.1650398
5. Behnami D, Luong C, Vaseli H, et al. Automatic Detection of Patients with a High Risk of Systolic Cardiac Failure in Echocardiography. Springer International Publishing; 2018:65-73.
6. Zhang J, Gajjala S, Agrawal P, et al. Fully Automated Echocardiogram Interpretation in Clinical Practice. *Circulation*. Oct 2018;138(16):1623-1635. doi:10.1161/CIRCULATIONAHA.118.034338
7. Tromp J, Bauer D, Claggett BL, et al. A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nature Communications*. 2022-11-09 2022;13(1)doi:10.1038/s41467-022-34245-1
8. Crockett D, Kelly C, Brundage J, Jones J, Ockerse P. A Stress Test of Artificial Intelligence: Can Deep Learning Models Trained From Formal Echocardiography Accurately Interpret Point-of-Care Ultrasound? *J Ultrasound Med*. Dec 2022;41(12):3003-3012. doi:10.1002/jum.16007
9. Jafari MH, Girgis H, Van Woudenberg N, et al. Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. *Int J Comput Assist Radiol Surg*. Jun 2019;14(6):1027-1037. doi:10.1007/s11548-019-01954-w

DISCLOSURES: No disclosures

APPENDIX:

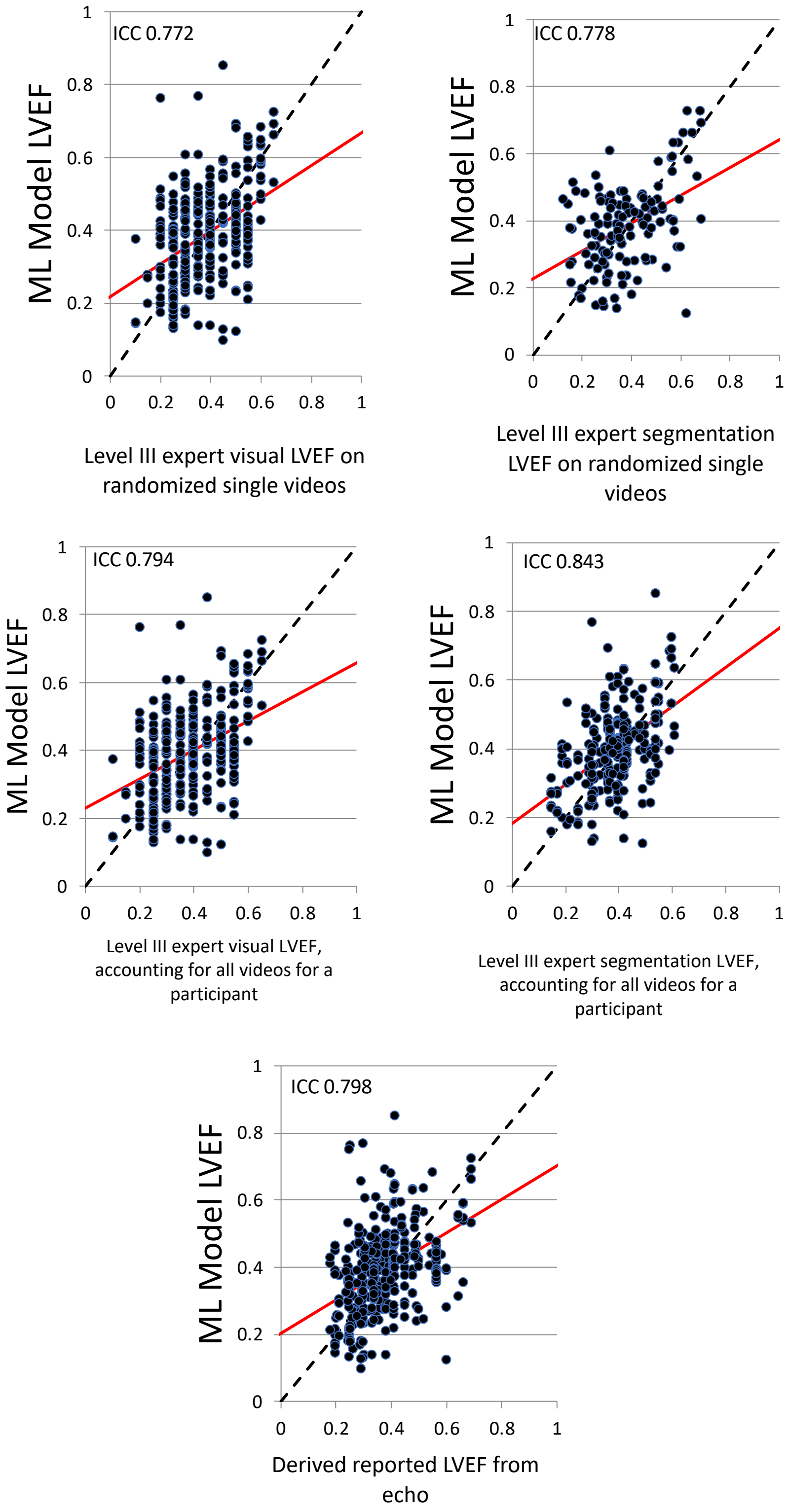
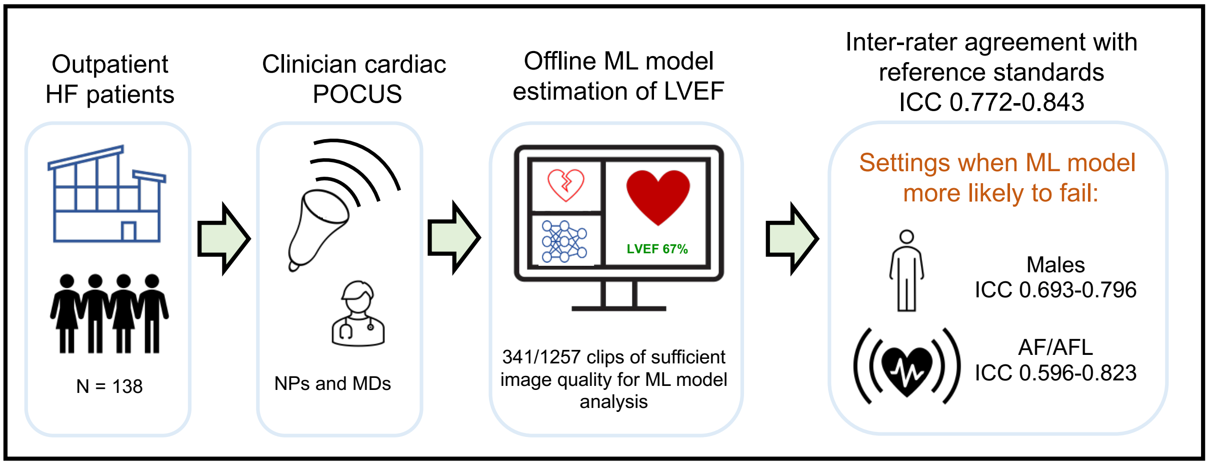


Figure 1: Linear regression plots comparing the ML model to the reference standards. The intraclass correlation coefficient (ICC) for ML model LVEF and level III echocardiographer LVEF was 0.772 [0.501,1.000] and 0.778 [0.578,1.000] for randomized single videos by visual estimate and segmentation, respectively. The ICC for single video ML model LVEF and level III echocardiographer LVEF was 0.794 [0.173, 1.000] for visual assessment and 0.843 [0.310, 1.000] by segmentation when the expert was able to review all clips for a participant. The ICC for ML model LVEF and derived reported LVEF was 0.798 [0.143, 1.000].

Central Illustration: Performance of Machine Learning Model for Left Ventricular Ejection Fraction on Clinician Scanned Point of Care Ultrasound in Heart Failure Clinic.


Outpatient HF patients
N = 138
Clinician cardiac POCUS
NPs and MDs
Offline ML model estimation of LVEF
LVEF 67%
341/1257 clips of sufficient image quality for ML model analysis
Inter-rater agreement with reference standards ICC 0.772-0.843
Settings when ML model more likely to fail:
Males
ICC 0.693-0.796
AF/AFL
ICC 0.596-0.823

AF/AFL = atrial fibrillation/flutter; HF = heart failure; ICC = intraclass correlation; LVEF = left ventricular ejection fraction; ML = machine learning; POCUS = point of care ultrasound;

Table 1: Participant demographic data

Characteristics	**Proportion**
Male	119*/138 (86.2%)
Scanned by nurse	91/138 (65.9%)
Scanned by physician	47/138 (34.1%)
Rhythm atrial fibrillation or atrial flutter at the time of scan	54/138 (39.1%)
LVEF > 50%	27/138 (19.6%)
Type of cardiomyopathy • NICMO • ICMO • Unknown	Type of cardiomyopathy • 73/138 (52.9%) • 53/138 (38.4%) • 12/138 (8.7%)
Variable	**Mean ± SD**
Age (y)	66.2 ± 14.3
Weight (kg)	81.4 ± 18.6
BMI (kg/m^2)	27.0 ± 5.5
Heart rate at time of scan (BPM)	73.9 ± 16.6
Systolic BP (mmHg)	121.7 ± 19.8
Diastolic BP (mmHg)	68.7 ± 10.1

*1 individual identified as a transgender man

Table 2: Single video imaging data split by type of rater

Single video imaging data split by type of rater			
Rater	**Number of videos assigned an LVEF**	**Number of videos of insufficient quality for LVEF estimation**	**Mean estimation of LVEF ± SD**
ML model	341	916	0.39 ± 0.13
Level III expert visual LVEF on randomized single videos	851	406	0.41 ± 0.13
Level III expert segmentation LVEF on randomized single videos	245	1012	0.40 ± 0.14
Level III expert visual LVEF, accounting for all videos for a participant	1175*	82	0.40 ± 0.13
Level III expert segmentation LVEF, accounting for all videos for a participant	754#	503	0.41 ± 0.13
Derived LVEF from echo reports	N/A	N/A	0.39 ± 0.12

* All videos for a participant were included in this category if at least one video in the study was assigned an LVEF by visual assessment

# All videos for a participant were included in this category if at least one video in the study was assigned an LVEF by segmentation

Table 3: Participant imaging data split by type of rater

Rater	Number of studies assigned an LVEF	Number of studies of insufficient quality for LVEF estimation	Mean estimation of LVEF ± SD
ML model	91	47	0.39 ± 0.11
Level III expert visual LVEF on randomized single videos, averaged per patient	120	18	0.40 ± 0.13
Level III expert segmentation LVEF on randomized single videos, averaged per patient	67	71	0.40 ± 0.14
Level III expert visual LVEF, accounting for all videos for a participant	124	14	0.40 ± 0.13
Level III expert segmentation LVEF, accounting for all videos for a participant	72	66	0.40 ± 0.12
Derived LVEF from echo reports	138	0	0.39 ± 0.12

Table 4: Inter-rater agreement for single video data

Observation	Rater 1 of LVEF	Rater 2 of LVEF	ICC (95% CI)
1	ML Model	Level III expert visual LVEF on randomized single videos	0.772 (0.501, 1.000)
2	ML Model	Level III expert segmentation LVEF on randomized single videos	0.778 (0.578, 1.000)
3	ML Model	Level III expert visual LVEF, accounting for all videos for a participant	0.794 (0.173, 1.000)
4	ML model	Level III expert segmentation LVEF, accounting for all videos for a participant	0.843 (0.310, 1.000)
5	ML Model	Derived LVEF from echo reports	0.798 (0.143, 1.000)

Table 5: Inter-rater agreement for single video data, subgroup analyses

Impact of BMI on LVEF estimation: BMI ≥ 30 or BMI <30:

Observation	Rater 1 of LVEF	Rater 2 of LVEF	ICC (95% CI) BMI ≥ 30 (n=80)	ICC (95% CI) BMI < 30 (n=260)
1	ML Model	Level III expert visual LVEF on randomized single videos	0.813 (0.247, 1.000)	0.749 (0.740, 1.000)
2	ML Model	Level III expert segmentation LVEF on randomized single videos	0.829 (0.165, 1.000)	0.709 (0.098, 1.000)
3	ML Model	Level III expert visual LVEF, accounting for all videos for a participant	0.822 (0.129, 0.999)	0.771 (0.551, 1.000)
4	ML model	Level III expert segmentation LVEF, accounting for all videos for a participant	0.909 (0.481, 1.000)	0.802 (0.243, 1.000)
5	ML Model	Derived LVEF from echo reports	0.870 (0.610, 1.000)	0.741 (0.071, 1.000)

Impact of sex on LVEF estimation: Male or female

Observation	Rater 1 of LVEF	Rater 2 of LVEF	ICC (95% CI) Male (n=293)	ICC (95% CI) Female (n=49)
1	ML Model	Level III expert visual LVEF on randomized single videos	0.693 (0.089, 1.000)	0.901 (0.520, 1.000)

2	ML Model	Level III expert segmentation LVEF on randomized single videos	0.705 (0.073, 1.000)	0.869 (0.293, 1.000)
3	ML Model	Level III expert visual LVEF, accounting for all videos for a participant	0.740 (0.067, 0.999)	0.877 (0.503, 1.000)
4	ML model	Level III expert segmentation LVEF, accounting for all videos for a participant	0.796 (0.176, 1.000)	0.901 (0.477, 1.000)
5	ML Model	Derived LVEF from echo reports	0.758 (0.131, 1.000)	0.859 (0.279, 1.000)

Impact of atrial fibrillation (AF) or atrial flutter (AFL) on LVEF estimation:

Observation	**Rater 1 of LVEF**	**Rater 2 of LVEF**	**ICC (95% CI) AF or AFL (n=108)**	**ICC (95% CI) Non-AF or non-AFL (n=234)**
1	ML Model	Level III expert visual LVEF on randomized single videos	0.684 (-0.143, 1.000)	0.809 (0.346, 1.000)
2	ML Model	Level III expert segmentation LVEF on randomized single videos	0.596 (-0.067, 0.999)	0.829 (0.135, 0.999)
3	ML Model	Level III expert visual LVEF, accounting for all videos for a	0.708 (0.182, 1.000)	0.831 (0.210, 1.000)

		participant		
4	ML model	Level III expert segmentation LVEF, accounting for all videos for a participant	0.823 (0.350, 1.000)	0.860 (0.428, 1.000)
5	ML Model	Derived LVEF from echo reports	0.673 (-0.043, 1.000)	0.841 (0.271, 1.000)

Deep learning-based segmentation of coronary arteries in x-ray coronary angiography

Authors:

M.A. Molenaar[1,2], J.L. Selder[1,2], J.O. Bescós[3], M.S. van Mourik[3], Y. Zhao[3], M.J. Schuuring[1,2], B.J. Bouma[1,2], S.A.J. Chamuleau[1,2], C.J. Verouden[1,2]

Affiliations:

[1] Amsterdam UMC, Department of Cardiology, University of Amsterdam, Amsterdam, The Netherlands.

[2] Amsterdam Cardiovascular Sciences, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands.

[3] Image Guided Therapy Systems – Philips, Best, The Netherlands.

Presenting author:

Mitchel Molenaar

Amsterdam UMC, Department of Cardiology, University of Amsterdam, Amsterdam, The Netherlands.

m.a.molenaar1@amsterdamumc.nl

Keywords:

Coronary angiography, image-guided interventions, computer aided diagnosis

Key information:

1. Research question: Visual assessment of stenosis grade in invasive coronary angiography is highly operator-dependent due to vessel foreshortening, vessel overlap and poor image quality by low-dose x-ray radiation. The aim of this study was to evaluate a deep learning algorithm to segment coronary arteries on invasive coronary angiography.
2. Findings: The trained deep learning models demonstrate accurate segmentation of the left circumflex artery, left anterior descending artery and right circumflex artery. The performance of the model to automatically segment the left main artery was poor.
3. Meaning: Deep learning enables accurate segmentation of coronary arteries in invasive coronary angiography, which is a crucial step towards the development of automated methods for stenosis detection and quantification.

