Medical Image Analysis 86 (2023) 102765

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Beyond rankings: Learning (more) from algorithm validation

Tobias Roß [a,b,c,*,1], Pierangela Bruno [a,d,1], Annika Reinke [a,c,e], Manuel Wiesenfarth [f], Lisa Koeppel [g], Peter M. Full [b,h], Bünyamin Pekdemir [a], Patrick Godau [a,e], Darya Trofimova [a,k], Fabian Isensee [c,h,k], Tim J. Adler [a], Thuy N. Tran [a], Sara Moccia [i], Francesco Calimeri [d], Beat P. Müller-Stich [j], Annette Kopp-Schneider [f], Lena Maier-Hein [a,b,c,e,l]

[a] Intelligent Medical Systems (IMSY), German Cancer Research Center (DKFZ), Heidelberg, Germany
[b] Medical Faculty, Heidelberg University, Heidelberg, Germany
[c] Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany
[d] Department of Mathematics and Computer Science, University of Calabria, Rende, Italy
[e] Faculty of Mathematics and Computer Science, Heidelberg University, Germany
[f] Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[g] Section Clinical Tropical Medicine, Heidelberg University, Heidelberg, Germany
[h] Division of Medical Image Computing (MIC), German Cancer Research Center (DKFZ), Heidelberg, Germany
[i] The BioRobotics Institute and Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Italy
[j] Department for General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany
[k] HIP Applied Computer Vision Lab, MIC, German Cancer Research Center (DKFZ), Heidelberg, Germany
[l] Germany and National Center for Tumor Diseases (NCT), Heidelberg, Germany

ARTICLE INFO

Keywords:
Surgical data science
Image characteristics driven algorithm development
Minimally invasive surgery
Endoscopic vision
Grand challenges
Biomedical image analysis challenges
Generalized linear mixed models
Instrument segmentation
Deep learning
Artificial intelligence

ABSTRACT

Challenges have become the state-of-the-art approach to benchmark image analysis algorithms in a comparative manner. While the validation on identical data sets was a great step forward, results analysis is often restricted to pure ranking tables, leaving relevant questions unanswered. Specifically, little effort has been put into the systematic investigation on what characterizes images in which state-of-the-art algorithms fail. To address this gap in the literature, we (1) present a statistical framework for learning from challenges and (2) instantiate it for the specific task of instrument instance segmentation in laparoscopic videos. Our framework relies on the semantic meta data annotation of images, which serves as foundation for a General Linear Mixed Models (GLMM) analysis. Based on 51,542 meta data annotations performed on 2,728 images, we applied our approach to the results of the Robust Medical Instrument Segmentation Challenge (ROBUST-MIS) challenge 2019 and revealed underexposure, motion and occlusion of instruments as well as the presence of smoke or other objects in the background as major sources of algorithm failure. Our subsequent method development, tailored to the specific remaining issues, yielded a deep learning model with state-of-the-art overall performance and specific strengths in the processing of images in which previous methods tended to fail. Due to the objectivity and generic applicability of our approach, it could become a valuable tool for validation in the field of medical image analysis and beyond.

1. Introduction

Comparative performance assessment of image analysis algorithms is typically performed by either reimplementing state-of-the-art methods or by international benchmarking competitions, so-called challenges (Maier-Hein et al., 2018). As the re-implementation of other methods is prone to errors (e.g. errors in the implementation or suboptimal choice of hyperparameters) and time-consuming, challenges are nowadays the de facto standard for benchmarking new methods.

To date, however, relatively little effort has been put into the systematic analysis of results, as summarized in Wiesenfarth et al. (2021).

Specifically, most reports neglect a particularly relevant question for the medical domain:

What characterizes images on which algorithms fail?

Or, more broadly speaking:

* Corresponding author.
E-mail address: t.ross@dkfz-heidelberg.de (T. Roß).
[1] First two authors contributed equally to this paper.

https://doi.org/10.1016/j.media.2023.102765
Received 17 June 2021; Received in revised form 24 May 2022; Accepted 8 February 2023
Available online 1 March 2023
1361-8415/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

