Temporal Explainable AI Models for Surgery Evaluation

Haadia Amjad[1]

[1]Chair of Fundamentals of Electrical Engineering, TUD | Dresden University of Technology, Dresden, Germany

Abstract

This research explores video data explanations, incorporating temporal information, via XAI methods to enhance reliability for surgical training. It aims to detect learning biases in DNNs used for surgical skill assessment. The broader objective is to evaluate XAI performance with complex models and real-world data. This document details the context, methodology, preliminary findings, and future contributions.

1. Research Context and Motivation

1.1. Research Context

The research context of this work includes video data, specifically video trials for surgical skill assessment. These videos contain tasks, such as balloon-cutting and knot-tying, that students of surgery practice. In the medical context, these videos are rated by expert surgeons as beginner, intermediate, or expert performance (of students). We select deep learning models, that automate this process, and possible explainable AI (XAI) methods that can explain these algorithms effectively. Overall, this research involves surgical-skill-assessment video data, with their complex annotations, DNN models that predict the rating of these videos, and explainable AI techniques that can represent the reasoning behind model decisions and the possible impact of using temporal data in explanations.

1.2. Research Motivation

Explainable AI (XAI) methods and techniques are crucial in making high-risk AI applications understandable and reliable. While many such methods exist, they have to be adapted for different intended tasks, illustrated in Figure 1. Based on these intended tasks, the representation of explanations differs as the goal is to be meaningful to the end user. This context and usability involves applicative research on explainable AI methods that describe these methods' challenges, usefulness and considerations in real-world applications. One of the important domains where explainability is crucial is the medical domain. The decisions made by the AI systems deployed in the medical domain directly impact human life. These decisions have to be considered with more certainty and understandability.

Data in the medical domain is of various types. Videos are essentially a set of frames that represent information over time. One kind of video data in the medical domain is associated with surgery and surgical skill assessment. Surgical skill assessment is the process of evaluating the surgery capabilities of a surgery student or resident [1]. This assessment involves a set of tasks that target specific skills that are required to perform surgery. These tasks are recorded and then presented to expert surgeons who rate the performance of students based on the OSATS score [2]. These scores are a combination of many rated factors. To some extent, we can assume that one expert surgeon may give a slightly different score than another expert surgeon but overall, the difference is not too drastic.

These videos combined with these scores are being used as datasets for deep learning models to automate this assessment process. Generally, the deep neural networks (DNNs) for this task are used to

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

haadia.amjad@tu-dresden.de (H. Amjad)

0009-0001-9227-9496 (H. Amjad)

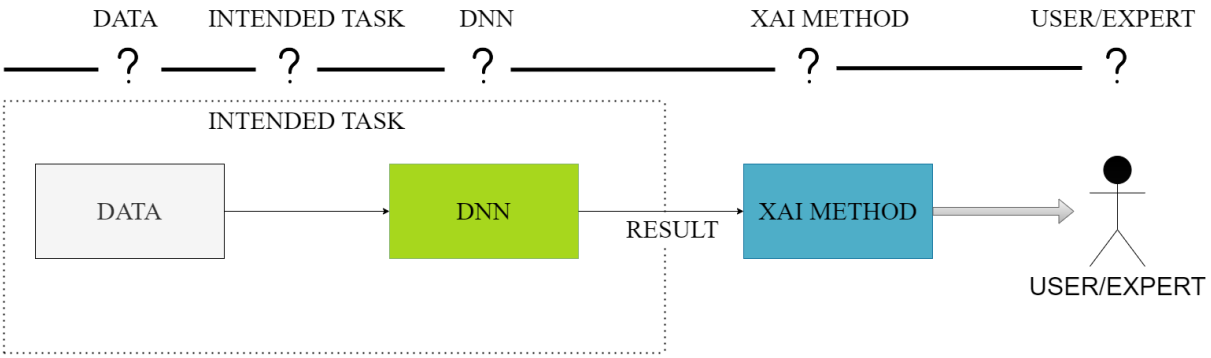

Figure 1: A representation of the kind of questions that need to be determined for an effective XAI pipeline.

sort the final scores into beginner, intermediate or expert categories of surgeons. Some models utilize datasets with instrument positioning, others are less specific. Funke et al. suggest that utilizing the temporal information of the video data and exploiting the information dependencies between frames produces better model performance [3]. Temporal Convolutional Networks (TCNs) are known for their power as they process information from both past and future time steps, making them effective for modelling the dynamics of temporal data.