MANUSCRIPT *(included in the word count)*

Introduction

Invasive coronary angiography (ICA) is the gold standard to diagnose coronary artery disease (CAD)[1]. Visual assessment of stenosis grade in ICA is highly operator-dependent due to vessel foreshortening, vessel overlap and poor image quality by low-dose x-ray radiation[2,3]. Deep learning may assist in stenosis assessment. To enable stenosis assessment first robust coronary artery detection is needed as a prerequisite. Therefore, the aim of this study was to evaluate a deep learning algorithm to segment coronary arteries on ICA.

Material and methods

ICA studies of patients who underwent ICA or percutaneous coronary intervention in a tertiary center between 2015-2017 were retrospectively collected. ICA cine runs were manually selected for each study in a way that all the major coronaries were clearly visible with minimum overlap in one of the cines runs and stenosis grade could be assessed (stenosis degree >50%). If the patient did not have any significant stenosis, ICA cine runs that would suffice to assess the coronary anatomy were selected. One ICA cine frame was manually selected per run with vessels filled with contrast agent and preferably in end-diastolic phase. In addition, one ICA cine frame was selected prior to administration of the contrast agent. The contours of the left circumflex artery (LCx), left anterior descending artery (LAD), right circumflex artery (RCA), and left main (LM) artery were manually segmented using dedicated software. For each coronary artery, the nnU-Net framework was employed to train a segmentation model. Performance was evaluated on an unseen test set (20% of images) by visual inspection and the dice similarity coefficient (DSC), a metric of segmentation performance between 0 (poor) and 1 (excellent)[4].

Results

A total of 3404 images obtained from 1148 patients were manually segmented (Table 1). Of these images, 2723 images (918 patients) were used to train the segmentation models. Testing of these models on 681 images (230 patients) showed accurate segmentation for the LCx, LAD and RCA, with a median DSC of 0.82, 0.91 and 0.94, respectively (Figure 1, Table 1). The LM segmentation model had a median DSC of 0.04. Upon visual inspection, the LM segmentation model incorrectly classified the proximal LAD and proximal LCx as the LM.

Discussion and Conclusion

Deep learning models to segment coronary arteries in x-ray coronary angiography demonstrated accurate identification of the LCX, LAD and RCA. However, the performance of automated LM segmentation was poor. These results suggest that deep learning can support clinicians in the diagnosis of CAD, and furthermore, it may extract relevant features that can contribute to risk stratification[5]. Further efforts are needed to externally validate the results, evaluate the accuracy of these models around stenosis, and automatically detect separate coronary segments, stenosis and stenosis grade.

References

1. Knuuti J, Wijns W, Saraste A, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. *European Heart Journal*. 2020;41(3):407-477. doi:10.1093/eurheartj/ehz425

2. I CA, F CS, Ms AG. A Novel Multiscale Gaussian-Matched Filter Using Neural Networks for the Segmentation of X-Ray Coronary Angiograms. Journal of healthcare engineering. doi:10.1155/2018/5812059

3. Kobayashi T, Hirshfeld JW. Radiation Exposure in Cardiac Catheterization. *Circulation: Cardiovascular Interventions*. 2017;10(8):e005689. doi:10.1161/CIRCINTERVENTIONS.117.005689

4. Zou KH, Warfield SK, Bharatha A, et al. Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Acad Radiol*. 2004;11(2):178-189. doi:10.1016/S1076-6332(03)00671-8

5. Molenaar MA, Selder JL, Nicolas J, et al. Current State and Future Perspectives of Artificial Intelligence for Automated Coronary Angiography Imaging Analysis in Patients with Ischemic Heart Disease. *Curr Cardiol Rep*. 2022;24(4):365-376. doi:10.1007/s11886-022-01655-y

Disclosures

None

APPENDIX

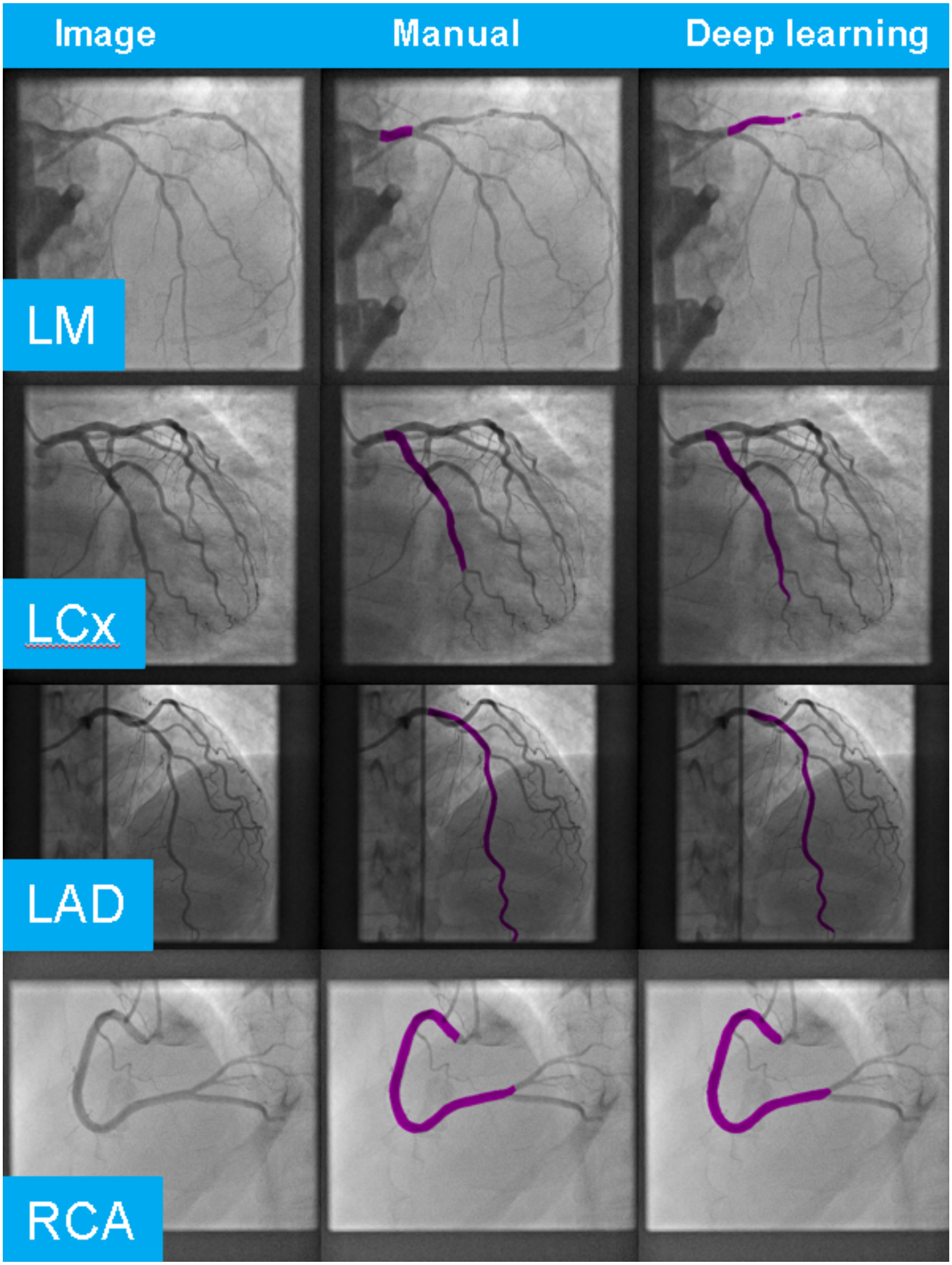
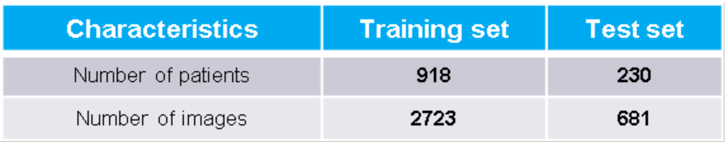


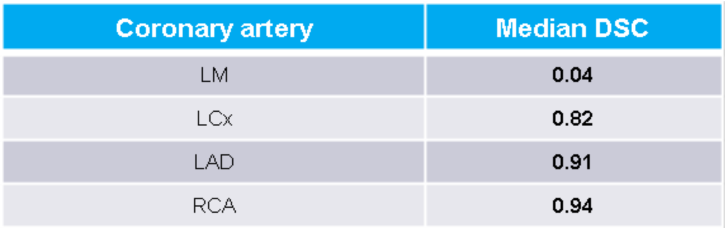
Figure 1: Representative examples of coronary artery segmentation: manual versus deep learning

Table1:Number of patients and images included in study.



Characteristics	Training set	Test set
Number of patients	**918**	**230**
Number of images	**2723**	**681**

Table 2: Performance of segmentation models.



Coronary artery	Median DSC
LM	**0.04**
LCx	**0.82**
LAD	**0.91**
RCA	**0.94**

DSC = dice similarity coefficient; LAD = left anterior descending artery; LCx =left circumflex artery; LM = left main artery; RCA = right circumflex artery.

Robot-Assisted SEEG Electrode Placement for Epilepsy in Pediatric Patients: Workflow Comparison between Frame-Based and Frameless approaches.

Authors:

Sandrine de Ribaupierre[1], Juan S. Bottan[1,2], Greydon Gilmore[1], Jonathan C. Lau[1] & Roy Eagleson[3].

Affiliations:

[1]Department of Clinical Neurological Sciences, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada.

[2]Division Neurosurgery, Hospital General de Niños "Pedro de Elizalde", Buenos Aires, Argentina.

[3]Faculty of Engineering, Western University, London, ON, Canada.

Presenting author:

Sandrine de Ribaupierre, MD, MSc, FRCSC

(sderibau@uwo.ca)

Keywords:

Stereo-electroencephalography, Epilepsy surgery, Stereotactic Robot-assisted surgery, pediatric epilepsy.

Key information:

1. Research question: *Is a simplified paediatric frame-less procedural workflow comparable to a traditional frame-based method in robot-assisted implantation of depth electrodes in children with drug-resistant epilepsy?*
2. Findings: *Although accuracy is less, the frame-less paediatric procedure seems to be comparable to the adult frame-based procedure in terms of safety and accuracy, simplifying the workflow and reducing radiation exposure.*
3. Meaning: *Frame-less procedures in robot-assisted implantation of depth electrodes may be a valid alternative in children with DRE.*

MANUSCRIPT *(included in the word count)*

Introduction

Robot-assisted depth electrode implantation is becoming the preferred method for invasive investigations in drug resistant epilepsy (DRE) due to its safety, accuracy and reducing error in addition to improved efficiency, and to reduce Operative Room time usage. Robot-assisted procedures in children require special technical considerations, such as avoiding the use of a stereotactic frame, and reducing the exposure to neuroimaging radiation. Several groups have published their techniques and accuracy results to establish this procedure. The advantages and disadvantages of implementing a frame-less procedure, specially tailored for children, have not yet been fully explored. Our primary objective is to compare our traditional frame-based procedure performed in adults with Neuromate® with a new frame-less procedure performed in children using the ROSA® stereotactic robot arm to assess its accuracy and associated complication rates. Our secondary objective is to describe a methodology for objective performance analysis applied to this specific scenario and in addition, to provide Surgeons' and Clinicians' perspectives on the usage of both devices.

Material and methods

We retrospectively reviewed a historic cohort of 145 adult frame-based robot-assisted procedures (Neuromate) and our pilot series of the first 10 frame-less robot-assisted procedures (ROSA) in children with DRE. Surgical timing, usage of radiation, workflow analysis, accuracy results (Entry point and target radial error), technical issues and complications (hemorrhage, infections, deaths) were collected. Additionally, we conducted a survey among Neurosurgeons who are skilled on both systems, to assess their experiences from a user perspective (hardware, software and technical support).

Results

A total of 1105 electrodes where implanted in the adult group vs. 96 in the pilot pediatric group. Extratemporal electrodes accounted for 52.7% in the adult series and 77.1% in children. Mean entry point radial error for the adult frame-based workflow was 1.17 mm (95% CI: 1.07-1.27), whereas for ROSA's frame-less procedure was 1.58 mm (95% CI: 1.42-1.75). Target radial error was 1.57 mm (1.32-1.81) vs. 1.74 mm (1.54-1.95) respectively. Target accuracy was higher in the frontal lobe for ROSA. In the adult workflow, there was one death, 0% infection rate and 2 cases of mild-moderate intracranial hemorrhage. Only one electrode in the pediatric series deviated, and no complications where registered. As expected, oblique and longer intracranial trajectories for both workflows showed greater error compared to shorter and/or orthogonal trajectories. Frame-less workflow resulted in lower overall radiation delivery by avoiding intraoperative fluoroscopy. The user questionnaire suggests a preference for the ROSA from a hardware usage perspective, while the Neuromate was preferred from a software usability perspective.

Discussion and Conclusion

Although limited series, both techniques seem to have comparable profile in terms of safety and accuracy. The accuracy of the frameless method on the ROSA was inferior to that of Neuromate, especially for parietal and occipital trajectories. Preliminary results suggest our frame-less robotic workflow applied in children is safe and involves less use of radiation. Larger samples will be needed to establish these findings, and this is a report on an on-going study.

Disclosures

None of the authors have any disclosures to report.

APPENDIX

Table 1: Mean error with 95% Confidence intervals. Deviated electrodes where not included. Eu= Eucledian distance, R= radial error; EP: Entry Point; T= Target.

Adult	**Eu EP**	**R EP**	**Eu T**	**R T**
N = 1105	**1,87**	**1,17**	**2,07**	**1,57**
	1,66	1,07	1,85	1,32
	2,09	1,27	2,29	1,81
Rosa				
N = 96	**1,75**	**1,58**	**2,17**	**1,74**
	1,56	1,42	1,94	1,54
	1,94	1,75	2,41	1,95

Feasibility of Ultrasound Screening for Hip Dysplasia in Primary Care Clinics Using AI

Authors:

Jacob L. Jaremko[1] , Ehsan Seyed Bolouri[2] , Rod Fitzsimmons Frey[3] , Sukhdeep Dulai[4] & Allan L. Bailey[5], Abhilash Hareendranathan[1]

Affiliations:

1. Department of Radiology & Diagnostic Imaging, University of Alberta, Edmonton, Canada.

2. Exo Inc, Santa Clara, USA.

3. Sonance AI Inc, Edmonton, Canada.

4. Department of Surgery, University of Alberta, Edmonton, Canada.

5. Department of Family Medicine, University of Alberta, Edmonton, Canada.

Presenting author:

Jacob L. Jaremko, Department of Radiology & Diagnostic Imaging, University of Alberta, Edmonton

jjaremko@ualberta.ca

Keywords:

hip dysplasia, ultrasound, point-of-care, artificial intelligence, screening.

Key information:

1. Research question: Is hip dysplasia screening by lightly trained users feasible in a primary care network (PCN) setting?
2. Findings: Our AI-aided workflow can be used by lightly trained users in a PCN setting. It detects cases of hip dysplasia at the expected rate (1.6% of babies), of which half would have been missed in the current care pathway, with specificity 99% and positive predictive value 61%.
3. Meaning: Our results make a strong case for universal screening of hip dysplasia in PCNs by lightly trained users such as nurses. Adding AI provides these users the much-needed confidence to perform a hip examination which can eventually lead to early and effective treatment of hip dysplasia.