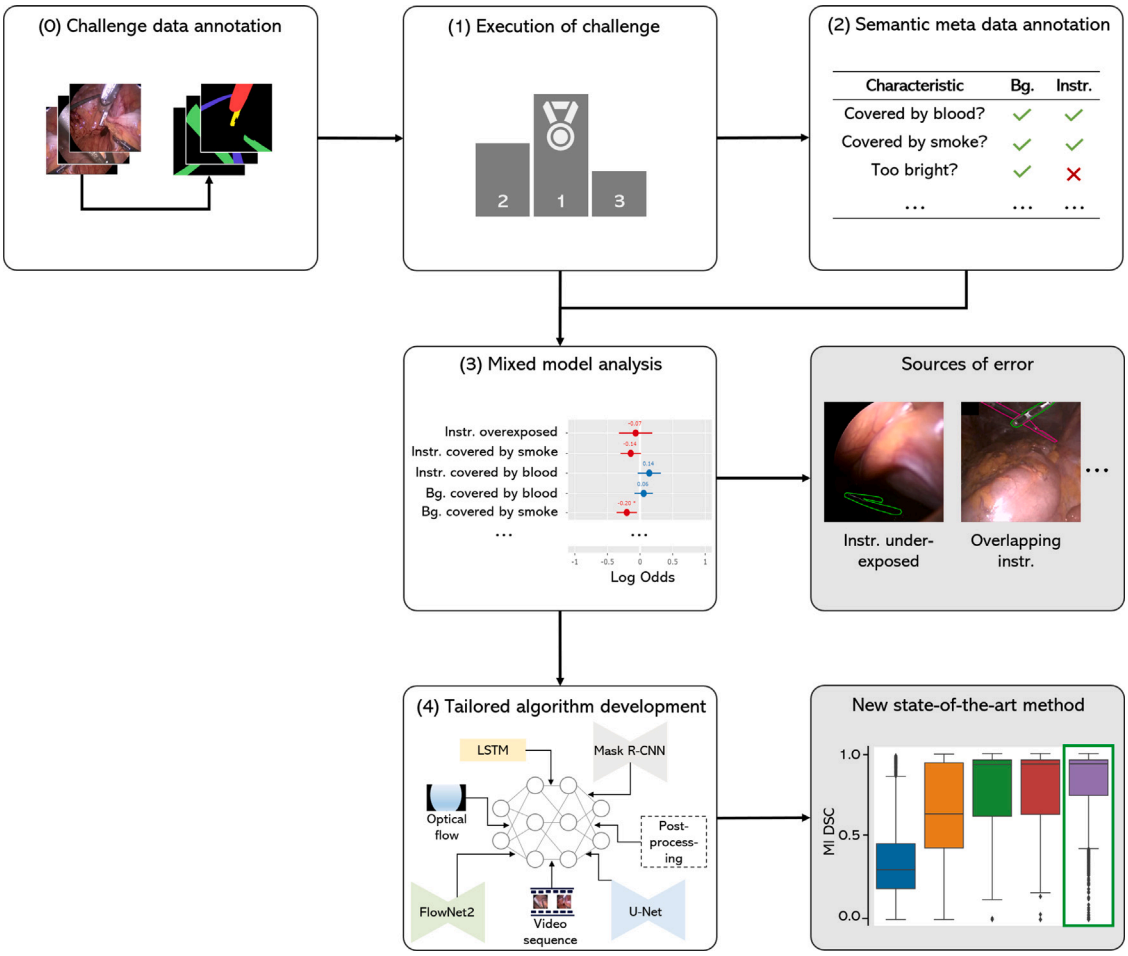

Fig. 1. Overview of the statistical approach to learning from a challenge. Based on the annotated challenge data set (0) and the results of the given challenge (1, here: Robust-MIS challenge), a semantic annotation of the challenge's test cases (2) is performed. This serves as the basis for a (generalized) linear mixed model ((G)LMM) analysis (3) to identify major sources of algorithm failure and quantify the respective impact. The algorithm development is then tailored to the specific weaknesses (4) with the goal of enabling a new state-of-the-art performance. (Bg: Background; Instr: Instrument).

How can we learn from challenge results in a way that enables us to tailor future algorithm development to the specific remaining needs?

Some challenge organizers have recognized this problem and carried out a laborious manual analysis, e.g., by reporting the best and worst cases based on the participants' performances and identifying a set of image characteristics that could lead to worsening or improving performance (i.e., over-/underexposed images) (Roß et al., 2020; Allan et al., 2021, 2020, 2019). However, this approach is rather subjective and does not allow for reliable quantification of the effects of different sources. Furthermore, it may be subject to confirmation bias. Given the lack of systematic analysis methodology, the contribution of this paper is threefold:

1. We present a statistical framework for learning from challenges, which focuses on the identification of sources for algorithm failure (Fig. 1).
2. To demonstrate potential benefit of the new concept, we apply it to the recently published challenge on multi-instance laparoscopic instrument segmentation ROBUST-MIS[2] (Roß et al., 2020) (see Fig. 2).
3. We demonstrate that knowledge on the identified sources of error can help improve algorithm performance in common failure cases.

[2] Robust Medical Instrument Segmentation (Robust-MIS) Challenge 2019, https://www.synapse.org/#!Synapse:syn18779624/wiki/591266.

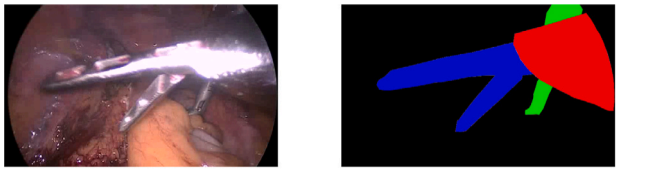

Fig. 2. (a) Sample video frame with multiple overlapping instruments and (b) corresponding annotation (red: trocar, green/blue: graspers). Challenges include presence of specular reflections, noise, motion blur, varying illumination levels, overlapping instruments and instrument occlusions.

Note that this approach was specifically designed for challenges but is similarly applicable to more basic validation studies, in which the performance of only a single algorithm is assessed.

The remainder of this paper is structured as follows: After presenting the related work in Section 2, we describe our framework for challenge analysis, its application to the task of multi-instance segmentation as well as our strength–weakness-driven algorithm development in Section 3. The performed experiments and results are presented in Section 4 and discussed in Section 5. Section 6 concludes this paper by summarizing our main findings.

2. Related work

In this section, we present the related work on systematic challenge analysis in the field of biomedical image analysis as well as a brief summary on the state-of-the-art in multi-instance medical instrument segmentation.

2.1. Challenge analysis

The literature on the analysis of image analysis challenges is extremely sparse, both within and outside the medical community. In fact, the meta science papers published to date have focused on ranking instabilities (Maier-Hein et al., 2018; Reinke et al., 2018), standards (Mendrik and Aylward, 2019; Maier-Hein et al., 2018, 2020) and challenge visualization (Wiesenfarth et al., 2021), while the topic of results analysis has been given extremely little attention. The closest work to ours was only recently published and presents a framework for visualizing challenge results in an uncertainty-aware manner (Wiesenfarth et al., 2021). While sources of error are not addressed within the framework, the paper provided an important motivation for our work: An analysis of numerous challenge reports in the field of biomedical image analysis revealed that a large number (66% of those investigated) report only final ranks or aggregated performance measures (Wiesenfarth et al., 2021) without providing further analyses. This finding is in line with our more recent observations: Challenge reports often only provide a website with the rankings (e.g., MISAW,[3] SurgVisDom,[4] EndoVis-WorkFlowChallenge[5]), and corresponding publications concentrate on the presentation of aggregated metric values (Al Hajj et al., 2019; Bodenstedt et al., 2018), visual examples (Allan et al., 2019, 2021) or manual inspection of best/worst cases (Roß et al., 2020). In fact, we are not aware of any prior work on identifying sources of algorithm failure in a systematic manner.

