Annals of Surgery Publish Ahead of Print
DOI:10.1097/SLA.0000000000006577

OPEN

Artificial Intelligence Assisted Surgical Scene Recognition: A Comparative Study Amongst Healthcare Professionals

Simon C Williams MRCS[1,2*], Jinfan Zhou PhD[1,3,*], William R Muirhead FRCS[1,2], Danyal Z Khan MRCS[1,2], Chan Hee Koh MRCS[1,2], Razna Ahmed MBBS[1], Jonathan P Funnell MRCS[1], John G Hanrahan MRCS[1,2], Alshaymaa Mortada Ali MSc[4], Shankhaneel Ghosh MBBS[5], Tarık Sarıdoğan MBBS[6], Alexandra Valetopoulou MBBS[7], Patrick Grover FRCS[2], Danail Stoyanov PhD[1], Mary Murphy FRCS[2], Evangelos B Mazomenos PhD[1,□], Hani J Marcus FRCS[1,2,□]

Affiliations

[1]Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), London, United Kingdom

[2]Victor Horsley Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, United Kingdom

[3]Robotics Institute, University of Michigan, Ann Arbor, MI, USA

[4]Ain shams university, Cairo, Egypt

[5]Institute of Medical Sciences and SUM Hospital, Bhubaneswar, India

[6]Uludağ University, Faculty of Medicine, Bursa, Turkey

[7] Department of Neurosurgery, Imperial College Healthcare NHS Trust, London, UK

* Denotes joint first authorship

□ Denotes joint senior authorship

Corresponding Author:

Mr Simon Williams

simon.williams32@nhs.net

Key words:

Artificial Intelligence; Computer Vision; Machine Learning; Aneurysm; Neurosurgery; Vascular Neurosurgery

Counts:

Abstract: 363

Manuscript: (excluding references, tables, figures, headings): 2993

Figures/Tables: 2/3

Videos: 0

Number of references: 36

Running head:

Automated Artificial Intelligence Detection of Cerebral Aneurysms

Funding:

No specific funding was received for this piece of work. SW, JF, JH, DZK, WM, DS, EM, HJM, are supported by the Wellcome (203145Z/16/Z) EPSRC (NS/A000050/1) Centre for Interventional and Surgical Sciences, University College London. HJM is also funded by the NIHR Biomedical Research Centre at University College London. DZK and JGH are funded by an NIHR Academic Clinical Fellowship. This research was funded in whole, or in part, by the Wellcome Trust [203145Z/16/Z].

Author contribution statement:

Simon Williams – conceptualisation, data curation, formal analysis, methodology, project administration, data analysis, writing – original draft preparation.

Jinfan Zhou – Conceptualisation, data curation, software, writing – review and editing.

William R Muirhead – conceptualisation, data duration, methodology, project administration, supervision, writing – review and editing.

Danyal Z Khan – formal analysis, methodology, writing – review and editing.

Chan Hee Koh - formal analysis, methodology, data analysis, writing – review and editing

Razna Ahmed – formal analysis, methodology, writing – review and editing.

Jonathan P Funnell – methodology, project administration, writing – review and editing.

John G Hanrahan – methodology, project administration, writing – review and editing.

Alshaymaa Mortada Ali – data curation, project administration, writing – review and editing.

Shankhaneel Ghosh – data curation, project administration, writing – review and editing.

Tarık Sarıdoğan - data curation, project administration, writing – review and editing.

Alendra Valetopoulou - data curation, project administration, writing – review and editing.

Patrick Grover – data curation, methodology, project administration, supervision, writing – reviewing and editing.
Danail Stoyanov – project administration, supervision, software, writing – reviewing and editing.
Mary Murphy - data curation, methodology, project administration, supervision, writing – reviewing and editing.
Evangelos B Mazomenos – Conceptualisation, data curation, formal analysis, methodology, project administration, software, supervision, writing – reviewing and editing.
Hani J Marcus – Conceptualisation, data curation, formal analysis, methodology, project administration, software, supervision, writing – reviewing and editing.

Abstract

Objective: This study aimed to compare the ability of a deep-learning platform (the MACSSwin-T model) with healthcare professionals in detecting cerebral aneurysms from operative videos. Secondly, we aimed to compare the neurosurgical team's ability to detect cerebral aneurysms with and without AI-assistance.

Background: Modern microscopic surgery enables the capture of operative video data on an unforeseen scale. Advances in computer vision, a branch of artificial intelligence (AI), have enabled automated analysis of operative video. These advances are likely to benefit clinicians, healthcare systems, and patients alike, yet such benefits are yet to be realised.

Methods: In a cross-sectional comparative study, neurosurgeons, anaesthetists, and operating room (OR) nurses, all at varying stages of training and experience, reviewed still frames of aneurysm clipping operations and labelled frames as 'aneurysm not in frame' or 'aneurysm in frame'. Frames then underwent analysis by the AI platform. A second round of data collection was performed whereby the neurosurgical team had AI-assistance. Accuracy of aneurysm detection was calculated for human only, AI only, and AI-assisted human groups.

Results: 5,154 individual frame reviews were collated from 338 healthcare professionals. Healthcare professionals correctly labelled 70% of frames without AI assistance, compared to 78% with AI-assistance (OR 1.77, $p<0.001$). Neurosurgical Attendings showed the greatest improvement, from 77% to 92% correct predictions with AI-assistance (OR 4.24, $p=0.003$).

Conclusion: AI-assisted human performance surpassed both human and AI alone. Notably, across healthcare professionals surveyed, frame accuracy improved across all subspecialties and experience levels, particularly among the most experienced healthcare professionals. These results challenge the prevailing notion that AI

primarily benefits junior clinicians, highlighting its crucial role throughout the surgical hierarchy as an essential component of modern surgical practice.

Introduction

Artificial Intelligence (AI) offers novel solutions to surgical problems.[1,2] Computer vision, a domain of AI, enables images and videos to be analysed and contextualised by computer technology.[1] Advances in computer vision analysis of radiographic and diagnostic imaging modalities have been significant in the past two decades, yet these benefits are yet to be translated to the operating room (OR).[3] Increasingly, however, intraoperative video is being viewed as a mass of untapped data with huge potential. Indeed, recent advancements in intraoperative CV include phase recognition[4,5], navigation[6], and instrument segmentation[7]. In this study, we aimed to demonstrate the benefit of surgical computer vision, using microsurgical clipping of cerebral aneurysms as an exemplar.