MANUSCRIPT (included in the word count)

Introduction

Developmental dysplasia of the hip (DDH) is a common cause of premature osteoarthritis in young adults[1]. DDH incidence averages 1-2% [1,2], up to 30x higher among Indigenous populations. Early screening is crucial to allow non-surgical treatment and improve outcomes since missed DDH necessitates multiple surgeries. Currently, DDH screening is done in most countries based on risk factors (eg., breech birth, family history). We have previously shown that the high false-positive rates and user dependency that limit wide use of hip ultrasound can be reduced by using partially automated [3,4] or fully automated image analysis[5,6]. Novice users learn more easily to acquire ultrasound videos than single 2D frames[7]. Providing real time feedback on image quality could further improve diagnostic accuracy as inadequate images can be flagged up-front[8,9]. This study examines the feasibility of an AI-augmented workflow with image quality feedback and diagnostic suggestions for universal DDH screening in primary care networks (PCNs).

Material and methods

Ultrasound videos were analysed in real-time using the FDA-cleared MEDO hip app (Exo Inc.). The hip app uses a Convolutional Neural Network (CNN) model similar to U-Net to segment the acetabulum and femoral head, assesses image quality based on the number of frames with necessary imaging landmarks, and suggests a diagnosis for physician review.

With ethics board approval we performed hip ultrasound in three Alberta PCN clinics using a low-cost handheld ultrasound (Philips Lumify) connected to an Android tablet. Consecutive infants aged 4-16 weeks presenting for already-scheduled infant wellness check visits were scanned by registered nurses or physicians at the clinic. Based on AI recommendation, infants with possible DDH from AI screening were sent for follow-up first internally at the same centre 1-2 weeks later, and if still suspicious for DDH, externally to a tertiary hospital orthopedic clinic for gold-standard assessment.

Results

Of 697 patients scanned (2 hips each), we had 57 scan failures (8%) where no scan could be obtained due to uncooperative infant, software issues or clinic logistic issues. Of the 92% of eligible infants successfully scanned, 591 had normal or typically developing hips (Figure 1a). The AI tool detected 37 cases with suspected DDH (Figure 1b) out of which 19 did not persist on internal follow-up. The remaining 18 cases were sent for external follow-up: 7 resolved and 11 were confirmed as dysplastic by the consulting radiologist (Table 1), a rate of 1.6% of eligible infants. Of the DDH cases, 5 had no associated risk factors. External referral specificity for DDH was 0.99 and positive predictive value 61%. Sensitivity could not be calculated since it was not feasible to perform gold-standard conventional ultrasound in all infants.

Discussion and Conclusion

This study conducted in three clinics in Alberta validates the use of AI-augmented hip ultrasound to screen for hip dysplasia in primary care settings. The tool provides feedback to the user when enough adequate-quality frames have been acquired. With this feature, nurses at the PCN were able to confidently perform the hip examination with minimal training, as low-quality images were flagged up-front. In this AI-aided workflow, we successfully scanned 92% of eligible infants (a rate which can be improved further with software refinements), detected the expected rate of DDH (1.6%), and picked up 5 babies with DDH that would have been missed in the current care pathway. We had specificity 99% and positive predictive value 61% vs. gold-standard referral. These results highlight the feasibility of implementing AI-augmented point-of-care hip ultrasound screening in infants by primary-care clinicians. Further studies are needed to establish scan sensitivity and expand to larger populations.

References

1. Bache CE, Clegg J, Herron M. Risk factors for developmental dysplasia of the hip: ultrasonographic findings in the neonatal period. J Pediatr Orthop B. 2002;11(3):212-218.

2. Furnes O, Lie SA, Espehaug B, Vollset SE, Engesaeter LB, Havelin LI. Hip disease and the prognosis of total hip replacements: a review of 53 698 primary total hip replacements reported to the Norwegian arthroplasty register 1987--99. J Bone Joint Surg Br. 2001;83(4):579-579.

3. Hareendranathan AR, Mabee M, Punithakumar K, Noga M, Jaremko JL. A technique for semiautomatic segmentation of echogenic structures in 3D ultrasound, applied to infant hip dysplasia. Int J Comput Assist Radiol Surg. 2016;11(1):31-42.

4. Zonoobi D, Hareendranathan A, Mostofi E, et al. Developmental Hip Dysplasia Diagnosis at Three-dimensional US: A Multicenter Study. Radiology. 2018;287(3):1003-1015.

5. Hareendranathan AR, Zonoobi D, Mabee M, et al. Toward automatic diagnosis of hip dysplasia from 2D ultrasound. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). ; 2017:982-985.

6. Ghasseminia S, Seyed Bolouri SE, Dulai S, et al. Automated diagnosis of hip dysplasia from 3D ultrasound using artificial intelligence: A two-center multi-year study. Informatics in Medicine Unlocked. 2022;33:101082.

7. Mostofi E, Chahal B, Zonoobi D, et al. Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. Eur Radiol. 2019;29(3):1489-1495.

8. Hareendrananthan AR, Mabee M, Chahal BS, Dulai SK, Jaremko JL. Can AI automatically assess scan quality of hip ultrasound? Appl Sci. 2022;12(8):4072.

9. Hareendranathan AR, Chahal BS, Zonoobi D, Sukhdeep D, Jaremko JL. Artificial Intelligence to Automatically Assess Scan Quality in Hip Ultrasound. Indian J Orthop. July 2021. doi:10.1007/s43465-021-00455-w

Disclosures

Dr. Jacob L. Jaremko is a co-founder of MEDO.ai, which develops ultrasound imaging AI products including the app evaluated in this manuscript. He is also a partner of Medical Imaging Consultants and receives fees for interpreting medical imaging including hip ultrasound. Seyed Ehsan Bolouri and Rod Frey were employees of MEDO.ai at the time of this study. Abhliash Hareendranathan was an employee with MEDO from 2018 to 2020. We have not received any funding from MEDO.ai for this study. The rest of the authors have no conflicts to declare.

APPENDIX

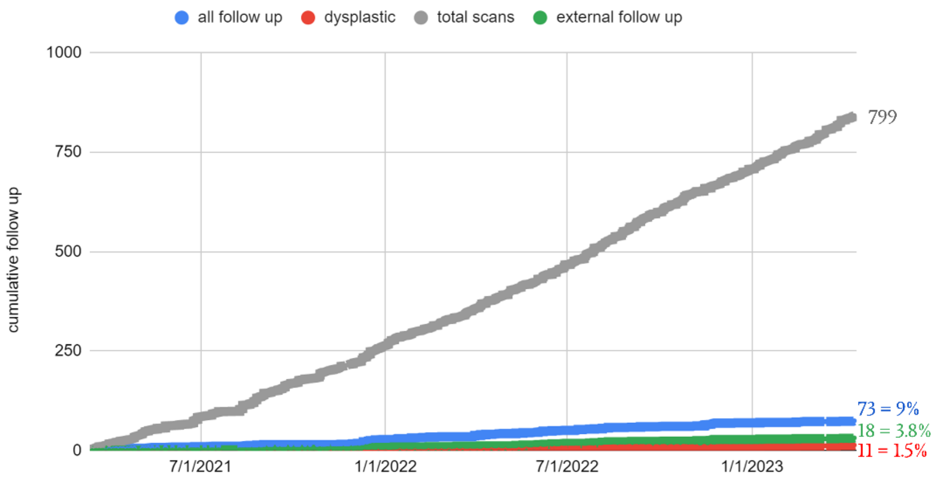


Figure 1: Results from scanning 697 patients (799 total scans) in 3 different clinics (with ~6 sonographers) around Alberta over 2 years.

Table 1: Hip ultrasound/AI hip dysplasia screening

Result	N	%	Description
Scan failure	57	8%	Could not obtain scan
Normal at first screen	591	85%	
Normal at internal FU	19	2.7%	Hips normal when re-scanned 1-2 weeks later in same clinic.
Normal at external FU	7	1%	i.e., false-positives from the screening program
Dysplastic at external FU	11	1.6%	Of which 5 had no risk factors for DDH and would not have been picked up otherwise.

Notes: (1) percentages are as a proportion of infants eligible for screening. (2) For screening including up to 1 internal follow-up scan, at external referral (total 18 cases) the specificity was 99%, and positive predictive value 61%.

H&E 2.0: deep learning-enabled identification of tumor-specific CD39$^+$CD8$^+$ T cells in marker-free images for predicting immunotherapy response

Authors:

Willa Wen-You Yim[1*], Felicia Wee[1*], Jia Meng[1*], Jeffrey Chun Tatt Lim[1], Craig Ryan Joseph[1], Xinru Lim[1], Kai Soon Ng[1], Jiang Feng Ye[1], Zhen Wei Neo[1], Li Yen Chong[1], Chan Way Ng[2], Tony Kiat Hon Lim[3†], Mai Chan Lau[2,4†], Joe Yeong[1,3†]

*Equal contribution

†Co-corresponding

Affiliations:

1. Institute of Molecular Cell Biology (IMCB), Agency for Science, Technology and Research (A*STAR), Singapore
2. Singapore Immunology Network, Agency for Science, Technology and Research (A*STAR), Singapore
3. Department of Anatomical Pathology, Division of Pathology, Singapore General Hospital, Singapore
4. Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore

Presenting author:

Full name: Willa Wen-You Yim

Affiliation: Institute of Molecular Cell Biology (IMCB), Agency for Science, Technology and Research (A*STAR), Singapore

Email: willa_yim@imcb.a-star.edu.sg

Keywords:

Cancer, checkpoint blockade immunotherapy, deep learning, immunofluorescence, tumor-specific T cells

Key information:

1. Research question:

   Can deep learning (DL) models be used to identify CD39$^+$CD8$^+$ T cells in H&E images of cancer tissues and brightfield images of blood cells?

2. Findings:

   A DL model was developed for H&E images of colorectal carcinoma (CRC) samples and another for peripheral blood mononuclear cells (PBMCs) from CRC mouse models. The F1 scores of the models are 0.83 and 0.80, respectively.

3. Meaning:

   Preliminary results indicate DL can identify CD39$^+$CD8$^+$ T cells, which suggests the presence of characteristic morphological features. The models should be further refined and tested on more samples before being used to predict PD-1 and PD-L1 blockade immunotherapy response.

MANUSCRIPT

Introduction

Accumulating evidence implicates CD39 as a tumor-specific CD8$^+$ T cell marker. Our group showed that CD8$^+$ T cells without CD39 expression are bystander tumor infiltrating leukocytes in colorectal carcinoma (CRC) and non-small cell lung cancer (NSCLC)[1] and that CD39-expressing CD8$^+$ T cells function as tumor antigen-specific CD8$^+$ T cells in treatment-naïve NSCLC[2] and triple-negative breast cancer (TNBC)[3]. These findings have been confirmed by other groups[4-6]. Therefore, the combination of CD39$^+$CD8$^+$ T cell abundance and spatial localization is a potential predictor of patient response to PD-1 and PD-L1 blockade immunotherapy in multiple types of cancers[3-6]. However, multi-marker assays for identifying specific immune phenotypes, including CD39$^+$CD8$^+$ T cells, in tissue or blood samples are laborious and costly, preventing high throughput implementation on patient samples. To overcome these issues, we sought to develop deep learning (DL) models that have been trained with data from multi-marker assays, namely multiplex immunofluorescence and imaging flow cytometry, to identify CD39$^+$CD8$^+$ T cells based on morphology in hematoxylin and eosin (H&E)-stained tissue images and brightfield images of immune cells from blood samples.

Material and methods

We developed DL pipelines to identify CD39$^+$CD8$^+$ T cells in CRC samples and peripheral blood mononuclear cells (PBMCs) from CT26 tumor-bearing mice (CRC mouse tumor models).

CD39$^+$CD8$^+$ T cells in CRC tissue samples were visualized with multiplex immunofluorescence. The samples were subsequently washed and stained with H&E. The mouse PBMCs were immunostained with fluorescent antibodies and visualized with imaging flow cytometry (Fig. 1A).

The H&E DL pipeline stages for the CRC samples are: (1) aligning fluorescence images with the H&E image, (2) cell segmentation, (3) manual identification of CD39$^+$CD8$^+$ cells that serve as ground truth labels, (4) extracting each cell as a small image patch with the cell in the center, and (5) training a DL model for CD39$^+$CD8$^+$ prediction (Fig. 1A). The current model ($\theta_{\text{H\&E}}$) was trained with 2,426 positive examples and 101,084 negative examples.

The DL pipeline stages for the mouse PBMCs are: (1) gating CD8$^+$ and CD39$^+$ positivity based on fluorescence intensity, and (2) train a DL model for CD39$^+$CD8$^+$ prediction (Fig. 1B). The current model (θ_{blood}) was trained with 1,985 positive examples and 4,639 negative examples.

Both DL models are convolutional neural networks with residual blocks. Data augmentation such as random flips, rotations, and brightness adjustments, were implemented. The models were evaluated with F1 scores.

Results

The current version of $\theta_{\text{H\&E}}$ has a test F1-score of 0.83, θ_{blood} has a test F1-score of 0.80.

Discussion and Conclusion

Both models are able to identify CD39$^+$CD8$^+$ T cells from marker-free H&E images and brightfield images, respectively, which indicates that there are morphological characteristics unique to these tumor-specific T cells. However, the models should be further improved, especially in terms of generalization. Ongoing work include testing the models on more cancer types and validating them with independent cohorts. These models would eventually be applied to pre-treatment patient samples of known outcomes from different cancers to evaluate their predictive capabilities for PD-1 and PD-L1 blockade immunotherapy response.

When fully realized, cell identification by virtual staining of H&E images ('H&E 2.0') and brightfield images of blood samples would be an inexpensive and easy-to-implement diagnostic pathology tool that clinicians and researchers can use to screen large numbers of clinical samples. Any samples of interest could then be validated with confirmatory analyses involving the more labor-intensive and costly methods of multiplex immunofluorescence and/or imaging flow cytometry (Fig. 1C). Hence, DL-enabled virtual staining would help attenuate the rapidly increasing amount of medical resources being spent on cancer care to bring immunotherapy to more patients in the research setting and potentially in the clinic.

References

1. Simoni Y, Becht E, Fehlings M, et al. Bystander CD8(+) T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*. 2018;557(7706):575-579.
2. Yeong J, Suteja L, Simoni Y, et al. Intratumoral CD39(+)CD8(+) T Cells Predict Response to Programmed Cell Death Protein-1 or Programmed Death Ligand-1 Blockade in Patients With NSCLC. *J Thorac Oncol*. 2021;16(8):1349-1358.
3. Jia M, Tira T, Jiangfeng Y, et al. 1039 The prognostic value of tumour-specific T cells in Asian TNBC: using CD39(+)CD8(+) T cells as a surrogate marker. *J Immunother Cancer*. 2022;10(Suppl 2):A1081
4. Laumont CM, Wouters MCA, Smazynski J, et al. Single-cell Profiles and Prognostic Impact of Tumor-Infiltrating Lymphocytes Coexpressing CD39, CD103, and PD-1 in Ovarian Cancer. *Clin Cancer Res*. 2021;27(14):4089-4100.
5. Lee YJ, Kim JY, Jeon SH, et al. CD39(+) tissue-resident memory CD8(+) T cells with a clonal overlap across compartments mediate antitumor immunity in breast cancer. *Sci Immunol*. 2022;7(74):eabn8390.
6. Attrill GH, Owen CN, Ahmed T, et al. Higher proportions of CD39+ tumor-resident cytotoxic T cells predict recurrence-free survival in patients with stage III melanoma treated with adjuvant immunotherapy. *J Immunother Cancer*. 2022;10(6):e004771.

Disclosures

None.

