Langenbeck's Archives of Surgery (2025) 410:53
https://doi.org/10.1007/s00423-025-03626-7

BRIEF REPORT

Can surgeons trust AI? Perspectives on machine learning in surgery and the importance of eXplainable Artificial Intelligence (XAI)

Johanna M. Brandenburg[1,2] · Beat P. Müller-Stich[2,3] · Martin Wagner[4,5,6] · Mihaela van der Schaar[7,8,9,10]

Received: 1 October 2024 / Accepted: 21 January 2025

Abstract

Purpose This brief report aims to summarize and discuss the methodologies of eXplainable Artificial Intelligence (XAI) and their potential applications in surgery.

Methods We briefly introduce explainability methods, including global and individual explanatory features, methods for imaging data and time series, as well as similarity classification, and unraveled rules and laws.

Results Given the increasing interest in artificial intelligence within the surgical field, we emphasize the critical importance of transparency and interpretability in the outputs of applied models.

Conclusion Transparency and interpretability are essential for the effective integration of AI models into clinical practice.

Keywords Artificial intelligence · Explainable artificial intelligence · Machine learning · Minimally invasive surgery

✉ Martin Wagner
martin.wagner@ukdd.de

✉ Mihaela van der Schaar
mihaela@ee.ucla.edu

1 Department of General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany

2 National Center for Tumor Diseases (NCT), Heidelberg, Germany

3 University Digestive Healthcare Center Basel, Basel, Switzerland

4 Department of Visceral, Thoracic and Vascular Surgery, University Hospital Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany

5 National Center for Tumor Diseases (NCT/UCC), Dresden, Germany

6 Centre for Tactile Internet with Human-in-the-Loop (CeTI), Technische Universität Dresden, Dresden, Germany

7 University of Cambridge, Cambridge, UK

8 Department of Electrical and Computer Engineering, University of California– Los Angeles, Los Angeles, CA, USA

9 Alan Turing Institute, London, UK

10 Centre for Mathematical Imaging in Healthcare Machine Learning and Artificial Intelligence, Faculty of Mathematics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK

Benefit or burden - how do surgeons perceive AI?

It is inarguable that there is immense potential in artificial intelligence (AI) and its most prominent subfield: machine learning (ML). This applies to surgery as well as to medicine in general, with healthcare being a highly data-generating and data-driven field. Experts in digital surgery [1] are convinced that emerging technologies can enhance preoperative planning, provide navigation assistance - similar to autonomous driving - by, e.g., highlighting the critical view of safety in the surgical field of minimally invasive procedures [2], assess surgical skills [3], and offer decision support, such as predicting intra- and postoperative complications [4] and personalizing therapy recommendations for complex patient cases.

Surgeons undergo extensive professional training, accompanied by high workload and stress levels. During years of collecting clinical experience, they gradually gain expertise regarding effective decision-making and managing situations with high demands on technical, cognitive, and communicative skills. This is especially important in the high-stakes environment of the operating room [5]. The potential integration of AI can thus justifiably evoke different reactions among the surgical community: doubt when it comes to a machine influencing what the correct patient diagnosis or treatment should be, e.g., in (surgical)

Published online: 28 January 2025

oncology or the intensive care unit [6]; fear that at some point your technical qualities or ability of clinical reasoning will be replaced, e.g., in defining the critical view of safety in laparoscopic cholecystectomy [2]; or simply irritation because, e.g., automatic robot-assisted camera guidance may not precisely display the desired viewpoint.

The importance of eXplainable artificial intelligence (XAI)

Given the challenge of trust in AI systems and the limited concrete usage in healthcare, particularly in surgery [7], how can we foster the acceptance of ML methods and their integration into clinical practice? A major challenge is to make the ML models accessible to surgeons in such a way that the information is presented in a clear manner within an intuitive user interface, and their output being transparent and interpretable. Building trust and understanding is of utmost importance. As surgeons often find themselves in situations without much time to reconsider one's actions or question recommendations, they need to be able to confidently rely on newly introduced AI-based assistance systems. This does not necessarily mean that the predictions of underlying ML models must be perfectly accurate, but rather that the surgeon understands why a certain output is generated by providing human-interpretable information including uncertainties and ambiguities that are inherent to clinical decision-making anyway. However, this is a major methodological challenge because the multi-layer architecture of neural networks used as ML models for high problem-complexity hardly allows humans to fully understand the models' conclusions [8]. Researchers try to address this need by developing the field of ML model interpretability defined as the extent to which an ML model can be made understandable to relevant human users [9]. The term eXplainable AI (XAI) can thus be summarized as tailored interpretability for different users. It can be employed in many ways depending on the type of input data, the applied ML algorithm, and of course the requirements and questions of the users themselves [10]. To establish a surgical understanding of the capabilities of computer science in providing XAI, we present an overview of various interpretability methods with potential applications in surgery (Fig. 1).