2.2. Multi-instance segmentation

While the task of binary instrument segmentation received a lot of attention over the last couple of years such as Lee et al. (2019), Shvets et al. (2018), Jin et al. (2019), Allan et al. (2015) and García-Peraza-Herrera et al. (2016), literature on multi-instance segmentation in applications for minimally invasive surgeries is extremely sparse. To our knowledge, the only peer-reviewed work published independently of the ROBUST-MIS challenge (Roß et al., 2020) (which this work is based on), was published by Shvets et al. (2018). Their work is on robotic instrument segmentation and features a comparatively simplistic data set with respect to image characteristics (e.g., blood, reflections). Hence, the methods competing in the ROBUST-MIS challenge can be regarded as representative for the state-of-the-art in the field.

3. Methods

The following sections present the proposed framework for learning from challenges (Section 3.1), its instantiation in the ROBUST-MIS challenge (Section 3.2), as well as the proposed deep learning method resulting from problem-tailored algorithm development (Section 3.3).

3.1. Framework for learning from challenges

This section introduces our concept for learning from challenges and details the underlying statistical approach.

[3] https://www.synapse.org/#!Synapse:syn21776936/wiki/601705.
[4] https://www.synapse.org/#!Synapse:syn22083820/wiki/606329.
[5] https://endovissub2017-workflow.grand-challenge.org/PastChallenges/.

3.1.1. Concept overview

We propose the following four-step procedure for learning from challenges.

1. **Hypothesis generation:** In an initial step, potential sources of algorithm failure are identified. These can relate to the image device (e.g. dirty endoscope lens), the imaging protocol (e.g. overexposure/underexposure), the handling of the equipment (e.g. motion blur), the target structure (e.g. crossing medical instruments in the case of ROBUST-MIS) and other application-specific features (e.g. smoke in the field of view of a laparoscope). To generate a list of image characteristics that may lead to poor performance, knowledge from the literature, expert knowledge, personal experience, as well as a manual analysis of the challenge results (as in Roß et al. (2020)) can be leveraged.
2. **Semantic meta data annotation:** (Part of) the challenge test cases are then semantically annotated with these image characteristics. This can be done by domain experts, or by leveraging crowdsourcing, for example. Please note that we use the term *test case* "to refer to a data set for which a participating algorithm produce one result" (Maier-Hein et al., 2020).
3. **Mixed model analysis:** The semantic labels on image characteristics along with the challenge results – represented by (aggregated) metric results per test case – are then leveraged to identify image characteristics leading to poor algorithm performance. To this end, a mixed effects model (West et al., 2014) is set up in which the (possibly transformed) metric values embody the outcome variable and the image-specific information is integrated as explanatory variables. In other words, the performance of an algorithm on a given image is represented as a function of the meta information available for the image. The method is detailed in Section 3.1.2.
4. **Tailored algorithm development:** Based on the identification of those error sources that have the biggest effect on algorithm performance, algorithm development is tailored to the specific problems identified.

Our approach to mixed model-based challenge analysis is detailed in Section 3.1.2, followed by an instantiation of this framework in the specific task of multi-instance instrument segmentation in laparoscopic video data in Section 3.2.

3.1.2. Mixed model analysis

Although mixed model analysis is a standard approach in statistics, we found that it is only very rarely employed in medical image analysis. More specifically, we reviewed all 5390 papers published between 2004 and 2021 in the proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), the largest conference in the field of medical image analysis. According to a systematic search (see Appendix B, Table B.4), only 1.7% of the papers applied mixed model analysis despite the fact that a hierarchical data structure (i.e. non-independent test cases) is very common in medical imaging. Due to the rare usage of mixed models in the field, this section will provide a light introduction into the topic.

As summarized in the previous section, the goal of the statistical analysis is to leverage the semantic meta data annotation to identify image characteristics leading to poor algorithm performance. Standard linear regression models are not suitable for this purpose whenever individual data points are not independent. Non-independence of the data is a typical characteristic of challenges, for example because multiple images from the same patient, or multiple frames from the same video, are used in the analysis. In such a situation, data are best represented by a hierarchical structure. In such a data tree (cf. Fig. 3), data corresponding to the same leaf can be assumed to be independent. Branches represent the source of non-independence, such as a specific hospital, device, or patient. To account for the correlation within challenge outcome data, we propose using Linear Mixed Models

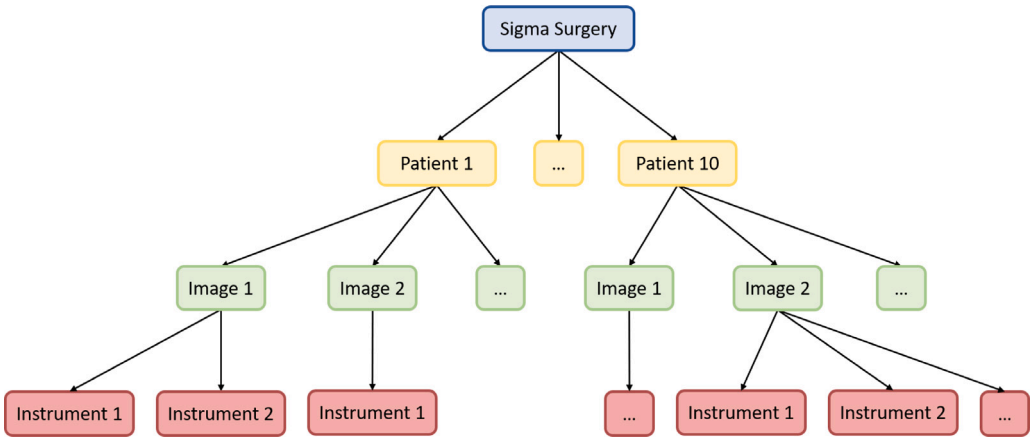

Fig. 3. Hierarchical structure of the ROBUST-MIS 2019 data. The stage 3 test set comprises solely the Sigma surgery that was performed for ten patients. For each patient $p \in P$, a varying number of images are acquired. Every image $i \in I$ itself contains a varying number of instrument instances J_i.