Surgical clipping of cerebral aneurysms is a high-risk procedure, with 30% of operations experiencing complications and 36% of poor outcomes attributable to intraoperative issues.[8,9] The most feared intraoperative complication is aneurysm rupture, occurring in 17% of cases.[8] Crucially, the greatest risk of aneurysm rupture occurs when the aneurysm is in the surgical field of view.[8] Intraoperative identification of aneurysms can be challenging owing to narrow surgical corridors, anatomical variation, surrounding vasculature, and dense arachnoid adhesions. Indeed, interpreting three-dimensional anatomy in real-time and identification of small cerebral aneurysms feature amongst the top technical challenges faced by neurovascular surgeons.[10] Recognition of cerebral aneurysms is not only relevant to the operating surgeon, but also to the wider theatre team. Effective teamwork amongst the surgical team is essential to reducing the risk of intraoperative complications. Intraoperative cohesion and understanding improves efficiency, reduces stress burden, and facilitates rapid solving of intraoperative problems in what are frequently dynamic and high stress scenarios.[11,12] Indeed, communication breakdown and misunderstanding between surgical team members have been shown to contribute directly to adverse events.[11,13] The notion of ‘shared mental models’, in which a team share a collective understanding of a situation, helps elucidate why group understanding is paramount to intraoperative safety.[14,15] Essential to this shared experience is scene recognition. Put simply, it is important that all personnel in theatre environments understand when the highest risk phase of an operation is occurring, such as the aneurysm clipping phase. Recognition enables preparedness and a heightened alertness amongst the surgical team and safeguards against unnecessary distraction, akin to the ‘sterile cockpit’ protocol adopted in the aviation industry.[16,17] Innovations to increase scene recognition and awareness, therefore, stand to benefit patient safety.

Previously, our group described MACSSwin-T, a deep-learning platform able to detect or exclude the presence of cerebral aneurysms from microsurgical clipping operation operative video, exploiting computer vision based on the Shifted-Windows Transformer architecture.[18] MACSSwin-T achieved an accuracy of 80.8% (precision

51.3% and recall 63.8%) and an average F1-score of 56.8% in multiple cross-fold validation.[18] An optimised model was produced, and in an initial expert validation assessment demonstrated non-inferiority when compared to a cohort of ten attending (consultant grade) neurosurgeons.[18]

Surgical technologies benefit from stepwise evaluation, such as described by the Idea, Development, Exploration, Assessment, Long Term (IDEAL) Framework.[19] Prior to first-in-human studies, a range of factors pertaining to the safety and efficacy of an innovation should be explored. This comparative, ex- vivo (IDEAL Stage 0) study builds on our previous work by comparing the efficacy of the MACSSwin-T platform against neurosurgical healthcare professionals in detecting cerebral aneurysms from microsurgical aneurysm clipping operations. This study aimed to validate our intervention by comparing the MACSSwin-T platform with neurosurgical healthcare professionals at identifying cerebral aneurysms in microsurgical clipping operations. Secondly, we aimed to compare neurosurgical healthcare professional's ability to detect cerebral aneurysms with and without AI-assistance. In doing so, we aim to demonstrate the benefits of computer vision in surgical contexts.

Methods

Overview of methods

An online survey was created and distributed to neurosurgical healthcare professionals (neurosurgeons, anaesthetists, and operating room (OR) nurses) worldwide, all at varying stages of training and experience. Participants reviewed 15 still frames (seven containing an aneurysm, eight without an aneurysm) derived from four aneurysm clipping videos and determined if each frame contained an aneurysm. The frames were analysed by the MACSSwin-T platform, which predicted whether the frames did or did not contain an aneurysm. Human and AI performance results were compared. In a second version of the survey, the neurosurgical healthcare professionals received AI assistance, and initial human performance was compared to AI-assisted performance. Survey methodology adhered to Good Survey Practice guidelines[20] and have been reported in-keeping with CROSS guidelines.[21] This comparative validation study represents an IDEAL Stage 0 study, according to the IDEAL framework for evaluating innovations.[19] At time of publication, TRIPOD-AI reporting guidelines were not published - in the absence of these, this study has adhered to TRIPOD guidelines where applicable.[22] All patients provided written informed consent for the research video recordings, adhering to General Medical Council guidelines; videos and images were anonymised, and formal governance approval was obtained (Registration Reference 85-202021-SE).

Model Development: Microsurgical Aneurysm Clipping Surgery (MACS) Dataset and the MACSSwin-T Platform

Zhou et al.[18] created a dataset of 16 aneurysm clipping operative videos with expert-labelled annotations, which was used to train a deep-learning architecture for detecting cerebral aneurysms. Frames were extracted from 16 operative videos at a rate of five frames per second (fps), resulting in a dataset of 356,165 images (the Training Dataset – publicly available at available online: https://doi.org/10.5522/04/23533731).[23] All frames were labelled by experts as containing (n=71,113 frames) or not containing an aneurysm (n=285,052), and labelled frames were used to train a deep-learning platform (MACSSwin-T, details of which can be found in the original publication[18]) in a supervised learning phase. The MACSSwin-T platform utilises a shifted-windows (Swin-T) transformer architecture to classify frames. MACSSwin-T is based on hierarchical, multi-scale self-attention which allows the generation of localised features from multiple frames. These were then aggregated enabling the platform to detect and distinguish the aneurysm from similar-looking adjacent vasculature. For clarity, the MACSSwin-T platform's function was exclusively to identify cerebral aneurysms from operative video; the platform had no bearing on patient selection for treatment. The MACSSwin-T model underwent four-fold cross validation in a 12:4 training:test split, achieving an accuracy of 80.8% (precision 51.3% and recall 63.8%). In an initial expert validation assessment, the platform demonstrated non-inferiority when compared to a cohort of ten attending neurosurgeons.[18]

Survey Development:

Frames of operative aneurysm clippings used in the online survey were taken from a secure database of operative video recordings (1280x1080 pixels) of four patients obtained at a single tertiary-academic centre in the UK between 2020 – 2021. Videos were derived from elective and emergency cases. No criteria were applied regarding location or morphology of aneurysm, or the type of surgical clip applied. Operative videos were recorded direct to a ZEISS Kinevo 900 operating microscope (Carl Zeiss Co, Oberkochen, Germany), or a ZEISS OPMI Pentero 800 operating microscope (Carl Zeiss Co, Oberkochen, Germany).

Frames were extracted from the four operative videos (unseen to the MACSSwin-T platform) at a rate of 5 fps, resulting in 96,129 total frames (the Validation Dataset). A random number generator (Random Number Generator; randomwordgenerator.com/number) was used to select 50 frames. Ground truth was established through blinded review in duplicate by two vascular neurosurgeons where frames were classified as 'Aneurysm-Present', 'Aneurysm-Absent', or 'Exclude'(Figure 1). Reasons for frame exclusion were:

i. microscope not pointing at patient,
ii. microscope moving,
iii. indocyanine green angiography in process,
iv. ambiguous image with partial view of the aneurysm making it inconclusive to assign either X or Y label,
v. instruments crossing the field-of-view,

vi. rapid changing view within the scene.