With data as complex as surgical skill assessment, and models as complex as TSNs, it is difficult to apply XAI methods correctly. Many XAI methods work well with DNN models that are not as complicated as recent developments, and datasets such as MNIST [4]. For practical use, accurate models and complex datasets are what we have to work with. Hence, the XAI methods have to produce acceptable results that work with complex model layers and they have to visualize the model decisions in ways that make sense. Ultimately, the explanations have to add value to the DNN decision for the end user, which in the case of surgical skill assessment is the surgeon. This added value, or trust in the model decision involves exploring surgical skill assessment data to determine what kind of explanations are useful for the surgeons.

Additionally, XAI models can be used to evaluate the performance of a DNN model. Their visualizations and activation scores can potentially uncover learning biases. This enhances the reliability that tasks such as surgical skill assessment and other high-risk applications require.

This project plans to explore the explanations of video data that include the influence of temporal information. The goal is to use XAI methods to produce reliable explanations for surgery trainers and trainees and potentially detect learning biases of the DNNs that perform surgical skill assessments. The broader objective is to understand how XAI methods perform with complex models and data for real-life applications.

2. Related Work

Surgical skill assessment using video data and deep neural networks that incorporate temporal information has become an important area of research. Several studies have explored this approach using various methodologies. Funke et al. [3] developed a method using 3D convolutional neural networks (CNNs) to assess surgical skills from video data of skill assessment surgery tasks. Their approach leveraged the temporal information in video sequences and demonstrated high classification accuracies ranging from 95.1% to 100% on the JIGSAWS dataset [5]. Similarly, Yanik et al. [6] have applied deep learning techniques to laparoscopic surgery videos for skill acquisition assessment. These methods aim to provide objective and automated evaluation of surgical performance using only video input, which can be easily collected during training scenarios.

More recently, attention mechanisms have been incorporated into spatial-temporal neural network architectures for video-based surgical skill assessment. Wan et al. [7] investigated the use of explicit supervision of spatial attention, guided by instrument tip locations, to improve the generalizability of their algorithm. Their best-performing model achieved an area under the ROC curve of 0.88, with supervised spatial attention improving specificity and discrimination measures when tested on unseen datasets. These approaches demonstrate the potential for deep learning models to provide objective and automated assessment of intraoperative surgical skills using video data alone.

Temporal explainable AI (XAI) techniques have been developed to provide insights into how AI models process and make decisions based on time series or temporal data. Canti et al. [8] have compared various techniques for temporal XAI, exploring methods to generate explanations for models that incorporate time-dependent information. In the context of video action recognition, Saha et al. [9] have investigated ways to explain the decision-making process of models analyzing video inputs. These approaches extend popular techniques like Grad-CAM to work with video data, allowing for visualization of important spatial and temporal regions that contribute to the model's predictions. Such methods aim to provide a more comprehensive understanding of how AI models interpret and classify actions in video sequences.

The application of explainable AI in surgery, particularly for video analysis, has gained significant attention due to the critical nature of surgical procedures. In the work of Brandenburg et al. [10], surgeons and researchers have emphasized the importance of developing interpretable AI models to build trust and ensure safe implementation in clinical settings. A survey of explainable deep learning models in medical video analysis by Kolarik et al [11]. highlighted various approaches and applications, summarizing key requirements for explainability in medical contexts. These efforts aim to address the "black box" nature of deep learning models and provide clinicians with insights into the reasoning behind AI-generated assessments or recommendations.