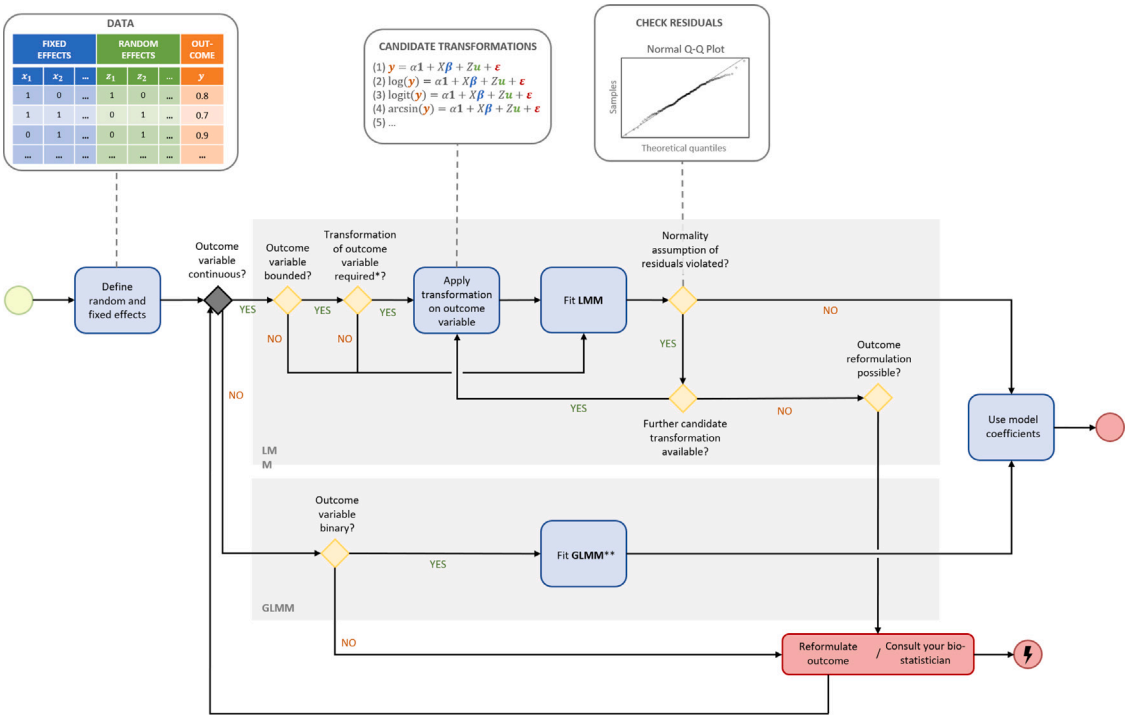

Fig. 4. Mixed model-based statistical analysis for a challenge with a continuous metric value as outcome. Initially, random and fixed effects are defined. Image characteristics that have a potential influence on algorithm performance are assumed to be binary and are represented as fixed effects in the model. Other factors, reflecting the hierarchical structure of the acquired data (e.g. the patient/hospital/image frame identifier) are represented as random effects. Depending on the distribution of the outcome, either a Linear Mixed Model (LMM) or a Generalized LMM (GLMM) for binary outcome is the model of choice. In the case of the LMMs, a transformation of the outcome may be required before the model is fitted to avoid violation of the normality assumption. Further details of the workflow are provided in Sections 3.1.2. and 3.2.
*Note that metric values may be bounded in theory but appear normally distributed in the specific data set. In this case, no transformation is needed.
**Note that a GLMM for binary outcome is a mixed effect logistic regression model.

(LMMs) (West et al., 2014), as a generalization of linear regression models.

LMMs enable regressing an outcome using a linear combination of explanatory variables weighted by regression coefficients. The outcome variable is also referred to as *dependent/target/response/explained variable* while the explanatory variables are also called *independent variables*. In contrast to standard regression models, mixed models not only incorporate the parameters of interest, referred to as *fixed effects*, but also so-called *random effects* of variables explaining the hierarchical structure.

In our specific (challenge) setting, image characteristics that have a potential influence on algorithm performance are by default assumed to be binary (present/not present; though categorical variables are also possible) and are represented as fixed effects in the model. Other factors, reflecting the hierarchical structure of the acquired data (e.g. the patient/hospital/image frame identifier) are represented as random effects. The fixed effects coefficients of the model then provide estimates of the impact of the provided image characteristics on the prediction performance. There are several ways to incorporate the challenge participant's algorithms in this setup. If the specific effects of the (typically small number of) algorithms are of interest, the algorithm can be modeled as a fixed effect, otherwise it is modeled as a random effect. Alternatively, as a first approximation, aggregated metric values across algorithms (possibly after transformation; see below) may be used.

Fig. 4 presents a simplified workflow for choosing the appropriate mixed model for a given problem. The specific choice of the mixed model and the process of instantiating depends primarily on the distribution of the metric values. In the case of continuous metric values with unbounded support, an LMM (West et al., 2014) is usually the most natural choice. Applied to the problem of challenge analysis, LMMs can be represented as:

$$\overbrace{y}^{N\times 1} = \overbrace{\underbrace{\alpha}_{1\times 1}\underbrace{\mathbf{1}}_{N\times 1}}^{N\times 1} + \overbrace{\underbrace{X}_{N\times p}\underbrace{\beta}_{p\times 1}}^{N\times 1} + \overbrace{\underbrace{Z}_{N\times q}\underbrace{u}_{q\times 1}}^{N\times 1} + \overbrace{\varepsilon}^{N\times 1} \quad (1)$$

with the following components:

Outcome. The vector y represents the N metric values on the test images, which may be aggregated over all algorithms or be provided separately for all of them. For clarity of presentation, we will assume a single value per image in the following explanation, but an example of a more fine-grained composition is provided in Section 3.2.2.

Fixed effects. The so-called *design matrix* X corresponds to the p image characteristics with corresponding p fixed effects β. Generally, each row of X represents an image and consists of p binary variables representing the presence or absence of a specific image characteristic on the corresponding image. β is the regression coefficient resulting from mixed model fitting, which can be used to predict the dependent variable y from the fixed effects. The fixed effects considered in our model setting are listed in Table 1 and illustrated in Fig. 8. Coefficients are easily interpretable in LMMs with untransformed dependent variables. Given, for example, presence of an image characteristic c, the expected metric value changes by β_c compared to the situation when the characteristic is not present. The so-called *intercept* α is a scalar representing an average outcome when all characteristics are absent (the arithmetic mean of y) and $\mathbf{1}$ is a vector of 1s.