****Figure 1:** Examples of 'Aneurysm-Present' Frames (Frames A, C, E) and 'Aneurysm-Absent' Frames (Frames B, D, F) Frames**

To ensure accurate ground truth during frame reviews, at least one of the reviewers had been present during the operation from which the frames were derived, and reviewers had access to the operative videos. Frames with conflicting labels (6/50) were excluded. A final dataset of 15 frames was selected from those with concordant reviews, to include eight 'Aneurysm-Absent' frames and seven 'Aneurysm-Present' frames (all frames can be found in Supplemental Digital Content 2, http://links.lww.com/SLA/F340). These 15 frames were used to create an online survey using SurveyMonkey (Momentive Inc., San Mateo, California, USA).

Data collection: Predictions from Healthcare Professionals and MACSSwin-T Analysis

Human performance in aneurysm detection was assessed using an online survey, comprising an initial set of demographic questions, followed by the 15 still frames. Participants were asked to answer whether each frame did or did not contain a cerebral aneurysm in view, in a "Yes" or "No" format. All participants reviewed the same 15 frames, in the same (randomised) order. Time to completion of the survey was recorded. Data collection was conducted over a four-week
period in September 2022. Participants were neurosurgical healthcare professionals defined as neurosurgeons and anaesthetists of senior (Attending/Consultant) or junior (Trainee/Resident/Fellow) grade, and OR nurses. Eligibility criteria for study participants included fulfilling one of these roles and working at a neurosurgical centre. No criteria was applied regarding neurosurgical subspecialty of practise. Participants were recruited from multiple centres internationally by local study collaborators. Collaborators were issued an information and instruction sheet including details of the research project, the host institution, and a QR code with a link to the survey.

Following collection of responses from the neurosurgical team, the 15 frames were analysed by the MACSSwin-T platform. The platform binarily classified each image as containing an aneurysm or not. Finally, a second independent round (Round Two) of data collection was performed during a four-week period in June 2023, using the same 15 frames, yet this time neurosurgical healthcare professionals were given the MACSSwin-T platform's predictions for each frame, along with the Gradient-weighted (GRAD) Cam class activation map for each frame. Human participants were informed that the platform has demonstrated an 81% accuracy in detecting cerebral aneurysms and were advised to use this to inform their decision making.

Exclusion Criteria

Incomplete survey responses (defined as <50% complete) were excluded, as were responses suspected to be falsified data.[24,25] Quantitative steps can be employed to identify such data, and established methodologies such as those described by Hernandez et al. were employed to ensure integrity of the dataset (Supplemental Digital Content 1, http://links.lww.com/SLA/F340).[24,25]

Data Analysis

Data were analysed using Microsoft Excel (Microsoft Corporation, USA), GraphPad (GraphPad Software, Inc.), and R (R Foundation for Statistical Computing, Vienna, Austria). Standard definitions were used for accuracy, precision, recall, and F1 score.[4] Comparative analysis was performed between human only and human with AI-assistance groups. Hierarchical mixed-effects regressions were employed for comparative analysis between groups with and without AI-assistance, that enabled accounting for confounders and within-cluster correlations.[26] The experimental group was set as the fixed-effect, and the random-effects were frame nested within video, occupation, and trainee/expert status. For binary outcomes, mixed-effects logistic regressions were conducted. For time, mixed-effected gaussian regressions with log link functions were conducted. Two-sided confidence intervals and p-values were calculated. P-values are reported to two significant figures up to values <0.001.[27] The type-1 error rate was set at $\alpha<0.05$, with Benjamini-Hochberge adjustments for multiple comparisons in any post-hoc testing.[28] Difficult and discrimination index for each frame were calculated and reported in Supplemental Digital Content 3, http://links.lww.com/SLA/F340. For time to completion data, surveys that were recorded to have taken more than 60 minutes were removed, with the reasoning that these were likely interrupted sessions on surveys that are not likely to take more than an hour otherwise.

Results

MACSSwin-T Performance

The MACSSwin-T platform binarily classified each frame as containing an aneurysm or not based on its predictive modelling. Results can be seen in the confusion matrix below (Table 1). The platform correctly predicted 87% (13/15) of all frames; 100% (8/8) of 'Aneurysm-Absent' frames, and 71% (5/7) of 'Aneurysm-Present' frames, giving an accuracy of 87%, precision of 100% and recall of 71%, and an F1 score of 83%.

****Table 1****

Neurosurgical Team Performance: Round One (no AI assistance)

Round One (no AI assistance) included 230 responses after exclusions, comprising 3396 individual frame reviews. Complete responses (i.e., reviewed all 15 frames) were submitted by 208/230 participants; 22/230

submitted incomplete surveys, but with >50% frames completed thus eligible for inclusion. The survey was completed by 88 neurosurgeons (38 senior grade, 50 junior), 77 anaesthetists (38 senior grade, 39 junior), and 65 OR nurses. Baseline characteristics of respondents can be found in Table 2.

Analysing responses from all healthcare professional respondents reveals an accuracy of 70% (2370/3396), specificity of 70% (2370/3396), sensitivity of 75% (1191/1587), positive predictive value of 65% (1191/1821), and negative predictive value of 75% (1179/1575).

Results by specialty are shown in Table 3. Neurosurgeons demonstrated an accuracy of 76% (993/1303), compared to 67% (753/1127) for anaesthetists, and 65% (624/966) for OR Nurse frame reviews. There were significant differences in accuracy between the groups on ANOVA (p=0.005), with post-hoc pairwise testing with Bejamini-Hochberg adjustment for false comparisons showing significant difference between neurosurgeons vs anaesthetists or OR nurses, but no significant difference between anaesthetists and OR nurses. Respondents were non-significantly better at identifying 'Aneurysm-Present' frames than 'Aneurysm-Absent' frames (Neurosurgeons 81% vs 72%; Anaesthetists 69% vs 65%; OR Nurses 74% vs 57%) (p = 0.09).