Global explanatory patient features

A ML model could predict the overall survival of patients undergoing different surgeries such as an esophagectomy based on preoperative, static patient data such as age, gender, comorbidities, neoadjuvant therapy, or smoker status. However, surgeons might need to know which data play the most important role for this prediction to optimize further treatment pathways. The interpretability method of explanatory (patient) features points out these features and their contribution to the model's prediction.

Individual explanatory patient features

However, the described global approach of outcome prediction in clinical practice is often not sufficient when it comes to specific patient cases with questions arising such as: "What are the most important features for the survival prediction of this specific patient I'm going to operate on tomorrow?". The surgeon needs to know which subset of features is relevant for each patient meaning individualized feature importance [11], e.g., to build a basis for an informative patient discussion with specific recommendations.

Explanatory features for imaging data

Image and video data are of particular importance in surgery, e.g., for surgical planning using radiological data or the surgical video itself during minimally invasive procedures. When these images are analyzed by means of computer vision, feature importance can be applied in the context of XAI by highlighting the parts of the image/video that have the greatest influence on the output when changed, so-called integrated gradients [12]. When developing a model to assess completeness of lymphadenectomy in esophagectomy, it is crucial for the surgeon to know exactly on which structures or features the model bases its rating in certain areas. In general, XAI may be more approachable for imaging tasks, as the visual nature of predictions often aligns with human interpretability, facilitating the detection and assessment of potential biases in the algorithm.

Explanatory features taking time series into account

Especially in surgery with long and complex treatment pathways and the highly important surgical procedure itself, static features alone are insufficient for certain predictions. Interpretability approaches taking the temporal sequence of clinical processes into account [13] need to be addressed when, e.g., postoperative complications shall be predicted. Thereby, the risk estimation of developing anastomotic insufficiency after pancreatic surgery could be more effectively explained by assessing dynamic data such as intraoperative variation in heart rate and blood pressure of the patient during surgery, as well as the perioperative trends in laboratory values. Moreover, e.g., continuous monitoring of blood levels might enable conclusions about total blood loss, potential complications like vascular injuries, and correlations with specific surgical phases and steps. By