APPENDIX

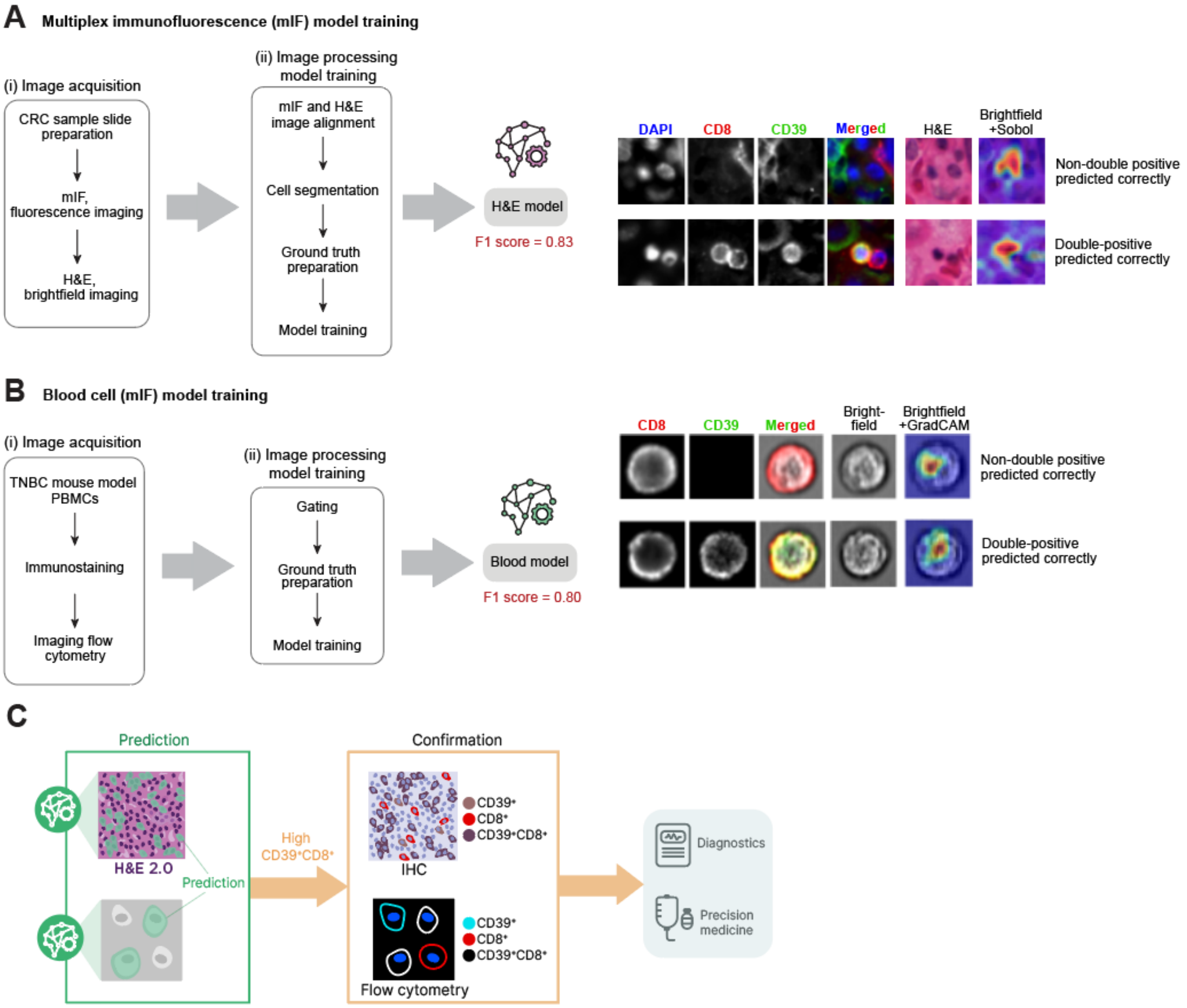


Figure 1: Two deep learning models for identifying tumor-specific CD39$^+$CD8$^+$ T cells (double-positive cells) in H&E images and single blood cell images. (**A**) Colorectal carcinoma (CRC) sample sections were visualized for CD8 and CD39 expression by multiplex immunofluorescence and subsequently for morphology by H&E. Individual cells were obtained from H&E images and marked for CD39-CD8 positivity based on immunofluorescence results. These ground truth images were then used to train the DL model. The current version has an F1 score of 0.83, indicating that it can distinguish double-positive cells from those that are not, which are shown with representative images. (**B**) Peripheral blood mononuclear cells (PBMCs) isolated from a CRC mouse model were immunostained and visualized with imaging flow cytometry. Double-positive cells were identified based on fluorescence intensity of CD8 and CD39. The model is trained on these ground truth images and has an F1 score of 0.80, indicating that it can distinguish double-positive cells from those that are not, which are shown with representative images. (**C**) DL models that reliably predict CD39+CD8+ cells can be used to screen large numbers of patient samples before expensive and time-consuming confirmatory analyses like immunohistochemistry and imaging flow cytometry. This will alleviate some pressure on medical resources in the immunotherapy era.

HE2.0 web server: an image database supports interactive visualization towards AI-empowered pathology training

Authors:

Joe Poh Sheng Yeong[1,2*], Minh N. Nguyen[3*], Willa Yim[1*], Felicia Wee[1], Marcia Zhang[3,4], Xinyun Feng[3,4], Menaka Priyadharsani Rajapakse[3,5], Jeffrey Chun Tatt Lim[1], Chandra Verma[3,6#], and Mai Chan Lau[3,5#]

Affiliations:

[1]Institute of Molecular and Cell Biology (IMCB), Agency of Science, Technology and Research (A*STAR), Singapore 138673.
[2]Department of Anatomical Pathology, Singapore General Hospital, Singapore 169856.
[3]Bioinformatics Institute (BII), Agency of Science, Technology and Research (A*STAR), Singapore 138671.
[4]National University of Singapore, Singapore 639798.
[5]Singapore Immunology Network (SIgN), Agency of Science, Technology and Research (A*STAR), Singapore 138668.
[6]Department of Biological Sciences, National University of Singapore, Singapore
These authors contributed equally to this work and share first authorship

[#]*Co-corresponding authors*

Presenting author:

Full name: Joe Yeong
Affiliation: IMCB, A*STAR
Email: yeongps@imcb.a-star.edu.sg

Keywords:

Haematoxylin and eosin, interactive visualization, web server, spatial omics

Key information:

1. Research question: To provide an interactive and integrated visualization of spatial features, such as multiple cell phenotypes, on clinically available histomorphological H&E images.
2. Findings: Visual appreciation of specific lymphocyte subsets detected by hyperplexed immunofluorescence overlaid on the morphologically identified lymphocytes in the H&E image space.
3. Meaning: We envision that H&E2.0 web server can serve as a data repository for researchers and clinicians to visualize various biological information collected through manual annotation, tissue-based multi marker assays, advanced spatial techniques, or deep learning prediction models. Potentially, it can provide an effective solution for integrative multi-omics spatial analysis, as well as pathology training.

MANUSCRIPT

Introduction

Growing evidence shows that, besides the abundance of immune cells, their spatial relationship in the tumor microenvironment (TME) play a critical role in anti-tumor immunity and patient survival (1, 2). Recent advancement of high-plex molecular profiling further facilitates the spatial interrogation of complex TME (3, 4), showing great potential for discovering novel spatial biomarkers. However, the high-plex techniques are expensive and often inaccessible. On the other hand, H&E-based deep learning (DL) model has gained great success in the past decade for diagnostic and prognostic utilities including cancer subtype classification (5) and prediction for genetic alterations (5). Such success substantiates the theory that histomorphological features in routine H&E images contain biological signal predictive of clinically actionable information, leading to the research initiative termed H&E 2.0 (6) for extended application of H&E digitized images via DL. Such efforts have been extended successfully for virtual staining studies for instance in-situ CD3$^+$ T-cell in non-small cell lung cancer(7), and gene expression of 28 different cancer types (8). To bring the advantage of H&E 2.0 to clinic, particularly for achieving pathology consensus or for effective pathology training, visual appreciation of various spatial features in H&E image space can give a great boost. Here, we present such a platform in a web server that facilitates integration of, theoretically, an unlimited number of spatial features/omics data in the same H&E space (Figure 1).

Material and methods

The H&E-stained and hyperplex immunofluorescence (IF) images of a colorectal cancer tissue section were first preprocessed (background noise removal, tissue contour identification and binary mask generation) for tissue region extraction. The processed image pair were then registered with the imreg_dft 2.0.0 package. Cells in H&E images were segmented using the *StarDist* extension on *QuPath* v0.4.3; while cells in the paired IF image were segmented using the Mesmer model from the *DeepCell* library in a Docker container from *steinbock* v0.16.1 toolkit. React which is a library for web and user interfaces was used to build the front end of H&E2.0 web server. The visualization functions of the web server were implemented with PlotyJS. Redux, an open-source JavaScript library, was used to manage the front-end data. The HE2.0 web server is freely accessible at https://mspc.bii.a-star.edu.sg/minhn/he2.html.

Results

In the demo image of a colorectal cancer tissue (Figure 2), we identified all the cells from H&E image which was overlaid with the individual cell type markers quantified by multiplexed immunofluorescence on the same tissue section. In this CRC tissue, we identified 297,000 cells of which 83,500 are tumour epithelial cells, 5900 are proliferative tumour epithelial cells, 24,500 are CD8$^+$ cytotoxic T-cells, 2400 are CD4$^+$ helper T-cells and 600 are CD68$^+$ macrophages. Interactive visualization of selected marker(s) on the H&E image reveals that there are lymphocytic aggregates near small tumour islands within the TME. This is confirmed with the IF mask of the interactive slider, where CD8$^+$ and CD4$^+$ T-cells are present in these lymphocytic aggregates.

Discussion and Conclusion

We propose an online database with interactive visualization capability, aiming to extend the clinical usability of routinely collected H&E digitized images for multi-scale spatial analysis and pathology training. With AI-enabled image processing including cell identifying and phenotyping as well as image registration, we can integrate additional spatial information, that is obtained via various spatial techniques, onto the H&E space. To unlock the full potential, DL-based virtual staining models can be incorporated to the web server to enable prediction of various spatial omics data from the query H&E images, offering a feasible solution for multi-omics spatial analysis as well as to bring about innovative pathology learning.

References

1. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet. 2018;391(10135):2128-39.
2. Väyrynen JP, Haruki K, Lau MC, Väyrynen SA, Ugai T, Akimoto N, et al. Spatial Organization and Prognostic Significance of NK and NKT-like Cells via Multimarker Analysis of the Colorectal Cancer Microenvironment. Cancer Immunol Res. 2022;10(2):215-27.
3. Bosisio FM, Van Herck Y, Messiaen J, Bolognesi MM, Marcelis L, Van Haele M, et al. Next-generation pathology using multiplexed immunohistochemistry: mapping tissue architecture at single-cell level. Frontiers in oncology. 2022;12.
4. Yeong J, Lim JCT, Lee B, Li H, Ong CCH, Thike AA, et al. Prognostic value of CD8+ PD-1+ immune infiltrates and PDCD1 gene expression in triple negative breast cancer. Journal for immunotherapy of cancer. 2019;7(1):1-13.
5. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature medicine. 2018;24(10):1559-67.
6. Comiter C, Vaishnav ED, Ciampricotti M, Li B, Yang Y, Rodig SJ, et al. Inference of single cell profiles from histology stains with the Single-Cell omics from Histology Analysis Framework (SCHAF). bioRxiv. 2023:2023.03.21.533680.
7. Abu Bakr A, Yu Qing C, Matthew Leong Tze K, Denise G, Jeffrey Chun Tatt L, Mai Chan L, et al. 818 Using deep learning approaches with mIF images to enhance T cell identification for tumor - automation of infiltrating lymphocytes (TILs) scoring on H&amp;amp;E images. Journal for ImmunoTherapy of Cancer. 2021;9(Suppl 2):A855.
8. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nature Communications. 2020;11(1):3877.

Disclosures

The authors declare no potential conflict of interest during the work.

APPENDIX

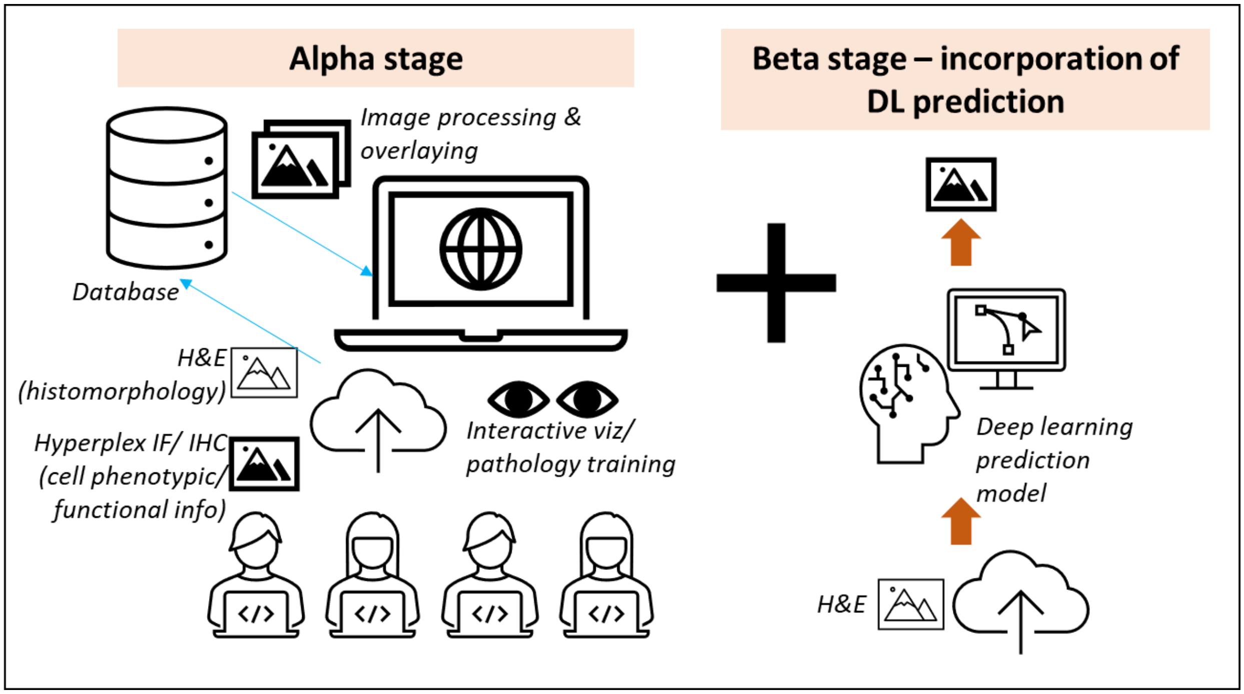
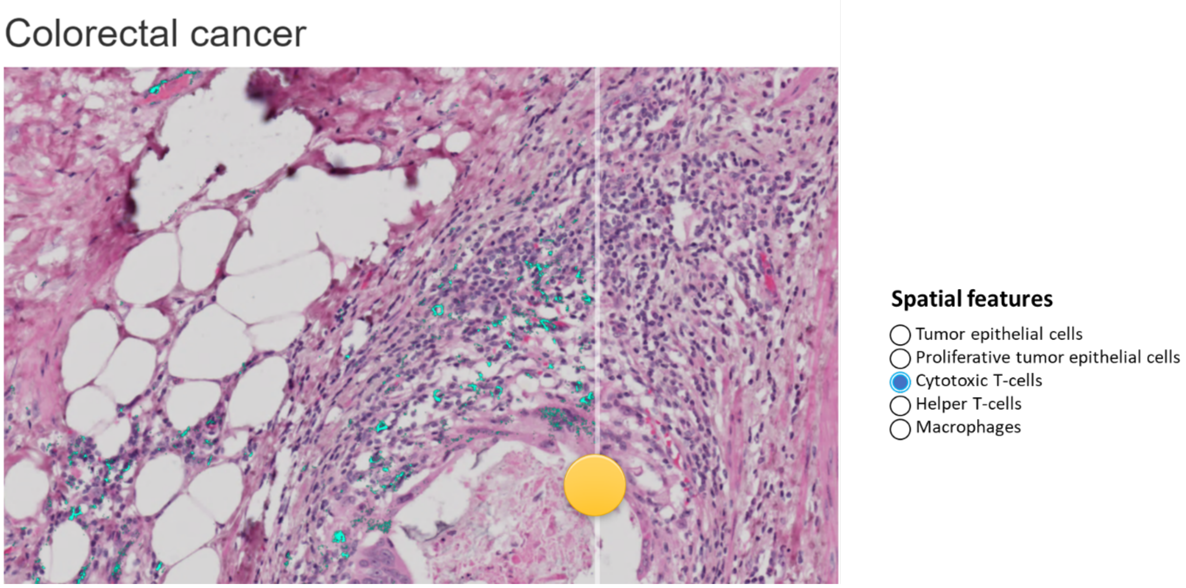

Alpha stage
Beta stage – incorporation of DL prediction
Image processing & overlaying
Database
H&E (histomorphology)
Hyperplex IF/ IHC (cell phenotypic/ functional info)
Interactive viz/ pathology training
Deep learning prediction model
H&E

Figure 1: The concept of H&E 2.0 web server serving as a shared data repository that enables interactive visualization of selected spatial features (like cell phenotypes) overlaid on H&E image space. At future beta stage, DL model(s) will be incorporated to allow predictions of cell types, gene expression, or other omics using low-cost H&E images uploaded by users.