Random effects. The matrix Z is a so-called *design matrix for the random effects*. Assuming, for example, a hierarchical data structure, in which clustering of test images arises from different patients, each patient would represent one of q columns in Z, and each row in the matrix would have exactly one entry with 1 (0 otherwise), reflecting the fact that there is a unique assignment of outcome data to a specific patient. Similarly, further random effects, such as hospital IDs, can be incorporated by increasing the number of columns of Z (and dimension of u). u quantifies the random effect on the specific outcome. For example, the imaging device of a specific hospital could have a lower resolution, hence making predictions more difficult and thus leading to a worse metric value on average. Elements in u are assumed to be normally distributed with zero mean and a variance interpreted as the between-cluster-variability (e.g. the variability in performance when comparing images of different hospitals). If the between-cluster-variability is large when compared to the residual variance (see below), observations within the cluster are highly correlated.

Residuals. ε is a vector of residuals. The residuals are assumed to be normally distributed, $\mathcal{N}(0, \sigma_\varepsilon^2 I)$ capturing the variation in y unexplained by the random and fixed effects.

Fitting of the LMM (i.e. estimation of the coefficients) is carried out through restricted maximum likelihood methods (REML). LMMs rely on the assumption of normality of the residuals, i.e., the outcome variable given the values of the explanatory variables follows a normal distribution. To detect a potential violation of the normality constraint, we recommend studying the Q-Q-Plot (Thode, 2002) of the residuals after model fitting. Often, a transformation of the metric values must be applied to obtain approximate normality such that an LMM can be used. For a metric with values $\in [0, \inf]$, for example, the *log* transformation is a popular choice; if a metric is bounded by $[0, 1]$, the *logit* function, mapping the metric values to $[-\inf, \inf]$, is frequently used. Note that in case of such nonlinear transformation, additivity and linearity of effects on y is lost, possibly complicating interpretation of regression coefficients.

If the outcome variable follows a non-normal distribution such as Binomial/Bernoulli, Poisson or Gamma, GLMMs, as a generalization of LMMs can be used (McCulloch et al., 2011). Specifically, if the outcome variable is binary, a mixed effects logistic regression model as a special case of GLMMs can be used. An instantiation of such a model is provided in Section 3.2.2.

Table 1
Meta data information provided by a human annotator for the entire background (Bg.) and for each instrument instance (Inst.), and/or globally.

Local characteristics		
Characteristic	*Bg.*	*Inst.*
Covered by blood?	✓	✓
Covered by smoke?	✓	✓
Covered by tissue?	✗	✓
Subject to motion artifacts?	✓	✓
Covered by specular reflections?	✓	✓
Covered by another instrument?	✗	✓
Covered by any other object (non-surgical)?	✓	✓
Too bright?	✗	✓
Too dark?	✗	✓
Global characteristics		
Characteristic	*Img.*	
Is the image too bright?	✓	
Is the image too dark?	✓	
Does the lens seem dirty?	✓	

3.2. Instantiation of the framework for multi-instance instrument segmentation

As a proof of concept, we instantiated our proposed framework for the ROBUST-MIS challenge 2019 (Roß et al., 2020). This challenge was based on 10,040 endoscopic images that had been extracted from three different surgery types (Maier-Hein et al., 2021). For the ROBUST-MIS challenge, a test case was defined as the last frame of a 10 s video snippet of 250 endoscopic image frames. For each test case, the segmentation results of five (one algorithm was excluded due of a non-compliant training process that would affect comparability) participating algorithms were available.

3.2.1. Annotation of image characteristics

To identify potential sources of error, we analyzed the general literature on endoscopic image analysis (Maier-Hein et al., 2014a; Ali et al., 2021b) as well as the specific literature on artifacts in endoscopy (Ali et al., 2020, 2021a; Funke et al., 2018; Soberanis-Mukul et al., 2020) and on endoscopic vision challenges (Bodenstedt et al., 2018; Allan et al., 2020; Roß et al., 2020). Combined with personal experience gained during the annotation process for the challenge data (Maier-Hein et al., 2021), we identified twelve relevant sources of error (see Table 1) which were classified as global – characterizing the whole image (here: dirty lens, overexposure and underexposure) – or local (e.g. blood on specific instrument instance). The local features were provided for all instrument instances and/or the background individually. A trained engineer with experience in endoscopic image annotation (annotator was part of the annotation team of Maier-Hein et al. (2021)) then annotated the presence of such characteristics for all images (see Fig. 6). For the relevant test set (stage 3) this resulted in a total of $(5 + 3) \cdot 2,728 = 21,824$ image related and $9 \cdot 3,302 = 29,718$ instrument instance annotations. We performed the statistical analysis

solely on the test cases, as algorithm results were only available for the challenge's test set. However, Fig. 6 shows the comparison of the image characteristics between training and test set to emphasize that both sets contained a similar distribution of characteristics.

3.2.2. Statistical analysis

The Dice Similarity Coefficient (DSC) (Dice, 1945) is a widely used metric in medical image analysis (Maier-Hein et al., 2018; Reinke et al., 2018) and also served as a basis for the ranking in the ROBUST-MIS challenge (Roß et al., 2020). Yet, as DSC values are in the range of $[0, 1]$, modeling the algorithm performances directly on the challenge metric would violate the normality assumption of the residuals ε. As mentioned in Section 3.1.2, the problem can potentially be overcome by applying a transformation $f(\cdot)$, such as the *logit* to metric values bounded in $[0, 1]$, with the goal of mapping the values to the range of $[-\inf, \inf]$. As this process did not yield an approximate normal residual distribution of ε for our data, we propose regarding the segmentation problem as a pixel-level classification problem and applying a GLMM to model the target metrics precision and recall as a function of image characteristics, as detailed in the following paragraphs.