****Table 2****

****Table 3****

Neurosurgical Team Performance: Round Two (AI assisted)

The dataset for Round Two (AI assisted) included 118 responses, comprising 1758 individual frame reviews. Complete responses (i.e., reviewed all 15 frames) were submitted by 114/118 participants, 4/118 submitted incomplete surveys, but with >50% frames completed thus eligible for inclusion. Baseline characteristics for respondents can be found in Table 2. Analysing responses from all respondents reveals an accuracy of 78% (1370/1758), which was statistically significantly greater than for Round One (without AI-assistance) (70% (2370/3396) (Odds Ratio (OR) 1.77, CI 1.44–2.17, p<0.001). Specificity was 77% (719/939), sensitivity was 80% (651/819), positive predictive value was 75% (652/872), negative predictive value was 81% (719/887). Results by specialty are shown in Table 3. Accuracy in frame prediction increased with AI-assistance for neurosurgeons (76% correct in Round 1 vs 88% in Round 2) (OR 2.66, CI 1.62-4.39, p<0.001)), anaesthetists (67% vs 77%) (OR 1.75, CI 1.32-2.34), p<0.001)), and non-significantly with OR nurses (65% vs 70%) (OR 1.34, CI 0.98-1.84, p=0.066)). Difficult and discrimination index for each frame is reported in Supplemental Digital Content 3, http://links.lww.com/SLA/F340.

Time to Survey Completion

Median time to complete the survey was 5.8 minutes (IQR 4.5 - 8.4) in Round One (no AI-assistance), and 5.5 minutes (IQR 3.5 – 7.9) in Round Two (AI-assisted) (P=0.16). Neurosurgeons were non-significantly quicker to complete the survey with AI-assistance (Round Two time 5.0 minutes (IQR 3.7 – 6.5)) than without AI-assistance (Round One time 6.1 minutes (IQR 4.5 – 8.8)) (p=0.26). This improvement was driven by Attending neurosurgeons, who's time to completion near-significantly improved with AI-assistance (Round One time 6.2 minutes (IQR 4.9 – 8.8) vs Round Two time 4.0 minutes (IQR 3.3 – 5.6) (p= 0.060). No significant difference in timing was noted for anaesthetists or OR nurses when comparing AI versus no AI-assistance.

Discussion

Principal Findings

This study presents the key findings from a comparative study comparing the accuracy of a deep-learning platform and neurosurgical healthcare professionals (with and without AI-assistance) in the detection of cerebral aneurysms from microsurgical aneurysm clipping operations.

First, our data demonstrate that AI-assisted human performance is superior to both human performance and AI performance alone. In this study, AI-assistance was provided through offering the MACSSwin-T model's prediction along with frame activation-maps. Improvements in accuracy were noted for all three occupations – neurosurgeons improved from 76% to 88%, anaesthetists 67% to 77%, and OR Nurses 65% to 70%. These results serve as a quantitative demonstration of the breakdown in shared mental models within the operating theatre, and support the use of AI technology to enhance collaborative orientation amongst the neurosurgical team.[29] Breakdown in shared understanding is particularly pertinent in neurosurgery, a high-stress, high-risk specialty in which communication breakdown influences patient outcomes and risk of litigation.[30]

Second, and perhaps most significant – our findings suggest that the greatest benefits were observed in the most experienced healthcare professionals, neurosurgical Attendings, whose frame accuracy significantly improved, and time-to-completion near-significantly benefited. This finding purports the value of clinician-AI collaboration in improving surgical safety, and contradicts assertions that the benefits of AI-assistance are confined to junior clinicians.[31] The reasons for this are intriguing. One possible explanation pertains to the adage *'the more you know, the more you see'*. In 1993, Ericsson et al. published their seminal paper on the attainment of expert performance through deliberate practise[32]; a key upshot of expert performance is the ability to interpret and understand more with a given dataset or situation than non-experts. In the case of small, obscured, and challenging aneurysms, AI-assistance may provide experts with the necessary nudge to confirm their suspicion that an aneurysm is present in the field of view.

Finally, we demonstrate that the MACSSwin-T platform outperformed humans (no AI-assistance), and showed a high precision and recall in keeping with our previous work.[18] Analysis of frame-level discrepancies between the MACSSwin-T platform and human performance reveals interesting findings. The platform was highly accurate in its identification of 'Aneurysm-Absent' frames (100%), in part due to the imbalance in 'Aneurysm-Absent' to 'Aneurysm-Present' frames in the initial dataset (80% vs 20%), as a result of aneurysm exposure making up only a small phase of the operation.[18] Yet, with inverse-proportional weighting of the two classes in the training loss function and adjustment of the model's decision threshold, the platform was optimised to detect 'Aneurysm-Present' frames without increasing erroneous 'Aneurysm-Absent' detections (i.e. false positives). This was apt in cases where the aneurysm was partially obscured, where humans found identification troublesome, but the platform was able to accurately locate the aneurysm (Figure 2, frame C). Figure 2 shows challenging frames for MACSSwin-T and the neurosurgical team, along with activation maps demonstrating the focus point of the platform. 'Aneurysm-Absent' frames were particularly challenging for the neurosurgical team. Notably, correct identification of 'Aneurysm-Absent' frames was increased with AI-assistance, from 65% to 77% ($p < 0.001$).

****Figure 2: challenging frames for MACSSwin-T and the neurosurgical team with associated MACSSwin-T activation maps:** Frames (A) and (B) are 'Aneurysm-Present' frames that the MACSSwin-T platform incorrectly classified as 'Aneurysm-Absent' frames. Frame (A) is a complex operative scene with numerous instruments and vessels in frame, as well as the aneurysm partially obscured by Cottonoids, which may account for the incorrect classification by MACSSwin-T. Human performance was also poor for this frame, with 57% of the neurosurgical team correctly classifying the frame. Frame (B) shows the atherosclerosed aneurysm dome in clear view – this variation in appearance may account for the incorrect platform classification, as the activation map shows little focus on the white dome; in contrast, 85% of neurosurgeons correctly identified the aneurysm. Frame (C) is a 'Aneurysm-Present' frame correctly classified by the MACSSwin-T platform, yet was the frame on which human performance was worst, with 51% of respondents correctly identifying the frame as containing as an aneurysm. Frame (D) was amongst the most commonly correctly classified frames by human assessors, with 84% of the neurosurgical team correctly labelling the 'Aneurysm-Present' frame (the frame was also correctly classified by MACSSwin-T). Only two other frames had a greater correct percentage, both of which showed contained an aneurysm clip around the aneurysm.**