Colorectal cancer

Spatial features

- ○ Tumor epithelial cells
- ○ Proliferative tumor epithelial cells
- ● Cytotoxic T-cells
- ○ Helper T-cells
- ○ Macrophages

Figure 2: A snippet of H&E 2.0 web server on a colorectal cancer tissue where various tumor and immune cell types quantified using hyperplex IF technique are overlaid on the H&E image space.

Radiology

Moderators: Mariam Aboian, MD PhD and Hersh Sagreiya, MD

Time	Presenting Author	Title
15:30-15:40	Andrew L Wentland (University of Wisconsin School of Medicine & Public Health)*	Transformer-Based Image Synthesis for Radiation Dose Reduction in Multi-Phase CT Imaging of the Kidneys
15:40-15:50	Ali Nabavizadeh (University of Pennsylvania)*	Automated Brain Tumor Subregion Segmentation on Multi-Parametric MRI Sequences of Pediatric Brain Tumors Across Multiple Institutions and Histologies
15:50-16:00	Dietmar Frey (Charité University Medicine Berlin)*	Prediction of hematoma expansion in acute intracerebral hemorrhage using a multimodal neural network model
16:00-16:10	Philip Edgcumbe (University Of British Columbia)*; Duncan Ferguson (University Of British Columbia); Joshua F Ho (University Of British Columbia)	Diagnosis of Pulmonary Emboli in Low Resource Settings with Rapid Serial Radiographs and IV Contrast Dual-Subtraction Radiography
16:10-16:20	Jacob Jaremko (University of Alberta)*	Automated segmentation of the humeral cortex and subacromial bursa with rotator cuff tear detection on shoulder ultrasound using deep learning
16:20-16:30	Discussion/Slush time	

Transformer-Based Image Synthesis for Radiation Dose Reduction in Multi-Phase CT Imaging of the Kidneys

Authors:

Andrew L. Wentland (MD, PhD)[1,2], Syed Jamal Safdar Gardezi (PhD)[1]

Affiliations:

University of Wisconsin School of Medicine & Public Health,

Department of Radiology[1], Department of Medical Physics[2]

Presenting author:

Andrew L. Wentland (MD, PhD)[1,2]

University of Wisconsin School of Medicine & Public Health,

Department of Radiology[1], Department of Medical Physics[2]

alwentland@wisc.edu

Keywords: Multiphase CT, urography, image synthesis, radiation reduction, kidneys

Key information:

1. Research question: To synthesize high-quality images for one of the phases in a 3-phase kidney CT. This goal will be achieved by employing a model that can leverage the redundant information in these CT examinations.
2. Findings: A two-channel transformer model can synthesize high-quality images in a 3-phase kidney CT. The synthesized images demonstrate high structural similarity indices relative to the ground-truth images.
3. Meaning: The ability to accurately synthesize one of the three phases in a kidney CT obviates the need to acquire all three phases. Radiation can be reduced 33% by synthesizing rather than acquiring this set of images.

MANUSCRIPT

Introduction

CT urography (CTU) is commonly performed for the evaluation of hematuria[1]. CTU studies include first a non-contrast set of images, second a nephrographic phase acquired ~90 seconds following the intravenous administration of iodinated contrast, and finally a urographic phase 5-10 minutes after the initial contrast injection. This 3-phase CTU study requires three times the radiation dose compared to conventional single-phase CTs[2]. There is inherently redundant information within the urographic phase set of images given the continued enhancement of the kidneys. Deep learning methods have the potential to exploit this redundancy. The purpose of this study was to develop a deep learning model to synthesize nephrographic phase images from the non-contrast and urographic phases (Figure-1). As a result, the nephrographic phase acquisition itself could be eliminated, effectively reducing a 3-phase to a 2-phase acquisition but still providing the unique nephrographic phase information. Moreover, the radiation dose can be reduced by ~33% of the original 3-phase CT.

Material and methods

A dataset of 101 patients (mean±SD age, 64±12 years; 61/40 males/females) with CTU studies was curated. The three phases were registered using affine rigid registration. A total of 4,080 registered images from the non-contrast, nephrographic, and urographic phases from the 101 patients were obtained and divided into an 80/20 train/test split.

A deep learning residual transformer (ResViT)[3] model was implemented with dual inputs of non-contrast and urographic sets of images. The model was tuned with 12 attention heads and 3,073 hidden units in each multilayer perceptron. The model was trained with a learning rate of {2×10^{-5}, 10^{-3}, 10^{-4}} and {200, 500, 700} epochs. The weighting of the pixel-wise, pixel consistency, and adversarial loss, respectively, were chosen as $\lambda_p = 100$, $\lambda_r = 100$ and $\lambda_{adv} = 1$. The structural similarity index measure (SSIM)[4] and peak signal-to-noise ratio (PSNR)[5] were computed for the synthesized images. Heat maps and kernel density estimate plots were used to compare real and synthesized images. The region of the kidneys was compared between real and synthesized images with cross correlation.

Results

Nephrographic phase images were successfully synthesized with the dual-input ResViT model, with negligible qualitative differences between synthesized and real images. The synthesized images yielded a mean PSNR of 28.1164±2.5806 as well as a mean SSIM of 0.9229±0.0406 as compared to real images. There were minimal differences in Hounsfield unit values between the real and synthesized images (Figure-2). A high cross correlation of 0.9642 was found between the real and synthesized images for the region of the kidneys.

Discussion and Conclusion

A ResViT model was implemented and customized for synthesizing nephrographic phase images in CTU studies from paired non-contrast and urographic phase images. The results demonstrate the robustness and effectiveness of the model in synthesizing nephrographic phase images. The synthesized images were highly similar to real images and achieved high SSIM and PSNR values.

The proposed ResViT model has great potential in clinical applications, particularly given that the method can yield a 33% radiation dose reduction in 3-phase CTU. Moreover, this study lays the groundwork for implementing the model in CTU examinations acquired with dual-energy CT (DECT). A standard component of DECT is the creation of virtual non-contrast images. The ResViT model can be re-trained to synthesize nephrographic images using the urographic and virtual non-contrast images. Thus, only a single dual-energy acquisition of the urographic phase would ultimately be needed. This approach yields a 66% radiation dose reduction as well as substantial scan time savings since only a single acquisition would now be needed instead of three. The ResViT model for synthesizing images in multiphase CT examinations of the kidney is well poised to translate quickly to clinical practice and substantially change the method with which these studies are performed.

Disclosures: None.

References

1. Silverman SG, Leyendecker JR, Amis Jr ES. What is the current role of CT urography and MR urography in the evaluation of the urinary tract? Radiology. 2009;250(2):309-323.
2. Nawfel RD, Judy PF, Schleipman AR, et al. Patient radiation dose at CT urography and conventional urography. Radiology. 2004;232(1):126-132.
3. Dalmaz O, Yurt M, Çukur T. ResViT: Residual vision transformers for multi-modal medical image synthesis. arXiv preprint arXiv:210616031. 2021.
4. Wang Z, Bovik AC, Sheikh HR, et al. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004;13(4):600-612.
5. Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM. 2010 20th international conference on pattern recognition: IEEE 2010:2366-2369.

APPENDIX

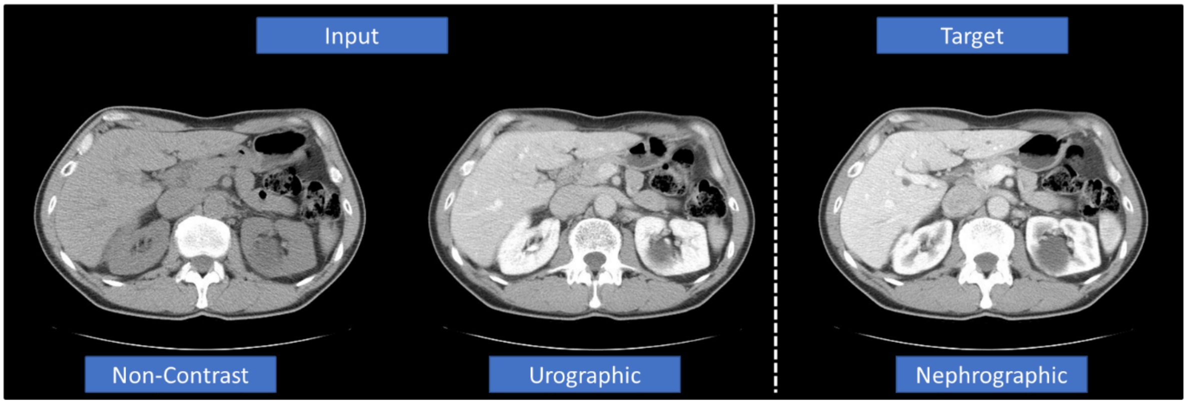


Figure 1: Paired non-contrast and urographic images are provided as input for the ResViT model to synthesize the nephrographic image.

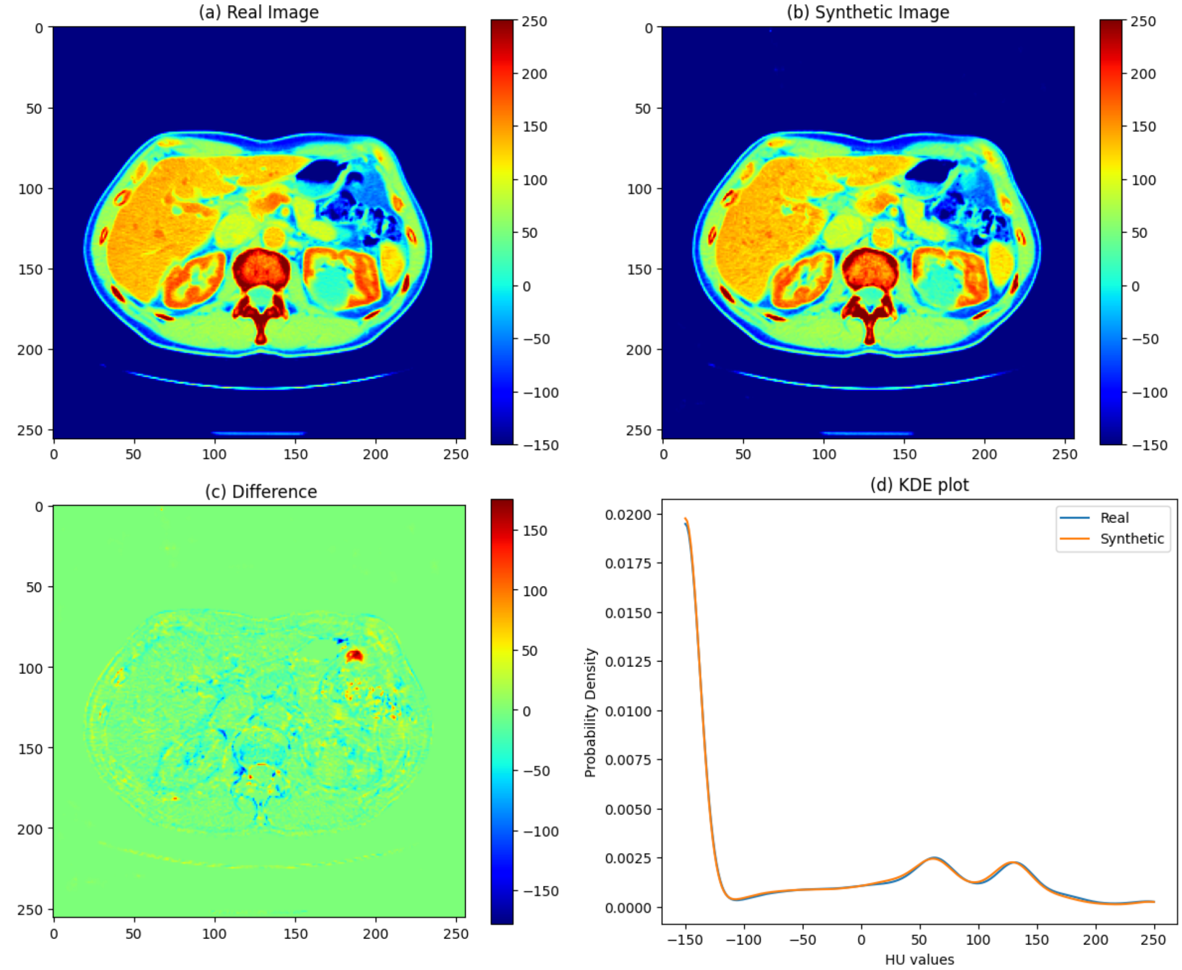


Figure 2: Heat maps of Hounsfield units (a, b) for the real and synthetic generated images and the difference between the real and synthetic images (c). The kernel density estimate (KDE) plot (d) shows high similarity between the real and synthetic images.

Automated Brain Tumor Subregion Segmentation on Multi-Parametric MRI Sequences of Pediatric Brain Tumors Across Multiple Institutions and Histologies

Authors:

Ali Nabavizadeh [1,2], Jeffrey B. Ware[1], Nastaran Khalili[2], Debanjan Haldar[2], Ariana Familiar[2], Karthik Viswanathan[2], Sina Bagheri[2], Hannah Anderson[2], Phillip B. Storm [2,3], Adam Resnick[2,3], Christos Davatzikos [1,4], Arastoo Vossough [5], Anahita Fathi Kazerooni [1,2,4]

Affiliations:

[1] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, US

[2] Center for Data-Driven Discovery in Biomedicine (D3b), The Children's Hospital of Philadelphia, Philadelphia, PA, US

[3] Department of Neurosurgery, The Children's Hospital of Philadelphia, Philadelphia, PA, US

[4] Center for AI And Data Science for Integrated Diagnostics (AI2D), Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, US

[5] Department of Radiology, The Children's Hospital of Philadelphia, Philadelphia, PA, US

Presenting author:

Ali Nabavizadeh, MD, Department of Radiology, Perelman School of Medicine, University of Pennsylvania & Center for Data-Driven Discovery in Biomedicine (D3b), The Children's Hospital of Philadelphia, Philadelphia, US; Email: ali.nabavizadeh@pennmedicine.upenn.edu .