Practical applications of XAI in surgical contexts have also been explored. Chittajallu et al. [12] developed an explainable AI system for content-based retrieval of video frames from minimally invasive surgery videos. Their approach used a self-supervised deep learning model and incorporated human feedback to refine search results, demonstrating the potential for XAI in surgical education and training. In a multi-institutional study by Kiyasseh et al. [13], researchers used AI to provide reliable and fair feedback to surgeons based on video analysis. They compared AI-generated explanations to those provided by human experts and proposed a method called "training with explanations" (TWIX) to improve the reliability and fairness of AI-based feedback across different cohorts of surgeons. These studies highlight the potential of XAI systems to support surgical training, assessment, and quality improvement while addressing important considerations such as reliability and trust in AI-generated feedback.

3. Research Questions, Hypothesis and Objectives

3.1. Research Questions

- Can information dependency between one frame of the video and another be represented using XAI? For example: Is something happening at 3 seconds in a video relevant to what is happening at 28 seconds in the video? Can this be reliably explained via XAI?
- Can XAI methods help surgery students understand the decisions made by a model on their surgical skill assessment?
- Can XAI methods help determine where potential learning biases arise while using a DNN for surgical skill assessment?

3.2. Research Hypothesis

- XAI methods can effectively represent and visualize the information dependency between different frames in a video, revealing temporal relationships across the video timeline.
- XAI methods can identify and illustrate potential learning biases in deep neural networks used for surgical skill assessment, allowing for better understanding and mitigation of these biases.
- The application of XAI methods to surgical skill assessment models will lead to improved model performance and reduced bias across different surgeon experience levels (e.g., novices vs. experts).

3.3. Research Objectives

- To develop and evaluate XAI methods for representing information dependency between video frames in temporal sequences.
- To create a reliable XAI framework that can explain relationships between temporally distant but semantically connected events in video data.
- To utilize XAI approaches to identify, visualize, and analyze potential learning biases in deep neural networks used for surgical skill evaluation.
- To measure the impact of applying XAI methods on surgical skill assessment model performance and bias reduction across different surgeon experience levels
- To evaluate the potential of XAI methods in improving surgical training programs by providing more transparent and actionable feedback to trainees.

4. Research Approach

The crucial first step of any AI development pipeline is understanding the dataset. While annotations are present for videos of surgical skill assessment, they are not annotated frame by frame. They are annotated based on either key events or overall performance. Key events, in this case, are events that, for a particular task, are considered failure points. Spotting these events in a video is easy even for someone who is not a surgeon. But understanding the overall performance is harder without the experience. To be able to attain reliable explanations and explore potential biases of the DNN model, it is important to understand this data. Therefore, a thorough review of the annotation process is needed. Additionally, consulting experienced surgeons who are familiar with this process can be highly beneficial.

Since this research is not focused on robust DNN models for surgery, developing new state-of-the-art surgery skill assessment models is out of the scope of this project. These models are selected from existing literature provided they do not measure the kinematics of the instruments (to start with simpler data, even in surgical skill assessment) and include temporal information during learning. These model(s) have to be trained and tested to obtain a reasonable performance. Since the intention is to explore model biases and also observe the trends in explanations, versions can be created of the DNN models. These variations can be models performing relatively poorly, training with the exclusion of a subset of annotations, and modifying other learning patterns. These versions may prove to help verify explanations and explore learning biases via XAI methods.

To learn and use XAI methods for real-life applications, they have to be tested on datasets that are not uncomplicated. This involves exploring the trends spotted in explanations when using complex DNN models and more cluttered datasets. This helps in not only understanding the XAI methods but also in modifying these methods so they may work accurately with more advanced models. Once a more experienced understanding of XAI methods has been developed, they can be used on single video frames to determine their behaviour excluding temporal data.