Findings in the Context of the Literature

Current AI applications in the field of aneurysm detection primarily rely on radiomic-based methods, using imaging modalities such as CT, MRI, or Angiography.[33] In this study, we present a novel approach to intraoperative aneurysm detection. While intraoperative computer vision has been successfully applied in

several surgical contexts[3,3,34,35], its utilisation in neurosurgery is emerging. Previous works by Pangal et al. and Staartjes et al. have demonstrated the feasibility of using deep-learning platforms for real-time anatomy segmentation in endoscopic endonasal surgery, in cadaveric and in-vivo settings respectively.[36,37] Additionally, Das et al. introduced the PAINet model for anatomical structure identification during pituitary surgery, achieving 66% accuracy in sella identification.[38] Another significant contribution in neurosurgery comes from Choi et al., who employed a YOLACT-based architecture for anatomic detection in mastoidectomy.[39] Our study adds to this growing body of research, expanding the potential applications of AI in neurosurgical procedures. Further, this study showcases the feasibility of an attention-based learning architecture (the shifted-window Transformer model) in anatomic detection, distinguishing it from the prevailing use of convolutional neural networks (e.g. ENet, PSPnet, UNet, SegNet, YOLO, and ErfNET).[34,39,40] This substantiates an alternative, hierarchical strategy for developing novel AI architectures to tackle contemporary challenges in surgical contexts. Rapid, real-time anatomic recognition of intracranial aneurysms represents a leap in AI-healthcare capabilities, with numerous potential benefits in decision support, system efficiency, and education.

Strengths and Limitations

This study has strengths in its international scope, adherence to established frameworks[29,34,3521], and small-scale pre-clinical validation in a subset of neurosurgical attendings prior to wider distribution[18]. Limitations include training on a limited dataset from a single institution, reducing generalisability and enhancing risk of overfitting. Participants were asked to binarily classify frames as ‘Aneurysm-Absent’ or ‘Aneurysm-Present’, whereas surgeons typically take a probabilistic approach to a live, three-dimensional situation. Some images were of low-resolution, reflecting the challenge in the operating room. Some participants commented that anatomy is more readily identifiable when benefiting from binocular disparity, enabling a three-dimensional view through the microscope. This is undoubtedly true, and we expect that neurosurgeons would score higher if able to view the operative scenes through a microscope and interact with the surgical environment. This does not, however, impact the ability of anaesthetists or OR nurses who rely wholly on the microscope monitor. When selecting frames for the survey, frames with conflicting consultant labels were excluded, which may bias the selected frames towards easier instances. Data collection was conducted in the same centres for Round One and Two, raising potential for participants who had already completed the survey in Round One repeating it. To minimise this potential for bias, participants were not shown which frames they scored correctly/incorrectly after Round One, participants were asked if they had completed the survey before, and Round One and Two were conducted nine months apart. Due to anonymity requirements, we could not link individual scores between survey rounds, but we conducted a sensitivity analysis by accounting for whether participants had previously undertaken the survey as a random effect, which revealed no differences in the study’s results. Survey distribution in English may have introduced selection bias.

Conclusion

This IDEAL stage 0 study compared a novel deep-learning computer vision platform (MACSSwin-T) with neurosurgical healthcare professionals in identifying cerebral aneurysms from microsurgical clipping operation images. Our data demonstrate that AI-assisted human performance is superior human performance without AI-assistance. Senior neurosurgeons benefited the most from AI-assisted aneurysm detection, with improved frame accuracy and time-to-completion. This research contradicts the prevailing narrative within the AI-healthcare paradigm, which asserts that the benefits of AI-assistance are most notable in junior clinicians. Future research in this area should focus upon model architecture iteration prior to first-in-human validation, in accordance with IDEAL Stage 0 & 1 evaluation.

Competing Interests:

Author DS is an employee of Digital Surgery, Medtronic. HJM holds a patent in video image processing (International Publication Number WO 2023/017230 A1). All other authors declare no financial or non-financial competing interests.

Acknowledgements

We would like to thank our international collaborators for their assistance with data collection. Collaborators are listed below:

Alshaymaa Mortada Ali - Alshaymaa Ali, Ain shams university, Cairo, Egypt; Daniel Stephen Masunga - Faculty of Medicine, Kilimanjaro Christian Medical University College, Moshi-Kilimanjaro, Tanzania; Oliver Burton - Royal Victoria Infirmary, Newcastle Upon Tyne, United Kingdom; R.M.G.K Rathnayaka - Postgraduate Institute Of Medicine, University of Colombo, Sri Lanka; Rohan Bhate - St. George's, University of London, London, United Kingdom; Samar Adel Abdel Rahim - Faculty of Medicine Zagazig University, AlSharkia, Egypt; Shankhaneel Ghosh - Institute of Medical Sciences and SUM Hospital, Bhubaneswar, India; Sufyan Ibrahim - Dept of Neurologic Surgery, Mayo Clinic, Rochester, Minnesota, United States of America; Suheda Yavuz - Sağlık Bilimleri Üniversitesi (SBÜ) Gülhane Tıp Fakültesi, Ankara, Turkey; Tarik Sarıdoğan - Bursa Uludag University, Bursa, Turkey; Razan Eid, Ankara Yildirim Beyazit University, Ankara, Turkey.

Data availability

Our Training Dataset is publicly available at available online: https://doi.org/10.5522/04/23533731). The datasets generated and/or analysed during the current study are not publicly available but are available from the corresponding author on reasonable request.

References

1. Hashimoto DA, Rosman G, Rus D, et al. Artificial Intelligence in Surgery: Promises and Perils. *Ann Surg*. 2018;268:70–76.

2. Garrow CR, Kowalewski K-F, Li L, et al. Machine Learning for Surgical Phase Recognition: A Systematic Review. *Ann Surg*. 2021;273:684–693.

3. Gong J, Holsinger FC, Noel JE, et al. Using deep learning to identify the recurrent laryngeal nerve during thyroidectomy. *Sci Rep*. 2021;11:14306.

4. Khan DZ, Luengo I, Barbarisi S, et al. Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: Development and preclinical evaluation (IDEAL stage 0). *JNS*.;In press.

5. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer Vision Analysis of Intraoperative Video: Automated Recognition of Operative Steps in Laparoscopic Sleeve Gastrectomy. *Ann Surg*. 2019;270:414–421.

6. Laplante S, Namazi B, Kiani P, et al. Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy. *Surg Endosc*. . Epub ahead of print August 2, 2022. DOI: 10.1007/s00464-022-09439-9.

7. Shvets A, Rakhlin A, Kalinin AA, et al. Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA):624–628.

8. Muirhead WR, Grover PJ, Toma AK, et al. Adverse intraoperative events during surgical repair of ruptured cerebral aneurysms: a systematic review. *Neurosurg Rev*. 2021;44:1273–1285.

9. Fridriksson S, Säveland H, Jakobsson K-E, et al. Intraoperative complications in aneurysm surgery: a prospective national study. *J Neurosurg*. 2002;96:515–522.