Keywords:

Pediatric Brain Tumors, Response Assessment, Automatic Segmentation, Deep Learning, Magnetic Resonance Imaging

Key information:

1. Research question: Automatic Segmentation for Volumetric Assessment of Tumorous Subregions in Pediatric Brain Tumors
2. Findings: *Our proposed automated segmentation method shows high Dice Similarity Score in segmentation of pediatric tumor and its subregions. Furthermore, the predicted volumetric measurements are highly correlated with expert volumetric measurements.*
3. Meaning: *This model has the potential to be used in clinical assessment of treatment response in pediatric neuro-oncology trials.*

MANUSCRIPT *(included in the word count)*

Introduction

Pediatric brain tumors are the most common solid tumors with high level of heterogeneity [1]. The standard approach for response assessment in pediatric brain tumors is bidimensional (2D) measurements of the tumor size, as recommended by the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group [2, 3]. However, there have been studies suggesting an underestimation of tumor size using 2D measurements [4]. Volumetric assessment of tumors can measure changes of tumor size reliably, especially for irregularly shaped tumors, without assuming uniform change in tumor size in all dimensions. However, the complex structure of pediatric brain tumors and the mixed solid/cystic components pose a challenge on manual segmentation of tumorous subregions [5]. Aside from being time consuming and cumbersome, neuroradiologists without sufficient training pediatric neuroimaging may have difficulty in differentiating tumor subregions from each other, resulting in large intra and inter-rater variabilities. In this study, we utilized a fully automatic deep learning approach on a large cohort of pediatric brain tumors, collected across multiple tumor histologies, scanners, and institutions, to facilitate volumetric measurement of tumor subregions and thereby, assessment of tumor burden.

Material and methods

We utilized nnU-Net (‘no new net') self-configuring deep learning architecture for the task of brain tumor segmentation [6] on multiparametric standard MRI sequences (T1-pre, T1-post, T2, T2-FLAIR). The model was trained and validated on a large cohort of well-annotated pediatric brain tumors using 5-fold cross-validation to differentiate RAPNO-recommended subregions [5], including enhancing tumor (ET), non-enhancing tumor (NET), cystic components (CC), and peritumoral edema (ED). The data was collected through Children’s Brain Tumor Network (CBTN) [7] from internal and external institutions. The model was trained on an institutional cohort of 233 subjects and independently tested on 60 subjects from the internal and 46 from external institutions. Performance of the model in segmentation of tumor subregions as well as whole tumor, as a union of all subregions, was assessed by calculating Dice score and Hausdorff95 distance metrics. Furthermore, clinical validity of the proposed approach was tested by measuring the volumes of RAPNO-defined tumor subregions from the segmentations predicted by the model in comparison with the values obtained from expert manual segmentations, via Pearson’s correlation coefficient (significance level, $p < 0.05$).

Results

The trained model showed excellent performance with median Dice scores of 0.94±0.10/0.90±0.07 for whole tumor segmentation, 0.85±0.33/0.84±0.30 for ET subregion, 0.80±0.32/0.64±0.31 for NET, 0.79±0.37/0.67±0.33 for CC, 0.70 ±0.42/0.37±0.43 for ED, and 0.86±0.19/0.80±0.21 for all nonenhancing components (combination of NET, CC, and ED) in internal/external test sets, respectively (Table 1). Figure 1 showcases examples of predicted segmentations using our proposed model, compared to expert segmentation in subjects from internal and external sets. The automated segmentation demonstrated strong agreement with expert segmentations in volumetric measurement of tumor components, with Pearson’s correlation coefficients of 0.93/0.69, 0.94/0.93, 0.78/0.93, and 0.94/0.50 ($p<0.01$) for ET, NET, CC, and ED regions in internal/external test cohorts, respectively (Figure 2).

Discussion and Conclusion

Here, we proposed an automated segmentation method with accurate performance in segmentation of RAPNO-define subregions in pediatric brain tumors, with decent generalization to the data from external sites. This method may be incorporated with clinical neuro-oncology practice to generate reliable and reproducible volumetric measurements that can be used for treatment response assessment in pediatric brain tumors.

References

1. Ostrom QT, Price M, Ryan K, et al. CBTRUS statistical report: pediatric brain tumor foundation childhood and adolescent primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. Neuro-oncology. 2022;24(Supplement_3):iii1-iii38.
2. Erker C, Tamrazi B, Poussaint TY, et al. Response assessment in paediatric high-grade glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. The Lancet Oncology. 2020;21(6):e317-e329.
3. Cooney TM, Cohen KJ, Guimaraes CV, et al. Response assessment in diffuse intrinsic pontine glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. The Lancet Oncology. 2020;21(6):e330-e336.
4. Lazow MA, Nievelstein MT, Lane A, et al. Volumetric endpoints in diffuse intrinsic pontine glioma: comparison to cross-sectional measures and outcome correlations in the International DIPG/DMG Registry. Neuro-oncology. 2022;24(9):1598-1608.
5. Fathi Kazerooni A, Arif S, Madhogarhia R, et al. Automated tumor segmentation and brain tissue extraction from multiparametric MRI of pediatric brain tumors: A multi-institutional study. Neuro-Oncology Advances. 2023;5(1):vdad027.
6. Isensee F, Jaeger PF, Kohl SA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203-211.
7. Lilly JV, Rokita JL, Mason JL, et al. The children's brain tumor network (CBTN)-Accelerating research in pediatric central nervous system tumors through collaboration and open science. Neoplasia. 2023;35:100846.

Disclosures

Authors have nothing to disclose.

APPENDIX

Figures and pictures should be added in this section.

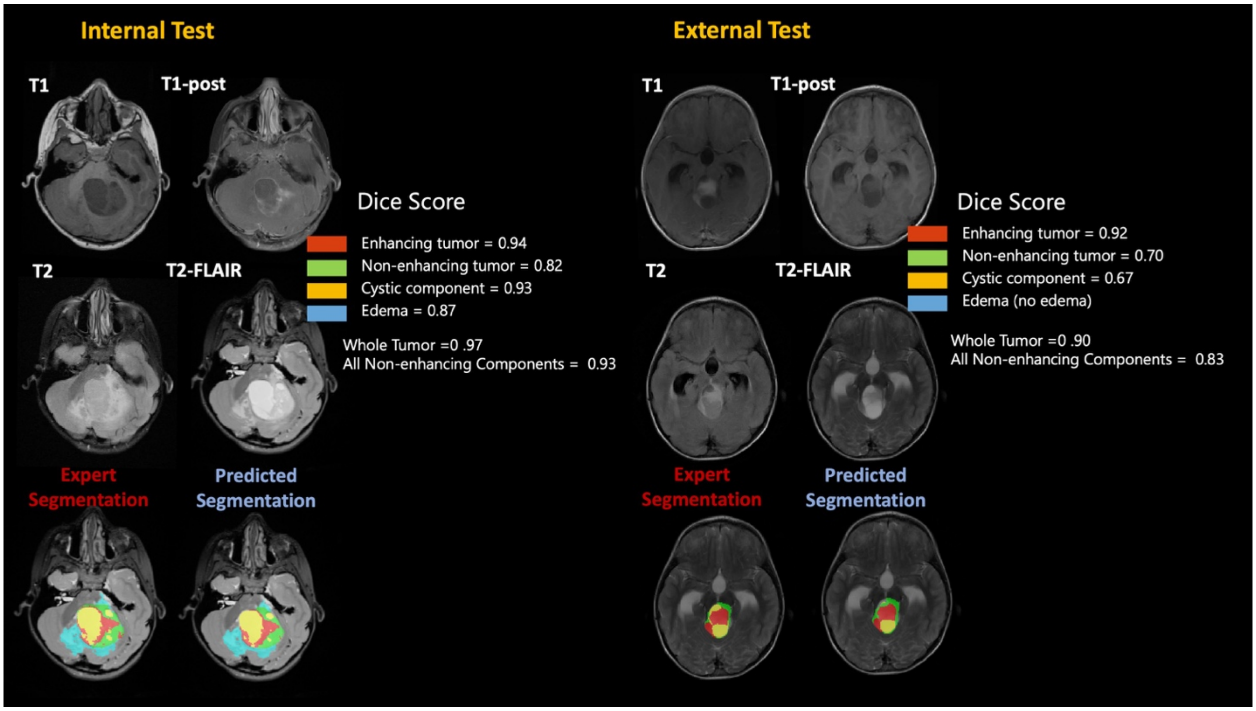
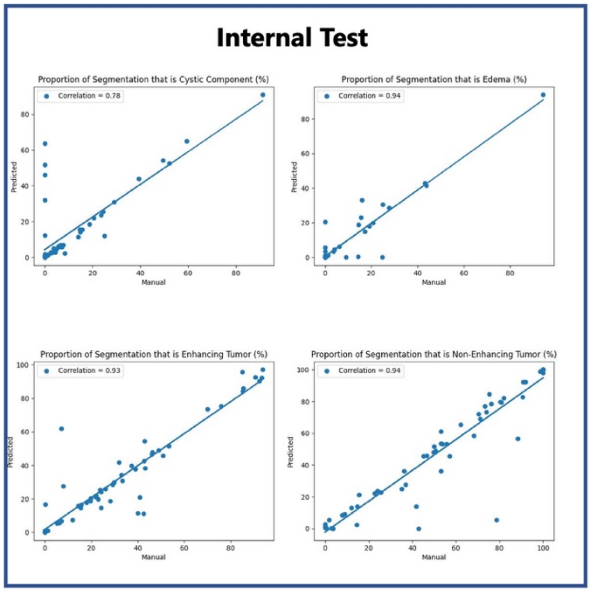


Figure 1: examples of subjects from internal and external test sets, with the predicted segmentation as compared to the expert (manual) segmentation results, as well as the Dice scores on the tumor subregions, the whole tumor, and all nonenhancing components (i.e., nonenhancing tumor, edema, and cystic component).



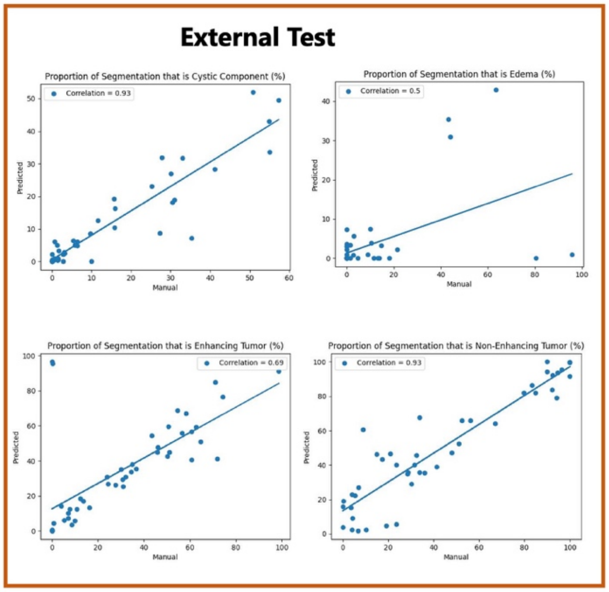


Figure 2: scatterplots indicating volumetric features computed from predicted versus manual segmentations for all subjects in the internal and external test sets. These features that are important for tumor response assessment according to RAPNO criteria include proportions of the whole tumor area that is cystic component, edema, enhancing tumor, nonenhancing tumor.

Tables should be added here.

Table 1. Summary of the performance of the proposed segmentation algorithm in terms of Median Dice Score and Hausdorff95 distance metrics, in the internal and external test sets, i.e., *$Test_i$* and *$Test_e$*, respectively.

Region	*Median Dice Score*		*Hausdorff95*	
	$Test_i$	*$Test_e$*	*$Test_i$*	*$Test_e$*
Whole Tumor	0.94	0.90	3.08	2.00
Enhancing Tumor	0.86	0.84	3.34	2.36
Nonenhancing Tumor	0.79	0.64	5.38	2.28
Cystic Component	0.78	0.67	6.40	3.60
Edema	0.70	0.36	25.97	6.40
All Nonenhancing Regions	0.86	0.80	3.61	2.24

Prediction of hematoma expansion in acute intracerebral hemorrhage using a multimodal neural network model

Authors: Dietmar Frey, MD [1,2], Orhun Utku Aydin[1], Adam Hilbert[1], Satoru Tanioka, MD PhD[1]

Affiliations: [1]Charité Lab for AI in Medicine, Charité Universitaetsmedizin Berlin, Germany
[2]Department of Neurosurgery, Charité Universitaetsmedizin Berlin, Germany

Presenting author: Dietmar Frey

Charité Lab for AI in Medicine, Charitéplatz 1, 10117 Berlin, Germany
Email: dietmar.frey@charite.de

Keywords: intracerebral hemorrhage, hematoma expansion, prediction, convolutional neural network, multimodal neural network

Key information:

1. Research question: Whether a multimodal neural network model using CT images and clinical variables as input is superior to a convolutional neural network model using CT images alone in predicting hematoma expansion in acute intracerebral hemorrhage.
2. Findings: A multimodal neural network model showed perfect sensitivity and better performance than a convolutional neural network model.
3. Meaning: Since hematoma expansion leads to neurological deterioration and the need for surgical treatment, its prediction by the built multimodal neural network model at the time of admission is very helpful in developing patient management strategies.

MANUSCRIPT

Introduction

In intracerebral hemorrhage (ICH), hematoma expansion occurs in 20-30% of cases, leading to neurological deterioration and the need for surgical treatment.[1] Although various CT markers and scoring systems have been proposed to predict hematoma expansion, CT image analysis using a convolutional neural network (CNN) has shown better predictive performance than these.[2,3] However, in addition to CT images, clinical information such as anticoagulant use and time from onset to baseline imaging are important factors in predicting hematoma expansion. In this study, we aimed to develop a multimodal neural network model using CT images and clinical variables as input and to validate the superiority of the multimodal model over the CNN model using CT images alone.

Material and methods

Patients with ICH admitted to 4 hospitals were retrospectively enrolled in the study (Table 1). Intraparenchymal hematoma was annotated using 3D Slicer, where hematoma volumes were calculated. Hematoma expansion was defined as an increase in volume between baseline and follow-up CT scans exceeding 6 cm^3 or 33%. 70% of the patients in 3 hospitals were randomly assigned to the training set and the rest in those hospitals to the validation set; patients in one other hospital were assigned to the test set. Clinical variables at admission were collected from each patient, for which univariate analyses were performed between expansion and no expansion cases in the training set, from which statistically significant variables were extracted (Table 2).

The computational environment and the preprocessing of the CT images were summarized in Table 3. First, two CNN models were created and compared using whole brain images (CNN model 1) and intraparenchymal hematoma images (CNN model 2) as input. The model architectures were shown in Figure 1. Then, whole brain images or intraparenchymal hematoma images, whichever was superior in the comparison, was used as one input for multimodal neural network models, where two more models were created using all clinical variables (multimodal model 1) and statistically significant variables (multimodal model 2) as the other input. In each of the 4 models, 70 epochs of training were performed, where 10 model weights that showed better sensitivity and area under the curve (AUC) were selected and used for testing. Among them, the one with the highest sensitivity was selected as the test result in each model.

Results

273 patients were enrolled in the training and validation sets and 106 patients in the test set; hematoma expansion occurred in 54 (19.8%) of the former and 14 (13.2%) of the latter. Both CNN model 1 and CNN model 2 showed a sensitivity of 1.000, but CNN model 2 showed higher AUC (Table 3). Therefore, intraparenchymal hematoma images were used for the multimodal neural network models. Multimodal model 2 performed best among 4 models (Table 4).

Discussion and Conclusion

We developed and validated a multimodal neural network model using CT images and clinical variables, showing perfect sensitivity and better performance than the CNN model using CT images alone. Since hematoma expansion leads to neurological deterioration and the need for surgical treatment, its prediction by this model at admission is helpful in developing patient management strategies. Although the sensitivity of 1.000 was achieved, the lower limit of its confidence interval was 0.681, indicating that more cases are needed for more reliable model validation.