Outcome reformulation. For this study, we leverage the fact, that the DSC is closely related to *precision* (PRE) and *recall* (REC). More specifically, DSC, PRE and REC can all be expressed as a function of the number of true(T)/false(F) positives(P)/negatives(N):

$$PRE = \frac{TP}{TP + FP};\ REC = \frac{TP}{TP + FN} \tag{2}$$

$$DSC = F1 = \frac{2TP}{2TP + FP + FN} = \frac{2 \cdot PRE \cdot REC}{PRE + REC} \tag{3}$$

To use *precision* and *recall* as target metrics, we convert each multi-instance reference annotation mask into a set of binary masks, each corresponding to one instrument instance. For each image $i \in I$, each instrument $j \in J_i$ that is present in i and each algorithm $k \in K$ we then determine both, the *recall*, defined as the probability $\pi_{i,j,k}$ of a pixel of the reference segmentation to be present in the mask provided by the algorithm, and the *precision*, defined as the probability $\tilde{\pi}_{i,j,k}$ of a pixel of the segmentation mask to be present in the reference segmentation. In other words, we relate the TP to either the reference mask of an instance (*recall*) or the mask provided by the algorithm (*precision*). Formally, the pixel-level classification (per instance) is binary and thus can be regarded as a Bernoulli experiment. Depending on the perspective (*recall* or *precision*) if a pixel is correctly classified as instrument follows a Bernoulli distribution $Bernoulli(\pi)$ with parameter $\pi = \pi_{i,j,k}$ resp. $\pi = \tilde{\pi}_{i,j,k}$.

GLMM fitting. In GLMMs, a *link function* g relates the expected outcome (here the parameter π) with the linear predictor. The canonical choice for this function in case of binary data is the logit link function $g(\pi) := \log \frac{\pi}{1-\pi}$. The complete equation is then given by

$$\overbrace{g(\boldsymbol{\pi})}^{N\times 1} = \overbrace{\underbrace{\alpha}_{1\times 1}\underbrace{\mathbf{1}}_{N\times 1}}^{N\times 1} + \overbrace{\underbrace{X}_{N\times p}\underbrace{\beta}_{p\times 1}}^{N\times 1} + \overbrace{\underbrace{Z}_{N\times q}\underbrace{u}_{q\times 1}}^{N\times 1}, \tag{4}$$

where $\boldsymbol{\pi}$ is a column-vector consisting of all the probabilities $\pi_{i,j,k}$ (resp. $\tilde{\pi}_{i,j,k}$) and g is applied element-wise. As described above, we define fixed effects β (see Table 1 and Fig. 8) for image and instrument characteristics (within X), resulting in $p = 17$. Furthermore, we model the participants' algorithms, as well as the patient, image and instrument as independent random effects (more detail in) combined in vector u (with indicator matrices combined in Z), resulting in $q = |K| + |P| + |I| + \sum_{i\in I} |J_i|$, where $|K|, |P|, |I|, \sum_{i\in I} |J_i|$ refer to the number of algorithms, the number of patients, the number of images and the total number of instances, respectively. The number of rows N in our model comes down to $N = |K| \cdot \sum_{i\in I} |J_i|$. Note that we inspected the cases of single and multiple instruments per image separately (see Table 2). So while $|K| = 5$ and $|P| = 10$ stayed constant there were 1184 images with 1 instrument and 1031 images with multiple instruments each (2118 instruments in total). By modeling the algorithms as random effects, the analysis is performed globally across all algorithms. However, the analysis could also be performed for comprehensive performance assessment of an individual algorithm. In this case, the algorithm can be removed as a random effect, as detailed in Appendix C.

Table 2
Sources of algorithm failures and successes, where + denotes a significant positive effect of $0 < OR \leq 1.25$, ++ a significant positive effect of $1.25 < OR \leq 1.50$ and +++ a significant positive effect of $1.50 < OR$. Analogously, − denotes a significant negative effect of $0.75 \leq OR < 0$, −− a significant negative effect of $0.50 \leq OR < 0.75$ and −−− a significant negative effect of $OR < 0.50$. Empty columns indicate no significant impact could be found. A significance level of 0.05 is used throughout. "x" means that the effect could not be assessed (e.g. the effect covered by instrument does only exist when $n > 1$). "inst." refers to instrument instance.

Characteristic	Local characteristics			
	Precision		Recall	
	Inst. > 1	Inst. = 1	Inst. > 1	Inst. = 1
Instrument overexposed	++			
Instrument covered by tissue	++		+	++
Instrument underexposed	−	−−	−−	−−−
Instrument covered by reflections		++		
Instrument covered by material		−−		−−
Instrument covered by smoke				
Instrument in motion		−−	−	
Instrument covered by blood				
Instrument covered by instrument	−−−	x	−−	x
Background contains other objects		−−		
Background in motion				
Background covered by blood				
Background covered by reflections				
Background covered by smoke		−	−	
Characteristic	**Global characteristics**			
	Precision		Recall	
	Inst. > 1	Inst. = 1	Inst. > 1	Inst. = 1
Image lens dirty				
Image too bright				
Image too dark			−	

Model interpretation. After fitting the model, coefficients β can be interpreted in terms of log odds ratios (OR) (McCulloch et al., 2011). The OR is a statistic measuring how two events are associated with each other regarding their presence or absence (McCulloch et al., 2011). Here, the OR measures the ratio of the odds (e.g. $\pi_{i,j,k}/(1-\pi_{i,j,k})$) in the presence and absence of a given image characteristic. However, the fact that the OR is not symmetrical around 1 (the value indicating no effect of the image characteristic) makes the comparison of the individual effects less intuitive (as can be illustrated by an OR of 2:3 ($0.\overline{6}$) and its inverted ratio of 3:2 (1.5)). Fig. 7 thus shows the log of the OR, making the values symmetrical around 0. For the interpretation, a positive log OR increases the chance that a high metric is measured, while a negative log OR decreases the chance.