10. Muirhead WR, Layard Horsfall H, Khan DZ, et al. Microsurgery for intracranial aneurysms: A qualitative survey on technical challenges and technological solutions. *Front Surg*.;9 Available from: https://www.frontiersin.org/articles/10.3389/fsurg.2022.957450. 2022. Accessed July 19, 2023.

11. Greenberg CC, Regenbogen SE, Studdert DM, et al. Patterns of communication breakdowns resulting in injury to surgical patients. *J Am Coll Surg*. 2007;204:533–540.

12. Westli HK, Johnsen BH, Eid J, et al. Teamwork skills, shared mental models, and performance in simulated trauma teams: an independent group design. *Scand J Trauma Resusc Emerg Med*. 2010;18:47.

13. Mazzocco K, Petitti DB, Fong KT, et al. Surgical team behaviors and patient outcomes. *Am J Surg*. 2009;197:678–685.

14. Nakarada-Kordic I, Weller JM, Webster CS, et al. Assessing the similarity of mental models of operating room team members and implications for patient safety: a prospective, replicated study. *BMC Med Educ*. 2016;16:229.

15. Gjeraa K, Dieckmann P, Spanager L, et al. Exploring Shared Mental Models of Surgical Teams in Video-Assisted Thoracoscopic Surgery Lobectomy. *Ann Thorac Surg*. 2019;107:954–961.

16. Broom MA, Capek AL, Carachi P, et al. Critical phase distractions in anaesthesia and the sterile cockpit concept. *Anaesthesia*. 2011;66:175–179.

17. Wadhera RK, Parker SH, Burkhart HM, et al. Is the “sterile cockpit” concept applicable to cardiovascular surgery critical intervals or critical events? The impact of protocol-driven communication during cardiopulmonary bypass. *J Thorac Cardiovasc Surg*. 2010;139:312–319.

18. Zhou J, Muirhead W, Williams SC, et al. Shifted-windows transformers for the detection of cerebral aneurysms in microsurgery. *Int J Comput Assist Radiol Surg*. 2023;18:1033–1041.

19. McCulloch P, Altman DG, Campbell WB, et al. No surgical innovation without evaluation: the IDEAL recommendations. *The Lancet*. 2009;374:1105–1112.

20. KELLEY K, CLARK B, BROWN V, et al. Good practice in the conduct and reporting of survey research. *Int J Qual Health Care*. 2003;15:261–266.

21. Sharma A, Minh Duc NT, Luu Lam Thang T, et al. A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). *J Gen Intern Med*. 2021;36:3179–3187.

22. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13:1.

23. Microsurgical Aneurysm Clipping Surgery (MACS) Dataset with image-level aneurysm presence/absence annotations . Epub ahead of print July 6, 2023. DOI: 10.5522/04/23533731.v1.

24. Hernandez I, Ristow T, Hauenstein M. Curbing curbstoning: Distributional methods to detect survey data fabrication by third-parties. *Psychol Methods*. 2022;27:99–120.

25. Seriousness checks are useful to improve data validity in online research | SpringerLink Available from: https://link.springer.com/article/10.3758/s13428-012-0265-2. Accessed November 8, 2022.

26. Brooks ME, Kristensen K, Benthem KJ van, et al. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *R J*. 2017;9:378–400.

27. Aguinis H, Vassar M, Wayant C. On reporting and interpreting statistical significance and p values in medical research. *BMJ Evid-Based Med*. 2021;26:39–42.

28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Methodol*. 1995;57:289–300.

29. Cannon-Bowers JA, Salas E, Converse S. Shared mental models in expert team decision making. In: Individual and group decision making: Current issues. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; 1993:221–246.

30. Hartley BR, Hong C, Elowitz E. Communication in Neurosurgery—The Tower of Babel. *World Neurosurg*. 2020;133:457–465.

31. Shen J, Zhang CJP, Jiang B, et al. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Med Inform*. 2019;7:e10010.

32. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev*. 1993;100:363–406.

33. M D, S A, M G, et al. Detection of cerebral aneurysms using artificial intelligence: a systematic review and meta-analysis. *J Neurointerventional Surg*.;15 . Epub ahead of print March 2023. DOI: 10.1136/jnis-2022-019456.

34. Madani A, Namazi B, Altieri MS, et al. Artificial Intelligence for Intraoperative Guidance: Using Semantic Segmentation to Identify Surgical Anatomy During Laparoscopic Cholecystectomy. *Ann Surg*. 2022;276:363–369.

35. Tokuyasu T, Iwashita Y, Matsunobu Y, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg Endosc*. 2021;35:1651–1658.

36. Dj P, G K, S S, et al. A Guide to Annotation of Neurosurgical Intraoperative Video for Machine Learning Analysis and Computer Vision. *World Neurosurg*.;150 . Epub ahead of print June 2021. DOI: 10.1016/j.wneu.2021.03.022.

37. Ve S, A V, L R, et al. Machine Vision for Real-Time Intraoperative Anatomic Guidance: A Proof-of-Concept Study in Endoscopic Pituitary Surgery. *Oper Neurosurg Hagerstown Md*.;21 . Epub ahead of print September 15, 2021. DOI: 10.1093/ons/opab187.

38. Das, Adrito, Khan, Danyal Z (second), Williams, Simon C (third), et al. A Multi-Task Network for Anatomy Identification in Endoscopic Pituitary Surgery. *MICCAI Proceedings*.;In Press.

39. Choi J, Cho S, Chung JW, et al. Video recognition of simple mastoidectomy using convolutional neural networks: Detection and segmentation of surgical tools and anatomical regions. *Comput Methods Programs Biomed*. 2021;208:106251.

40. Gumbs AA, Grasso V, Bourdel N, et al. The Advances in Computer Vision That Are Enabling More Autonomous Actions in Surgery: A Systematic Review of the Literature. *Sensors*. 2022;22:4918.