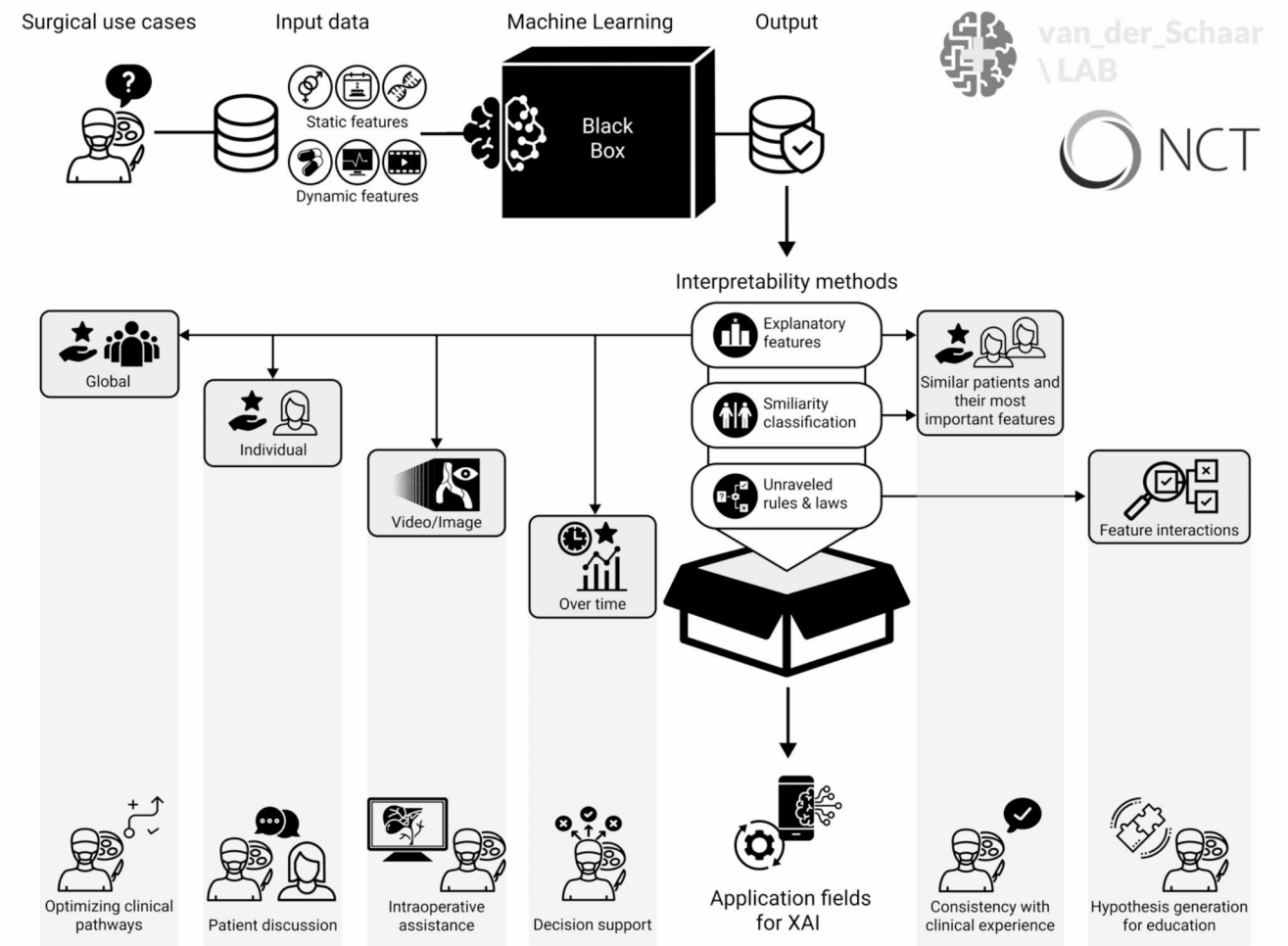

Fig. 1 Methods and clinical applications for eXplainable artificial intelligence in surgery

incorporating temporal data, clinicians gain a deeper understanding of physiological patterns and their relationship to surgical outcomes.

Similarity classification

In clinical practice, physicians learn a lot from comparing different but similar patient cases to each other, e.g., to find out whether their proposed intervention for the current patient is consistent with treated patients in the past. The interpretability method of similarity classification could derive similar patient cases when surgeons, e.g., must decide whether the risk of post hepatectomy liver failure after extended hemihepatectomy does or does not outweigh the potential benefits. Even more important may be information on the certainty a model has in its treatment recommendation for specific patients. XAI should thus further enable highlighting cases for which the model's prediction is uncertain, supporting the surgeon's right to be skeptical. Similarity classification could also help in this case, e.g., by being combined with individual explanatory patient features (see above). In this approach, similar patients and their most important features for the prediction for the actual patient are selected while showing if the model's prediction for similar patients has been true or false [14]. Based on this, the surgeon has more information on whether to follow the model's recommendation.

Unraveled rules and laws

When surgeons prepare themselves for future patients or revisit a specific patient case, "what if" scenarios often arise: "What if the tumor had not been so close to the aorta? What if the patient had been 5 years younger? Under these circumstances, might a different intervention have been recommended, or could the patient have avoided a postoperative complication?". These questions highlight the inherent complexity and variability in surgical decision-making. Scientists try to address these questions by uncovering previously unrevealed "rules" and "laws" in their models.

By leveraging advanced analytical techniques, they seek to identify variable interactions that enhance our understanding of complex clinical scenarios. This process enables hypothesis generation offering insights into alternative pathways and potential outcomes for diverse patient scenarios.