References

1. Al-Shahi SR, et al. Absolute risk and predictors of the growth of acute spontaneous intracerebral haemorrhage. *Lancet Neurol* 2018;17:885–894.
2. Zhong J, et al. Deep learning for automatically predicting early haematoma expansion in Chinese patients. *Stroke&Vascular Neurology* 2021;6:e000647.
3. Brouwers HB, et al. Predicting hematoma expansion after primary intracerebral hemorrhage. *JAMA Neurol* 2014;71:158–164.

Disclosures

Conflict of interest: None.

APPENDIX

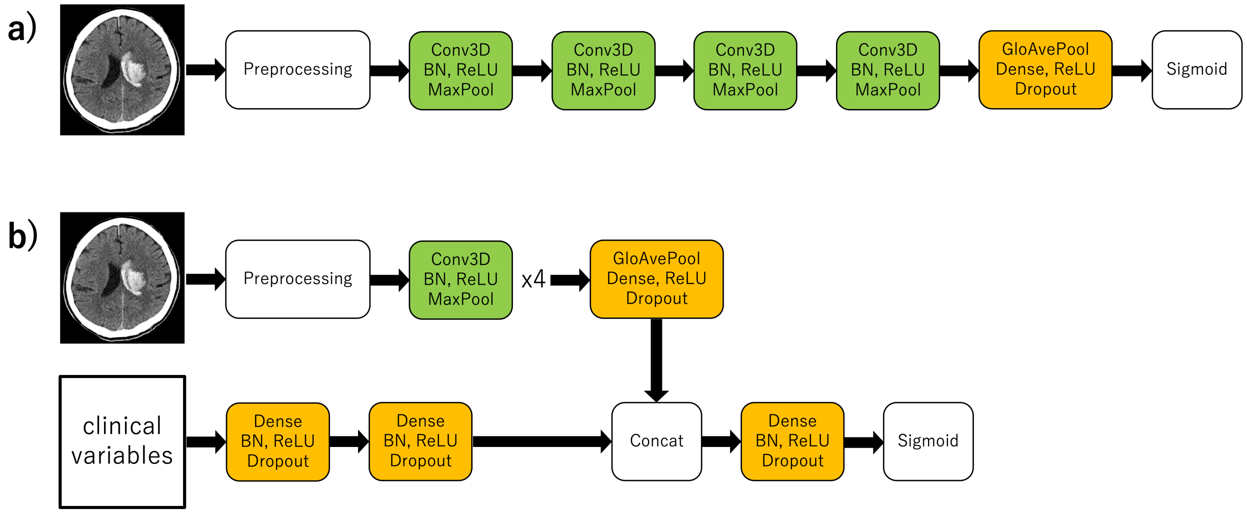


Figure 1. Neural network model architectures. **A**) A convolutional neural network (CNN) model using only CT images as input (CNN model). After preprocessing, 4 convolutional layers followed with kernel sizes of 19x19x7, 19x19x7, 14x14x5, and 11x11x4, respectively. **b**) A multimodal neural network model using CT images and clinical variables as input (multimodal model). The upper part was the same architecture as in **a**, and the lower part was added for clinical variables.

Table 1. Inclusion and exclusion criteria for the study

Inclusion criteria	Exclusion criteria
● ≥ 18 years of age ● Baseline CT scan within 24 hours of onset ● Follow-up CT scan within 30 hours of baseline CT scan ● Baseline and follow-up CT scans with a thickness of 0.5-2.0 mm and an image size of 512 x 512 or greater	● Traumatic ICH, ● Secondary cause of ICH ● Surgical evacuation before follow-up CT scan

ICH = intracerebral hemorrhage

Table 2. Clinical variables at admission

All variables	Statistically significant variables*
● Age ● Sex ● Medical history (ICH, cerebral infarction, ischemic heart disease, hypertension, diabetes mellitus, and dyslipidemia) ● Anticoagulant use ● Antiplatelet use ● Glasgow Coma Scale ● Systolic blood pressure ● Diastolic blood pressure ● Blood test (PT-INR, white blood cell count, hemoglobin, platelet count, serum creatinine, and serum total bilirubin) ● Time from onset to baseline CT scan	● Anticoagulant use ● Systolic blood pressure ● Diastolic blood pressure ● PT-INR ● Time from onset to baseline CT scan

*The variables with $p<0.05$ in the Mann-Whitney U test and Fisher's exact test between expansion and no expansion cases in the training data set were defined as statistically significant variables. ICH = intracerebral hemorrhage, PT-INR = prothrombin time-international normalized ratio

Table 3. Computational environment and preprocessing of the CT images

Computational environment	Preprocessing
● All data processing was done in Python and Keras ● The code was executed in Google Colab Pro (40 GB of GPU memory).	● Brain and hematoma extraction with the application of Haunsfield units from 0-100 and subsequent normalization ● Reslicing with a slice thickness of 2 mm, where the number of slices was 80 ● Standardizing the pixel size to 0.5 mm x 0.5 mm ● Resizing the image size to 256 x 256 ● Data augmentation by image flipping and rotation only for expansion cases due to imbalance between expansion and no expansion cases

Table 4. Result of 4 models for the test set

	Sensitivity	Specificity	AUC
CNN model 1	1.000 (0.681-1.000)	0.163 (0.094-0.255)	0.582 (0.544-0.614)
CNN model 2	1.000 (0.681-1.000)	0.511 (0.404-0.617)	0.755 (0.704-0.807)
Multimodal model 1	0.857 (0.572-0.972)	0.717 (0.614-0.806)	0.787 (0.682-0.893)
Multimodal model 2	1.000 (0.681-1.000)	0.598 (0.490-0.699)	0.799 (0.749-0.849)

Data are presented as value (95% confidence interval). AUC = area under the receiver operating characteristic curve

Diagnosis of Pulmonary Emboli in Low Resource Settings with Rapid Serial Radiographs and IV Contrast Dual-Subtraction Radiography

Authors:

Philip Edgcumbe[1], Duncan Fergusson[1], Joshua Ho[2]

Affiliations:

1: Department of Radiology, University of British Columbia, Vancouver, Canada
2: School of Biomedical Engineering, University of British Columbia, Vancouver, Canada

Presenting author:

Philip Edgcumbe.
Affiliation: Department of Radiology, University of British Columbia, Vancouver, Canada

Contact: pedgcumbe@gmail.com

Keywords:

1. Pulmonary Embolism
2. Healthcare in low-resource settings
3. Radiographs
4. Computed Tomography

Key information:

1. Research question: Can CT scan data be used to simulate radiographs? Can medium and large pulmonary embolisms (PEs) be diagnosed in low-resource settings without CT scanners? Can rapid serial radiography (RSR) and IV dual-subtraction radiographs (IV-DSR) be used to detect pulmonary emboli?
2. Findings: CT scan data can be used to simulate radiographs. It is feasible to retrospectively test rapid serial radiography (RSR) and IV dual-subtraction radiographs (IV-DSR) for PE detection using existing CT scan datasets and prospectively test the same using a simple clinical study design.
3. Meaning: In conclusion, we have described initial simulation data and a proposed framework for diagnosing PEs in low-resource settings where CT scanners are not available. In its success, RSR and IV-DSR for detection of PEs would reduce time to diagnosis of PEs and save lives.

MANUSCRIPT

Introduction

A pulmonary embolism (PE) is a clot that blocks the pulmonary arteries and prevents blood from traveling from the heart to the lungs. Patients who develop PEs often present with dyspnea and pleuritic chest pain as well as potentially life-threatening low blood oxygen levels.[1] In a study by Mansella et al, the mortality rate with early diagnosis of PEs was 1.6% vs a mortality rate of 43.2% for delayed diagnosis.[2] The gold standard for diagnosing PEs is computed tomography pulmonary angiography (CTPA). However, in low to middle income countries, and in rural settings around the world, there is limited or non-existant access to CT scanners.[3,4] When a CT scanner is not available, a viable alternative for detecting medium to large PEs is fluoroscopy-guided catheter angiography. Furthermore, a study by Musset et al showed a sensitivity of 94% for detection of medium and large PEs using IV-DSA, obviating the need of catheter angiography.[5] In this paper, we propose using IV contrast, rapid serial radiography (RSR) and IV contrast dual-subtraction radiographs (IV-DSR) for detection of PEs.

The aim of this paper is to:

1. Demonstrate how CT scan data can be used to simulate RSR and IV-DSR.
2. Discuss the design of prospective clinical study to test the accuracy of RSR and IV-DSR for detection of PEs.

Material and methods

A single negative CTPE CT study from the RSNA PE dataset was used to simulate a frontal radiograph.[6] The conversion from CT scan to frontal radiograph was done by taking an average of the Hounsfield Units (HU) along the AP dimension of the CT scan. Next, an axial slice of the CT scan was selected that intersected with the left pulmonary artery and a synthetic version of that same axial slice with PE was created by replacing the attenuation values with the pulmonary artery (approximately 400 HU) with the attenuation values of a PE (33 HU). The average linear attenuation for the x-ray that intersected the left pulmonary artery, and the synthetically added PE, was calculated for the axial slice with no PE and the synthetic addition of a PE.

Results

The average linear attenuation of the x-ray that intersected the pulmonary artery in the CTPE negative axial slice was -200 HU (Figure 1). The average linear attenuation of the x-ray that intersected the PE in the pulmonary artery in the synthetic dataset was -222 HU (Figure 2). The simulated chest radiograph is shown in Figure 3.

Discussion and Conclusion

We have simulated chest radiographs in which patients are given IV-contrast before the radiograph is taken. Furthermore, we have calculated the average linear attenuation for a specific x-ray that bisects a left pulmonary artery that is filled with IV-contrast vs a left pulmonary artery with a fully occlusive PE. The results showed that the difference in linear attenuation is 22 HU. Generally, a radiograph is windowed for display of average linear attenuation from -1,000 to 1,000 HU. Thus, the difference in contrast in a IV-contrast radiograph with and without a PE is about 1% (22/2,000). This shows that IV-DSR would be required for detection of PEs with radiographs. Our next step will be to identify patients at our institution who had triple-phase CT chest scans (part of our trauma imaging protocol) and were found to have PEs. We will then use the methods developed for this paper to convert the pre and post-contrast CT scans into radiographs and generate the equivalent of RSR IV-DSR images.

Finally, if the initial study with simulated radiographs is promising, we will undertake a prospective randomized control trial in which we enroll patients with and without PEs to undergo RSR and IV-DSR. An RSR IV-DSR study will include a non-contrast chest radiograph followed by RSR (5 radiographs/second over 4 seconds so as to cover approximately one respiratory cycle) after administration of IV-contrast and an appropriate delay to allow the contrast to reach the pulmonary artery. DSR images will be created by subtracting the initial non-contrast radiograph image from the images captured using RSR. The study’s interpreting radiologist will review the subtracted images and identify the one with the least patient movement and respiratory motion and highest diagnostic value and proceed to make a diagnosis.

In conclusion, we have described initial simulation data and a proposed framework for diagnosing PEs in low-resource settings where CT scanners are not available. In its success, IV-DSR for detection of PEs would reduce time to diagnosis of PEs and save lives.

References

1. Tapson VF. Acute Pulmonary Embolism. *New England Journal of Medicine*. 2008;358(10):1037-1052. doi:10.1056/NEJMra072753

2. Mansella G, Keil C, Nickel CH, et al. Delayed Diagnosis in Pulmonary Embolism: Frequency, Patient Characteristics, and Outcome. *Respiration*. 2020;99(7). doi:10.1159/000508396

3. Jamil H, Tariq W, Ameer MA, et al. Interventional radiology in low- and middle-income countries. *Annals of Medicine and Surgery*. 2022;77. doi:10.1016/j.amsu.2022.103594

4. Bergeron C, Fleet R, Tounkara FK, Lavallée-Bourget I, Turgeon-Pelchat C. Lack of CT scanner in a rural emergency department increases inter-facility transfers: A pilot study. *BMC Res Notes*. 2017;10(1). doi:10.1186/s13104-017-3071-1

5. Musset D, Rosso J, Petitpretz P, et al. Acute pulmonary embolism: Diagnostic value of digital subtraction angiography. *Radiology*. 1988;166(2). doi:10.1148/radiology.166.2.3275984

6. Colak E, Kitamura FC, Hobbs SB, et al. The RSNA Pulmonary Embolism CT Dataset. *Radiol Artif Intell*. 2021;3(2). doi:10.1148/ryai.2021200254

Disclosures

No potential conflict of interest. Nothing to disclose.

APPENDIX

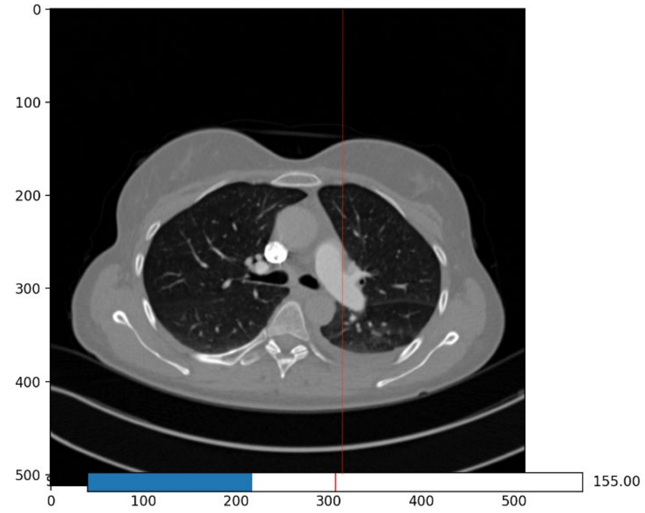


Figure 1: Axial slice from negative CTPE CT study from the RSNA PE dataset. The red line shows a ray that intersects the left pulmonary artery. The average linear attenuation of all the pixels in this ray's path was calculated to be -200 HU.

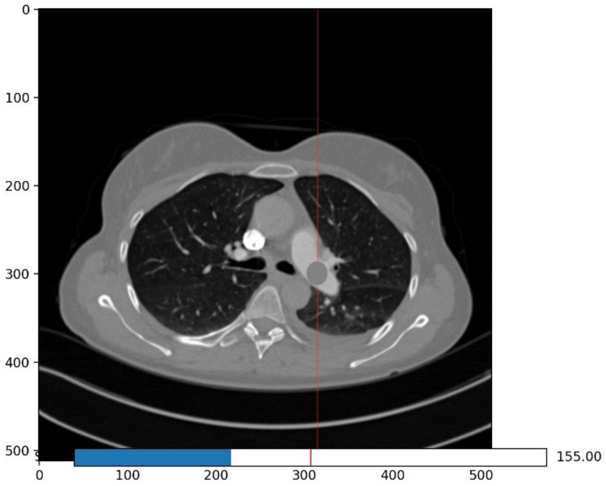


Figure 2: Axial slice from negative CTPE CT study from the RSNA PE dataset. A grey circle has been added to the left pulmonary artery to simulate a pulmonary embolism. The red line shows a ray that intersects the left pulmonary artery and the synthetic pulmonary embolism. The average linear attenuation of all the pixels in this ray's path was calculated to be -222 HU.

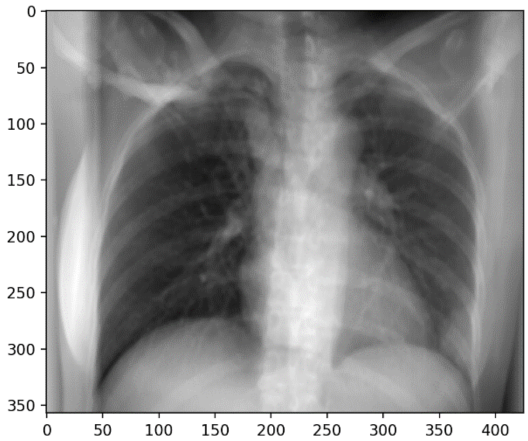


Figure 3: Radiograph generated from a negative CTPE CT study from the RSNA PE dataset. The conversion from CT scan to frontal radiograph was done by taking an average of the Hounsfield Units (HU) along the AP dimension of the CT scan.