3.3. Strength–weakness-driven algorithm development

As detailed in the results Section 4.1, the GLMM revealed several image characteristics with major impact on algorithm performance, namely: motion and underexposure of the instrument, crossing medical instruments as well as smoke or other medical equipment (e.g. swabs or bandages) in the field of view. We hypothesized that a majority of these issues can be addressed by going beyond single image analysis and *taking the temporal context into account*. In fact, the annotators of the Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room (HeiCo) (Maier-Hein et al., 2021), which served as basis for the ROBUST-MIS challenge, also reported that analysis of preceding frames was sometimes necessary to label a given

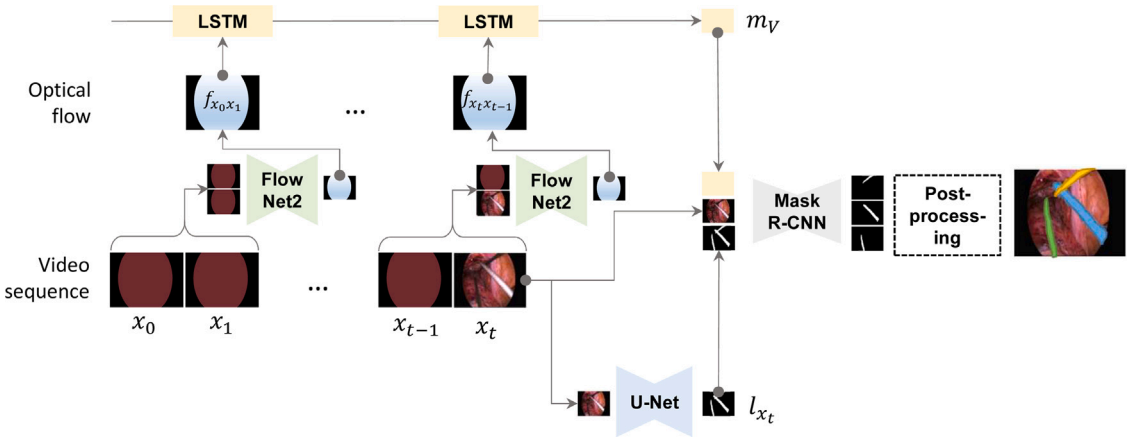

Fig. 5. Concept of the multi-instance segmentation approach for a video frame x_t of a video V. A Mask R-CNN performs instance localization based on three input channels: The current image frame x_t, the instrument likelihood l_{x_t} estimated with a U-Net (Ronneberger et al., 2015), and third channel m_V encoding motion information via optical flow. The optical flow of a video sequence $V = (x_0, \dots, x_t)$ of length $t+1$ is estimated by (1) determining the optical flow for each pair of consecutive frames via FlowNet2 (Ilg et al., 2017) and then (2) summarizing the pairwise information with an LSTM into m_V. A post-processing step applied to the output of the Mask R-CNN (see Sect. Appendix A.4) yields the final result.

frame. The proposed architecture resulting from the GLMM analysis is shown in Fig. 5. The core component of the presented deep learning architecture is a masked region-based convolutional neural network (Mask R-CNN) (He et al., 2017) that uses the following information as input: (1) the raw video frame, (2) the probability of a pixel to be an instrument and (3) the Long short-term memory (LSTM)-summarized (Hochreiter and Schmidhuber, 1997) information on object motion encoded in an optical flow. Details are provided in Appendix A.

4. Experiments and results

We validated our framework on the ROBUST-MIS challenge 2019 data set (Roß et al., 2020; Maier-Hein et al., 2021). The following sections present our findings with respect to image characteristics impacting algorithm performance as well as the validation of our algorithm tailored to the specific weaknesses of the state-of-the-art.

4.1. Effect of image characteristics on the performance of state-of-the-art algorithms

The frequency of special image characteristics captured by the meta data annotation is shown in Fig. 6.

Following the statistical methodology presented in Section 3.2, we analyzed the influence of image characteristics on the algorithm performance, using precision and recall as metrics. The main results are presented in Fig. 7 and Table 2.

According to our results, the following main characteristics had a statistically significantly ($p < 0.05$) negative influence on the results: instrument is underexposed, in motion or covered by material, or background is covered by smoke or other objects. Example frames are provided in Fig. 8. When two or more instrument instances were visible (see Fig. 7), the statistically significant characteristic with the largest impact was "instrument covered by another instrument".

4.2. Performance of algorithm developed based on strength–weakness -analysis

Our method resulting from the strength–weakness-analysis was compared to the ROBUST-MIS 2019 challenge's top-scoring methods based on the accuracy as reported in the ROBUST-MIS challenge (Roß et al., 2020) and using the challengeR analysis tool (Wiesenfarth et al., 2021). For the validation of our method, we used the metrics proposed by the ROBUST-MIS challenge as this allowed us to compare it to

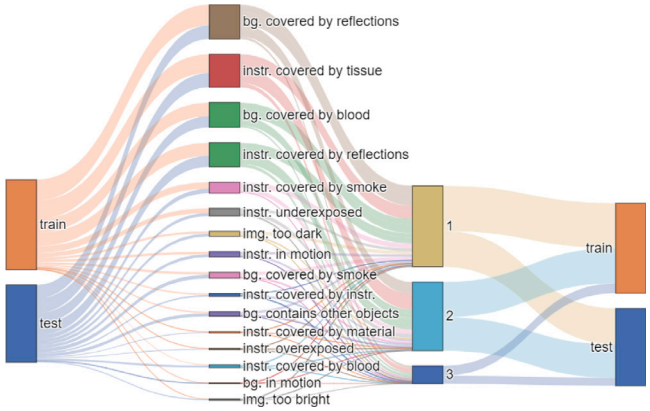

Fig. 6. Summary of the 195,148 meta data annotations performed on the ROBUST-MIS challenge data (training and test data). The frequency of the selected image characteristics (Table 1) in the test and training data set is shown along with information on the number of instruments per image. The size of the blocks and connecting lines in the image correspond to the proportion of images that have the respective property. The colors were picked to improve the legibility of the image ("bg.": background; "instr.": instrument; "img.": image).

the best-performing related methods. Following the challenge guidelines (Roß et al., 2020), we split the data into 5983 training and 4057 test images. The reason for the relatively high number of test images compared to training images was the fact that we reserved one surgery type exclusively for testing, as detailed in Roß et al. (2020). As shown in Fig. 10 and A.12, our method shows higher performances compared to the state-of-the-art in the majority of categories that represent typical failure cases. The metric value distributions for the top-performing ROBUST-MIS algorithms and our method are presented in Fig. A.12 for (a) the three stages of the test dataset and (b) varying numbers of visible instruments in the image. The proposed method achieved the highest median *multi-instance DSC (MI_DSC)* score (Roß et al., 2020) for the most difficult test stage (unseen procedure and patients), namely Stage 3.