Surgeons should embrace eXplainable AI (XAI)

If assistance systems based on AI are to become an integral part of clinical care especially in surgery, explainability and the utilization of XAI are inevitable. While AI systems are not infallible and require continuous validation and optimization, one significant challenge lies in the inherent difficulty of explaining and comprehending the processes leading to a model's output. XAI plays a crucial role in addressing this challenge, enabling surgeons to provide optimal care to their patients with the support of AI-based assistance systems in their final decisions. Simultaneously, incorporating XAI can simplify the process for users to articulate their doubts about specific ML-based recommendations and provide feedback to the model, thereby fostering a continuous building of trust. Thus, ML interpretability embedded in decision-support systems should be able to learn which interpretability modules are the most important and trustworthy ones for individual clinicians, paving the way for personalized usage [9]. Concerning the challenge of developing an intuitive interface for digital AI tools in surgery, the emerging field of Large Language Models and Visual Language Models must be highlighted. With their ability to process and generate human-interpretable textual data, they could enable the integration of multimodal information, thereby facilitating the interaction of surgeons with conventional AI and explainability tools [15]. In terms of the rapidly developing field of artificial intelligence, surgeons should embrace XAI in surgery. Interpretable models need to be developed with methods taking the complex, dynamic and individual surgical patient cases into account and adapting to the user's knowledge. This, however, is only possible if surgeons and computer scientists interact closely to bridge the gap between clinical need and technological feasibility.

Author contributions All authors contributed to the conception of the manuscript. J.M.B. prepared the main manuscript text and the figure. M.v.d.S., M.W., and B.P.M.S edited and reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.
None of the authors have a financial interest in any of the products, devices, or drugs mentioned in this manuscript. This manuscript is the authors' own work and was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy– EXC 2050/1– Project ID 390,696,704– Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Lam K, Abràmoff MD, Balibrea JM et al (2022) A Delphi consensus statement for digital surgery. Npj Digit Med 5:1–9. https://doi.org/10.1038/s41746-022-00641-6
2. Mascagni P, Vardazaryan A, Alapatt D et al (2022) Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Ann Surg 275:955–961. https://doi.org/10.1097/SLA.0000000000004351
3. Lavanchy JL, Zindel J, Kirtac K et al (2021) Automation of surgical skill assessment using a three-stage machine learning algorithm. Sci Rep 11:5197. https://doi.org/10.1038/s41598-021-84295-6
4. Bhandari M, Nallabasannagari AR, Reddiboina M et al (2020) Predicting intra-operative and postoperative consequential events using machine-learning techniques in patients undergoing robot-assisted partial nephrectomy: a vattikuti collective quality Initiative database study. BJU Int 126:350–358. https://doi.org/10.1111/bju.15087
5. Jung JJ, Jüni P, Lebovic G, Grantcharov T (2020) First-year analysis of the operating room black box study. Ann Surg 271:122–127. https://doi.org/10.1097/SLA.0000000000002863
6. Komorowski M, Celi LA, Badawi O et al (2018) The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 24:1716–1720. https://doi.org/10.1038/s41591-018-0213-5
7. Maier-Hein L, Eisenmann M, Sarikaya D et al (2022) Surgical data science– from concepts toward clinical translation. Med Image Anal 76:102306, Feb. 2022. https://doi.org/10.1016/j.media.2021.102306
8. Reyes M, Meier R, Pereira S et al (2020) On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2:e190043. https://doi.org/10.1148/ryai.2020190043
9. Lahav O, Mastronarde N, van der Schaar M (2019) What is interpretable? Using machine learning to design interpretable decision-support systems. https://doi.org/10.48550/arXiv.1811.10799. arXiv:1811.10799
10. Imrie F, Davis R, van der Schaar M (2023) Multiple stakeholders drive diverse interpretability requirements for machine learning

in healthcare. Nat Mach Intell 5:824–829. https://doi.org/10.1038/s42256-023-00698-2
11. Yoon J, Jordon J, van der Schaar M (2019) INVASE: Instance-wise variable selection using neural networks. Published as a conference paper at ICLR 2019. Available under: https://openreview.net/forum?id=BJg_roAcK7
12. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. https://doi.org/10.48550/arXiv.1703.01365. arXiv:1703.01365
13. Crabbé J, van der Schaar M (2021) Explaining time series predictions with dynamic masks. Published in: Proc 38th Int Conf Mach Learn PMLR 139. https://doi.org/10.48550/arXiv.2106.05303.
14. Crabbe J, Qian Z, Imrie F et al (2021) Explaining latent representations with a corpus of examples. Published as Conf Paper NeurIPS 2021. https://doi.org/10.48550/arXiv.2110.15355.
15. Imrie F, Rauba P, van der Schaar M (2024) Redefining digital health interfaces with large Language models. https://doi.org/10.48550/arXiv.2310.03560. arXiv:2310.03560