Table 1: Confusion Matrix for MACSSwin-T Model Performance on Fifteen Frames		
	Ground Truth Positive	Ground Truth Negative
Prediction Positive	5	0
Prediction Negative	2	8

Table 2. Baseline characteristics of respondents

	Round One (no AI assistance)	Round Two (AI assisted)
Total respondents (n =)	**230**	**118**
Gender (%)		
Male	54% (123/230)	45% (53/118)
Female	46% (106/230)	55% (65/118)
Prefer not to say	0.4% (1/230)	0% (0/118)
Non-binary	0% (0/230)	0% (0/118)
Age group (%)		
18 to 24	3% (7/230)	2% (2/118)
25 to 34	44% (102/230)	50% (59/118)
35 to 44	30% (68/230)	22% (26/118)
45 to 54	12% (28/230)	14% (16/118)
55 to 64	10% (22/230)	10% (12/118)
65 to 74	1% (3/230)	3% (3//118)
Occupation:		
Neurosurgeon	38% (88/230)	31% (36/118)
Consultant/Attending	*N = 38*	*N = 12*
Trainee/Resident/Fellow	*N = 50*	*N = 24*
Anaesthetist	33% (77/230)	31% (37/118)
Consultant/Attending	*N = 38*	*N = 18*
Trainee/Resident/Fellow	*N = 39*	*N = 24*
OR Nurse	28% (65/230)	34% (40/118)

Table 3: Percentage and number of correct frame reviews per specialty (%, N)
*OR = Odds Ratio; CI = Confidence Intervals; * denotes statistical significance.*

	All Frames			**'Aneurysm-Absent' Frames**			**'Aneurysm-Present' Frames**		
	Round One (no AI-assistance)	Round Two (AI assisted)	OR CI P value	Round One (no AI-assistance)	Round Two (AI assisted)	OR CI P value	Round One (no AI-assistance)	Round Two (AI assisted)	OR CI P value
Neurosurgeons	76% (993/1303)	88% (473/538)	2.66 (1.62-4.39) *<0.001	72% (500/694)	87% (251/288)	3.28 (1.70-6.31) *<0.001	81% (493/609)	89% (222/250)	2.22 (1.23-4.00) *0.0082
Consultant/Attending Grade	77% (438/566)	92% (166/180)	4.24 (1.63-11.1) *0.0031	73% (220/301)	91% (87/96)	5.12 (1.50-17.5) *0.0092	82% (218/265)	94% (79/84)	4.26 (1.25-14.5) *0.02
Trainee/Resident/fellow Grade	75% (555/737)	86% (307/358)	2.24 (1.26-3.98) *0.0056	71% (280/393)	85% (164/192)	1.03 (1.27-6.20) *0.010	80% (275/344)	86% (143/166)	1.74 (0.90-3.38) 0.1
Anaesthetists	67% (753/1127)	77% (475/620)	1.75 (1.32-2.34) *<0.001	65% (389/602)	75% (238/331)	1.84 (1.18-2.87) *0.0073	69% (364/525)	79% (227/289)	1.86 (1.19-2.90) *0.0067
Consultant/Attending Grade	69% (387/559)	75% (196/261)	1.43 (0.95-2.13) 0.08	69% (205/298)	78% (108/139)	1.72 (0.88-1.21) 0.11	70% (182/261)	72% (88/122)	1.22 (0.56-2.68) 0.62
Trainee/Resident/Fellow Grade	64% (366/568)	78% (279/359)	2.08 (1.41-3.06) *<0.001	61% (184/304)	73% (140/192)	1.96 (1.09-3.54) *0.025	69% (182/264)	83% (139/167)	2.50 (1.48-4.22) *<0.001
Operating Room Nurse	65% (624/966)	70% (422/600)	1.34 (0.98-1.84) 0.066	57% (290/513)	69% (220/320)	1.86 (1.17-2.96) *0.0088	74% (334/453)	72% (202/280)	0.91 (0.58-1.42) 0.67

Table 4: Difficulty Index and Discrimination Index for Individual Frames				
	Round One (no AI)		Round Two (AI assisted)	
	Difficulty Index	Discrimination Index	Difficulty Index	Discrimination Index
Image 1 (aneurysm absent)	0.70	0.34	0.78	0.13
Image 2 (aneurysm absent)	0.57	0.43	0.70	0.18
Image 3 (aneurysm present)	0.76	0.31	0.66	0.13
Image 4 (aneurysm absent)	0.66	0.33	0.68	0.21
Image 5 (aneurysm present)	0.79	0.25	0.81	0.13
Image 6 (aneurysm present)	0.51	0.05	0.65	0.18
Image 7 (aneurysm absent)	0.79	0.37	0.81	0.14
Image 8 (aneurysm present)	0.85	0.17	0.94	0.06
Image 9 (aneurysm present)	0.57	0.17	0.58	0.14
Image 10 (aneurysm absent)	0.54	0.26	0.73	0.17
Image 11 (aneurysm absent)	0.52	0.41	0.76	0.17
Image 12 (aneurysm present)	0.84	0.22	0.96	0.05
Image 13 (aneurysm absent)	0.65	0.33	0.81	0.11
Image 14 (aneurysm present)	0.93	0.17	0.97	0.03
Image 15 (aneurysm absent)	0.78	0.30	0.86	0.09