AUTOMATED SEGMENTATION OF THE HUMERAL CORTEX AND SUBACROMIAL BURSA WITH ROTATOR CUFF TEAR DETECTION ON SHOULDER ULTRASOUND USING DEEP LEARNING

Authors:

Jacob L Jaremko[1], Shrimanti Ghosh[1], Jessica Knight[1], Natasha Akhlaq[1], Abhilash R Hareendranathan[1]

Affiliations:

[1]Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, Canada

Presenting author:

Jacob L. Jaremko, MD, PhD, FRCPC
Radiologist, Professor and Canada CIFAR AI Chair
Dept. of Radiology and Diagnostic Imaging
Faculty of Medicine and Dentistry
University of Alberta Hospital
2A2.41 WMC, 8440 – 112 St. NW
Edmonton, AB, Canada T6G 2B7
http://www.ualberta.ca/~jjaremko/
E-mail: jjaremko@ualberta.ca

Keywords:

Shoulder Ultrasound, Rotator Cuff, Deep Learning, CNN, Segmentation, Classification

Key information:

1. Research question: We seek to develop a novel deep learning (DL) framework to automatically segment the humeral cortex, subacromial bursa, including detection of full thickness rotator cuff tears automatically from shoulder ultrasound (US) cine sweeps. The purpose is to determine whether Artificial Intelligence (AI) can be utilized to detect rotator cuff tendon tears from shoulder US images.
2. Findings: Our proposed architecture based on a modified version of U-Net achieved 92% accuracy in segmenting the rotator cuff tendons, humeral cortex, and subacromial bursa. After that, a CNN architecture VGG-16 achieved 78% accuracy in classifying the rotator cuff tendons as intact or torn from US images in a preliminary data set.

3. Meaning: AI detection of rotator cuff tears from ultrasound directly adds value to the health care system as it could allow non-specialized providers to quickly detect rotator cuff tears from ultrasound, reducing long wait times for patients and decreasing overall health care system costs.

MANUSCRIPT

Introduction

Chronic shoulder problems are the second-most prevalent orthopedic complaint after knee problems worldwide [1]. Anatomical assessment for rotator cuff injury requires accurate segmentation of the humeral cortex, rotator cuff tendon and subacromial bursa [2]. Currently, this is performed manually by a musculoskeletal radiologist, which is expensive, tedious, and time-consuming [3]. Ultrasound (US) is a faster and less expensive alternative to MRI [4]. We proposed a novel end-to-end deep learning (DL) based approach to segment clinically relevant regions like the humeral cortex, subacromial bursa and rotator cuff tendons and to identify whether the tendons are intact or torn based on these regions.

Material and Methods

In recent years, the U-Net [5] has achieved increasing success in the segmentation of MRI and CT data. However, ultrasound segmentation is more challenging due to the inherent speckle noise and artifacts. We proposed a modified version of U-Net [6] to automatically segment the humeral cortex, subacromial bursa, and rotator cuff tendons from 2D US image cine sweeps. The modified version makes use of a U-Net based backbone network incorporated with a bidirectional feature network for the task of segmentation [6]. After that, the original US images and the corresponding segmentation are passed inside another CNN architecture VGG-16 [7, 8] to detect the rotator cuff tear in the tendons. This study was performed on a dataset collected retrospectively, with institutional ethics approval, from 56 adult subjects with MRI-confirmed intact (n=29) or torn (n=27 full thickness tears) rotator cuffs, containing 8860 2D US images. The whole dataset was divided into 70% for training, 15% for validation, and the remaining 15% for testing. The training approach utilizes 5-fold cross-validation to obtain an accurate measure of the generalizing capability of the proposed model. Training parameters are chosen as follows: learning rate = $1x10^{-5}$, batch size = 16, epochs = 200. It took around 5 hours to train the network and the prediction took 4 ms per image on two NVIDIA GeForce GTX 1080 GPU processors.

Results

Table 1. reports the Dice Coefficient (DC) and the Hausdorff Distance (HD) between manual segmentation and the segmentation by automated methods. Our proposed method performed significantly better than the current state-of-the-art methods [7-10]. Our proposed architecture based on a modified version of U-Net achieved 0.92 Dice Coefficient in segmenting the rotator cuff tendons, humeral cortex and subacromial bursa. After that a CNN architecture VGG-16 achieved 78.6% accuracy (sensitivity 76.5%, specificity 75.3%) in classifying rotator cuff tendons as intact or torn from US images.

Discussion and Conclusion

Segmentation of the humeral cortex and subacromial bursa is difficult from ultrasound images due to noise, speckle, and artifacts, making it challenging to distinguish between different structures and boundaries accurately. This project aims to provide the core AI process needed to transform the current care pathway used for the assessment of shoulder injuries; specifically identifying full-thickness rotator cuff tears within minutes, rather than months, from when the patient consults a care provider. We developed an AI tool that identifies regions in the shoulder scans that are common sites of rotator cuff tears and detects the tears based on these regions. Compared to an end-to-end black box classifier, our two-stage approach is more explainable and focuses on clinically relevant landmarks. Our segmentation accuracy was high. Although the classification accuracy (79%) was not yet high enough for clinical use, this is a highly promising result in a small preliminary data set and provides motivation for us to expand this study to a larger data set (including some partial-thickness tears).

If successfully validated in a larger cohort, the automated tool could be used by lightly trained users at initial point-of-care facilities like family physician clinics and emergency rooms. In addition to our planned large-scale study with >200 subjects, we plan to explore weakly supervised and unsupervised approaches which could potentially improve prediction model accuracy and efficiency.

References

[1] Zheng F, Wang H, Gong H, Fan H, Zhang K, Du L. Role of Ultrasound in the Detection of Rotator-Cuff Syndrome: An Observational Study. Med Sci Monit. 2019;25:5856-5863.

[2] Rutten MJ, Jager GJ, Kiemeney LA. Ultrasound detection of rotator cuff tears: observer agreement related to increasing experience. AJR Am J Roentgenol. 2010 Dec;195(6):W440-6.

[3] Fischer CA, Weber MA, Neubecker C, Bruckner T, Tanner M, Zeifang F. Ultrasound vs. MRI in the assessment of rotator cuff structure prior to shoulder arthroplasty. J Orthop. 2015 Jan 28;12(1):23-30.

[4] Abdolali F, Kapur J, Jaremko JL, Noga M, Hareendranathan AR, Punithakumar K. Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. Comput Biol Med. 2020 Jul;122:103871.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on MICCAI, Springer, 2015.

[6] Keetha, Nikhil & Parisapogu, Samson Anosh Babu & Annavarapu, Chandra. U-Det: A Modified U-Net architecture with bidirectional feature network for lung nodule segmentation. EESS 2020.

[7] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, CVPR, 2015.

[8] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI Conference on Artificial Intelligence, 2017.

[9] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, UNet++: A Nested U-Net Architecture for Medical Image Segmentation, CVPR, 2018.

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, Yuyin Zhou, TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation, CVPR, 2021.

Disclosures

The authors declare that they have no financial or non-financial conflicts of interest in relation to this research study. Our research and findings are conducted independently and without bias.

APPENDIX

Figure 1.

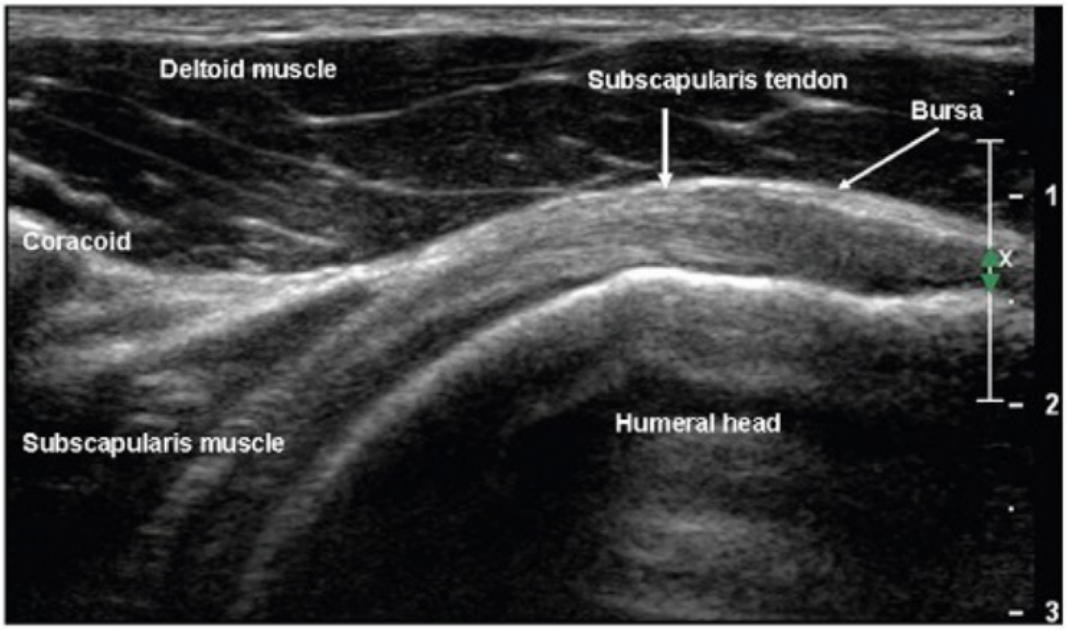


Figure 1: Shoulder ultrasound scan showing anatomical landmarks like humeral head cortex, subacromial bursa and subscapularis muscle.

Figure 2.

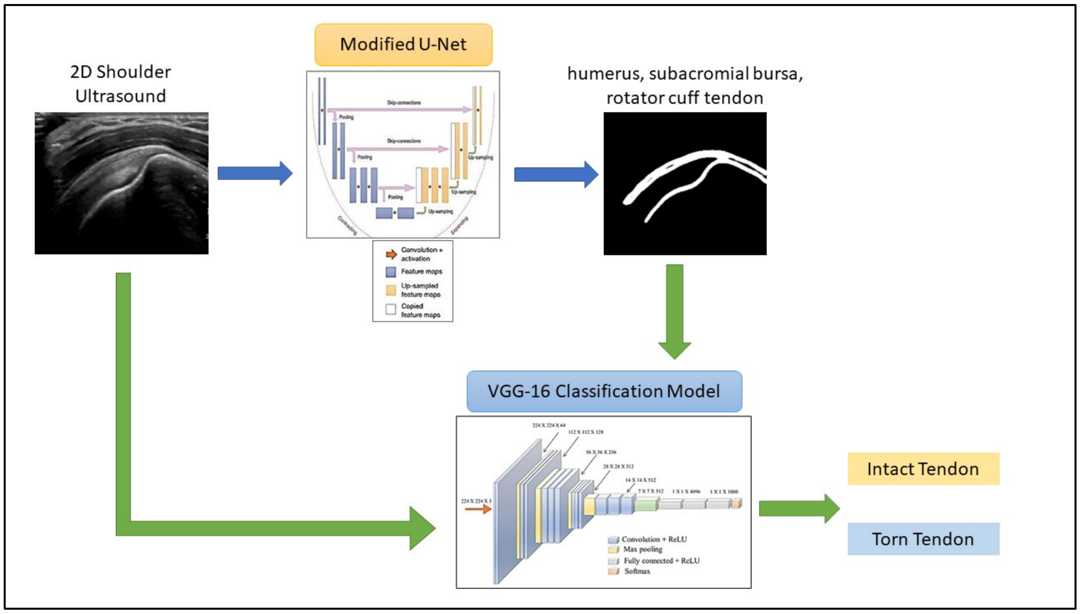
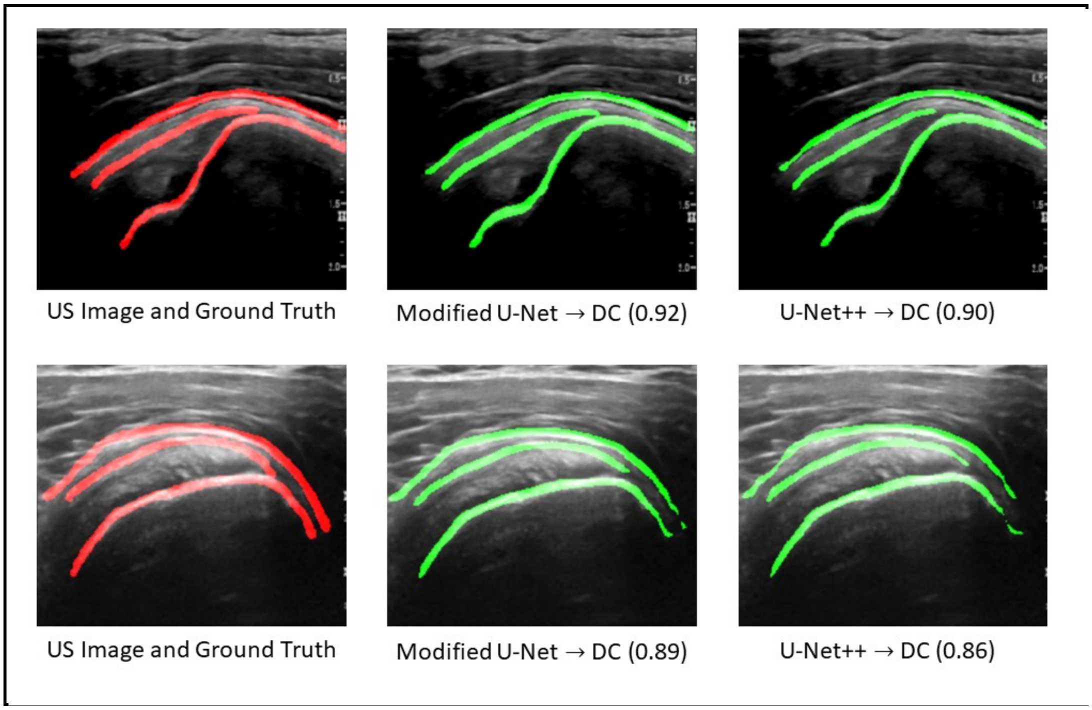


Figure 2: Illustration of our proposed solution combining the segmentation and classification models.

Figure 3.



US Image and Ground Truth

Modified U-Net → DC (0.92)

U-Net++ → DC (0.90)

US Image and Ground Truth

Modified U-Net → DC (0.89)

U-Net++ → DC (0.86)

Figure 3: The automated segmentation results of Modified U-net and U-Net++ with different Dice coefficient (DC) values. The red segmentation represents the manual segmentation or ground truth.

Table 1.

Evaluation of automated segmentation results in comparison to expert manual segmentation. The higher the Dice coefficient or the lower the Hausdorff distance the better the results.

Methods	Dice Coefficient (DC) (%)	Hausdorff Distance (HD) (mm)
Modified U-Net (Our method)	**92.4 ± 2.3**	**2.06 ± 2.8**
U-Net	88.5 ± 4.6	6.7 ± 3.5
TransUNet	82.2 ± 8.7	8.1 ± 4.2
U-Net++	90.3 ± 3.5	6.01 ± 4.9

Table 2.

Evaluation of automated segmentation results in comparison to expert manual segmentation. The higher the Dice coefficient or the lower the Hausdorff distance the better the results.

Methods	Accuracy (%)	Sensitivity	Specificity	Precision
VGG-16	**78.62**	**76.50**	**75.34**	**77.01**
ResNet	76.53	75.05	73.68	72.91
Inception-V4	72.03	70.46	73.24	71.46