After exploration of standard XAI methods and determining their performance, XAI methods developed for temporal data, such as time series, should be studied. These XAI methods can not

simply be plugged into the existing pipeline. They have to be modified to work with video data in the context of surgical skill assessment. Based on their performance, a direction can be obtained to produce better explanations for the task at hand. This can lead to enhancing the quality of the explanations by introducing improvements to the existing methods.

All of the above demands rigorous evaluations. The concept of a reliable explanation has to be quantified by performance metrics. Furthermore, properties such as faithfulness, completeness and correctness, and others, of the explanations have to be determined to express the quality of the work. Other than standard evaluations, task-based evaluations can be constructed to prove the hypothesis. Additionally, a human study can be conducted with experienced surgeons on their view on the usefulness of these explanations.

5. Preliminary Results

To begin with, I conducted a thorough review of XAI methods. Based on the citations of these methods, it was determined what kind of data and models they work with, so far. Additionally, their use cases in terms of intelligibility questions ("what", "where", "what-if" etc.) were determined.

To start with understanding basic images-in-sequence data, I chose to work with autonomous driving video frames. These frames (images) are easy to understand (no need for an expert in the field) and the motion of one target object (for example: a person) can be seen clearer in consecutive frames. Three DNN models of varying complexity, namely VGG-16 [14], ResNet50 [15] and ConvNext-Tiny [16], were trained on the OSDaR23 dataset [17]. Two versions of these models were created, one performing well and one performing poorly. The intention of this experiment was the following: 1) Determine how XAI methods produce and visualize explanations for complex models and complex (multi-label, for example) datasets. 2) Identify the trends in explanations based on model performance. 3) Apply standard and concept-based XAI methods to models of different complexities and observe their trends.

We selected two standard methods, namely Layer-wise Relevance Propagation (LRP) [18] and Grad-CAM [19] and two concept-based methods, Concept Relevance Propagation (CRP) [20] and Concept Recursive Activation FacTorization (CRAFT) [21] to work with our three DNN models. It was observed that the model learns many biases with the target class despite performing well. This emphasized the use of XAI methods to explore dataset biases. Additionally, LRP and Grad-CAM show that the explanations of a target class are not consistent in consecutive frames. This formalizes our hypothesis that standard XAI methods are not consistent over multiple frames in a sequence.

In accordance with the above, I conducted a review of XAI methods that are designed to work with temporal data. Additionally, I researched explainability in the field of surgery and surgical skill assessment models. Based on this study, I highlighted common techniques in surgical skill assessment models and temporal XAI.

To start with a basic approach, an RNN model with a 1D-ResNet feature extractor was created inspired by ResNet-LSTM in the work of Kasa et al. [22] This model was trained on the JIGSAWS (JHU-ISI Gesture and Skill Assessment Working Set) dataset which is a benchmark dataset containing synchronized video and kinematic data from robotic surgical tasks (Accuracy:89.41, F1: 82.13). Applying LRP and CRP to the decisions of this model resulted in the following observation: For the same video and classification of the same gesture, LRP shows mild relevance over temporal intervals to the target class where whereas CRP shows multiple detailed relevance, illustrated in Figures 2 and 3.

As of now, no publications have been made of these results but they are in progress.