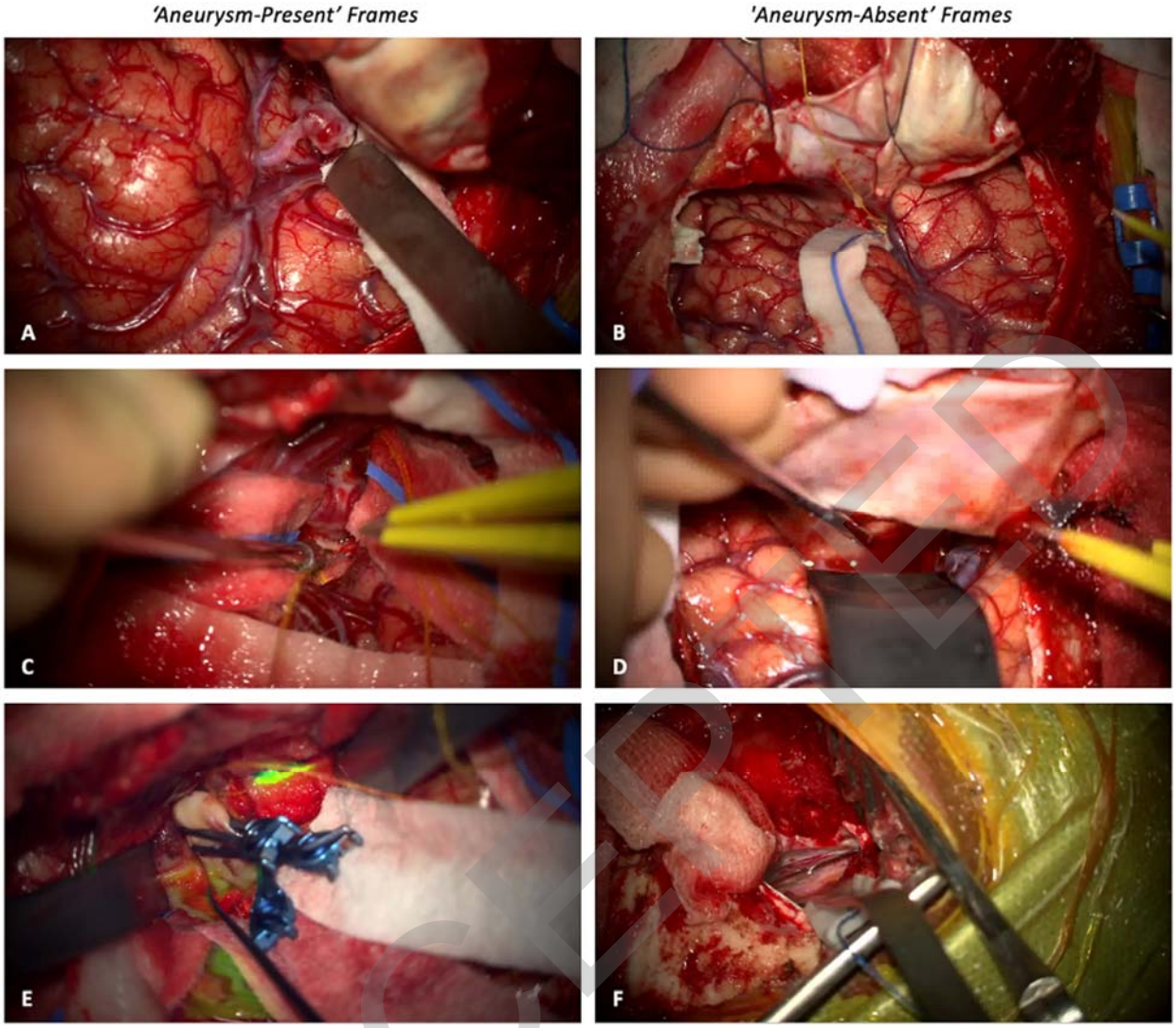

FIG 1

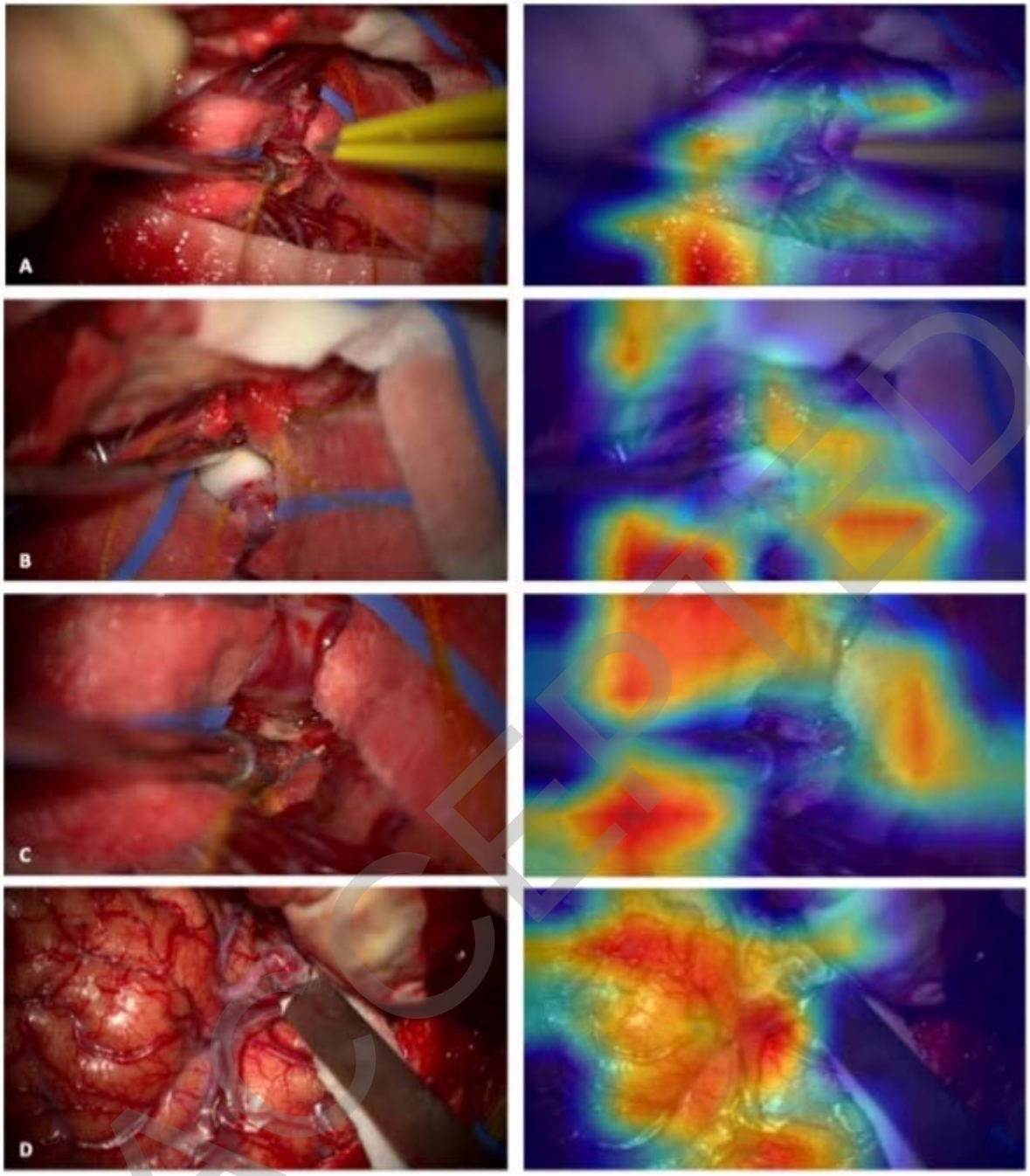

FIG 2

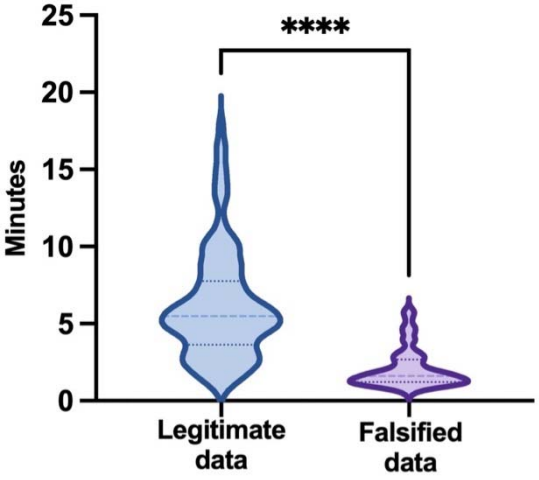

FIG 3