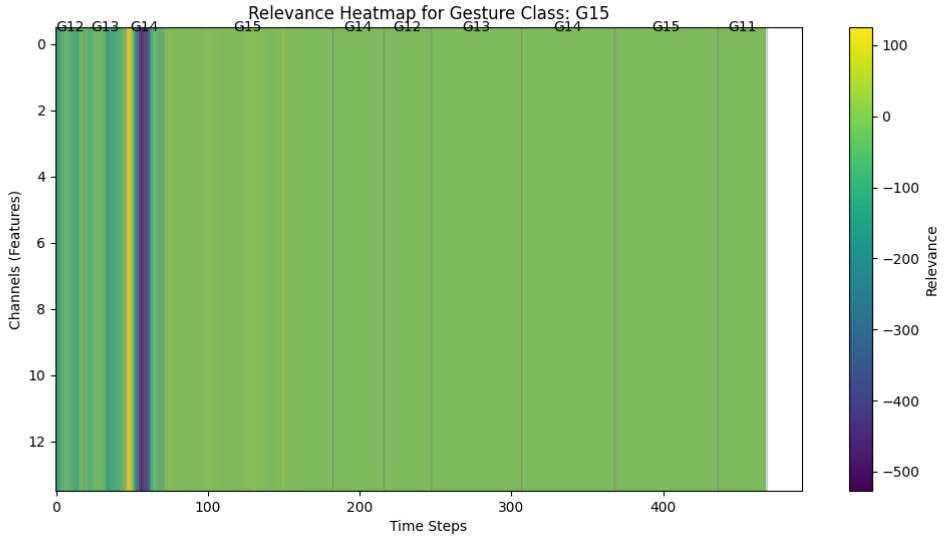

Figure 2: LRP shows mild relevance over temporal intervals of approximately 150 seconds. (for video of task balloon cutting)

Figure 3: CRP shows multiple detailed relevance over temporal intervals of approximately 150 seconds. (for video of task balloon cutting)

6. Next Research Steps

Currently, I am working on training more advanced surgical skill assessment models with a proprietary dataset obtained from a collaboration with the Department of Translational Surgical Oncology, National Center for Tumor Diseases, Dresden, a snippet illustrated in Figure 4. The goal is to obtain a model with high performance and then create variations of it based on performance and data subsets. Additionally, feature importance techniques are being used to store the statistical significance of the input data that may be referred back to after employing XAI methods on the model's decisions.

The next steps include selection of a group of temporal XAI methods and obtaining explanations (and visualizations) from these. The explanations can be variated to find the most suitable outcome

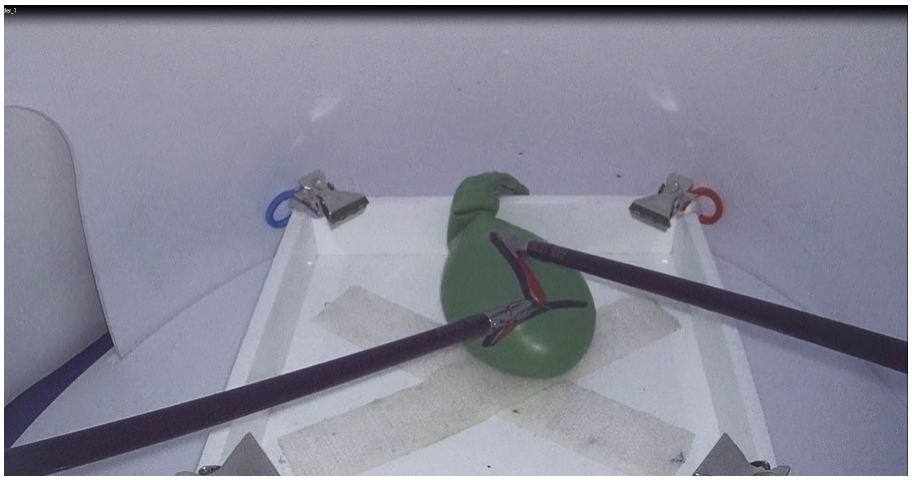

Figure 4: A snippet from a video containing surgical task balloon cutting.

based on the underlying technique of the temporal XAI methods. After rigorous evaluation of the explanations, they can be related to the properties of the dataset and the model to explore any potential biases. Furthermore, potential new techniques can arise by experimenting with this analysis. Lastly, a human study shall be conducted to obtain a subjective, yet useful view on the value of the explanations derived from the XAI methods and contributions of this study.

Acknowledgments

This work is partly supported by BMBF (Federal Ministry of Education and Research) in DAAD project 57616814 (SECAI, School of Embedded Composite AI, https://secai.org/) as part of the program Konrad Zuse Schools of Excellence in Artificial Intelligence.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] E. Yanik, X. Intes, U. Kruger, P. Yan, D. Diller, B. V. Voorst, et al., Deep neural networks for the assessment of surgical skills: A systematic review, The Journal of Defense Modeling and Simulation 19 (2022) 159–171.

[2] H. Niitsu, N. Hirabayashi, M. Yoshimitsu, T. Mimura, J. Taomoto, Y. Sugiyama, et al., Using the objective structured assessment of technical skills (osats) global rating scale to evaluate the skills of surgical trainees in the operating room, Surgery Today 43 (2013) 271–275.

[3] I. Funke, S. T. Mees, J. Weitz, S. Speidel, Video-based surgical skill assessment using 3d convolutional neural networks, International Journal of Computer Assisted Radiology and Surgery 14 (2019) 1217–1225.

[4] L. Deng, The mnist database of handwritten digit images for machine learning research [best of the web], IEEE Signal Processing Magazine 29 (2012) 141–142.

[5] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, et al., Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling, in: MICCAI Workshop: M2CAI, volume 3, 2014, p. 3.

[6] E. Yanik, J. P. Ainam, Y. Fu, S. Schwaitzberg, L. Cavuoto, S. De, Video-based skill acquisition

assessment in laparoscopic surgery using deep learning, Global Surgical Education-Journal of the Association for Surgical Education 3 (2024) 26.
[7] B. Wan, M. Peven, G. Hager, S. Sikder, S. S. Vedula, Spatial-temporal attention for video-based assessment of intraoperative surgical skill, Scientific Reports 14 (2024) 26912.
[8] E. Canti, E. Collini, L. A. I. Palesi, P. Nesi, Comparing techniques for temporal explainable artificial intelligence, in: 2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService), IEEE, 2024, pp. 87–91.
[9] A. Saha, S. Gupta, S. K. Ankireddy, K. Chahine, J. Ghosh, Exploring explainability in video action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8176–8181.
[10] J. M. Brandenburg, B. P. Müller-Stich, M. Wagner, M. V. D. Schaar, Can surgeons trust ai? perspectives on machine learning in surgery and the importance of explainable artificial intelligence (xai), Langenbeck's Archives of Surgery 410 (2025) 1–5.
[11] M. Kolarik, M. Sarnovsky, J. Paralic, F. Babic, Explainability of deep learning models in medical video analysis: A survey, PeerJ Computer Science 9 (2023) e1253.
[12] D. R. Chittajallu, B. Dong, P. Tunison, R. Collins, K. Wells, J. Fleshman, et al., Xai-cbir: Explainable ai system for content based retrieval of video frames from minimally invasive surgery videos, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 66–69.
[13] D. Kiyasseh, J. Laca, T. F. Haque, B. J. Miles, C. Wagner, D. A. Donoho, et al., A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons, Communications Medicine 3 (2023) 42.
[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[16] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[17] R. Tagiew, P. Klasek, R. Tilly, M. Köppel, P. Denzler, P. Neumaier, et al., Osdar23: Open sensor data for rail 2023, in: 2023 8th International Conference on Robotics and Automation Engineering (ICRAE), IEEE, 2023, pp. 270–276.
[18] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLOS ONE 10 (2015) e0130140.
[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[20] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through concept relevance propagation, Nature Machine Intelligence 5 (2023) 1006–1019.
[21] T. Fel, A. Picard, L. Bethune, T. Boissin, D. Vigouroux, J. Colin, et al., Craft: Concept recursive activation factorization for explainability, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2711–2721.
[22] K. Kasa, D. Burns, M. G. Goldenberg, O. Selim, C. Whyne, M. Hardisty, Multi-modal deep learning for assessing surgeon technical skill, Sensors 22 (2022) 7328.