We further performed an ablation study to assess the benefit of the different architectures components, namely (1) the optical flow and (2) the instrument likelihood as additional input to the Mask R-CNN and (3) our post-processing method applied to the output of the Mask R-CNN (see Appendix A for more details). The results are presented in Fig. 9 and Table 3. Including the optical flow (T_R vs. T_{RF}) improved

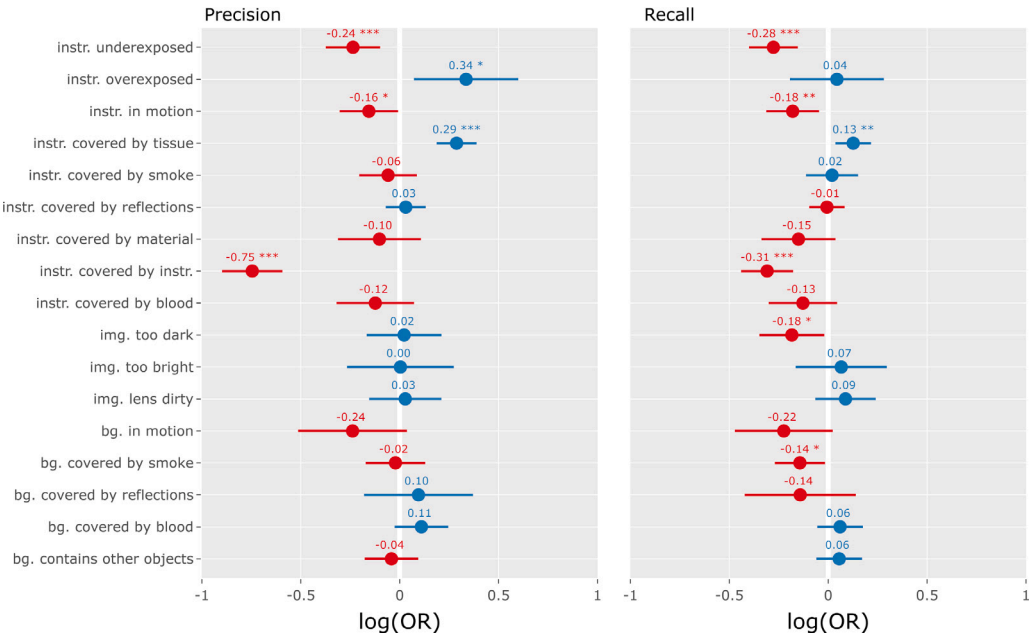

Fig. 7. Impact of image characteristics on the algorithm performance for ≥ 2 instances. The characteristics' effect is displayed in the form of the $\log(OR)$, representing the logarithms of the odds of occurrence of the outcome in presence of the image characteristic, compared to the odds of the outcome occurring in the absence of that exposure. Significant effects are marked with an asterisk (∗: $p \leq 0.05$, ∗∗: $p \leq 0.01$, ∗∗∗: $p \leq 0.001$).

Table 3
Effect of the different inputs for training the Mask R-CNN, showing the mean (μ), median ($\widetilde{x}$), 5th, 25th, and 75th quartile (Q_1, Q_3), and the interquartile range (IQR) of the multi-instance dice coefficient DSC_{MI}. The names of the trained model T with the indices R, L and F are referring to R = raw, F = flow and L = likelihood, as detailed in the Appendix A. The model T^{+}_{RFL} is the same model as T_{RFL}, but followed by the post-processing.

Model	μ	$\widetilde{x}$	Q_5	Q_{25}	Q_{75}	IQR
T_R	0.50	0.44	0.00	0.27	0.83	0.56
T_{RF}	0.73	0.91	0.0	0.48	0.95	0.47
T_{RL}	0.79	0.93	0.24	0.63	0.96	0.33
T_{FL}	0.80	0.93	0.29	0.63	0.96	0.33
T_{RFL}	0.80	0.93	0.29	0.63	0.96	0.33
T^{+}_{RFL}	*0.81*	*0.94*	*0.32*	*0.75*	*0.96*	*0.21*

the performance by a factor of 2, from a median *MI_DSC* of 0.44 to 0.91 (mean: 0.50 to 0.73 (146%)). Including the instrument likelihood in the network model (T_{RL}) also increased the median *MI_DSC* by a factor of 2 (0.44 to 0.93 (211%); mean 0.50 to 0.79 (158%)). Incorporating both the flow and the instrument likelihood as additional input (T_{RFL}) did not yield a substantial improvement compared to T_{RL}. The post-processing significantly ($p = 2.7E-7$) increases the mean performance of the model T_{RFL} at 1%, while simultaneously reducing the IQR from 0.33 to 0.21. Also, the robustness of the method (defined in Roß et al. (2020) as the 5th percentile) increased from 0.28 to 0.32. Further descriptive statistics can be found in Table 3.

5. Discussion

While a lot of research is currently invested in maximizing algorithm performance for various image analysis tasks, comparatively little effort is currently put into failure analysis. This holds especially true for the growing field of benchmarking via challenges. Although challenge results potentially encode crucial information with respect to typical failure cases as well as reasons for algorithm failure, most challenge organizers restrict their reports to plain ranking tables (Wiesenfarth et al., 2021) leaving the rich challenge data unexploited. To address this gap in the literature, we presented a statistical framework for systematically learning from challenge results. Rather than performing a black-box prediction of algorithm performance, we focus on gaining insights into what might be the problems algorithms encounter in medical image analysis and, specifically, when and why they fail. The following sections discuss mixed models in medical imaging (Section 5.1), the general approach (Section 5.2), the specific findings for the sample application (Section 5.3) as well the new algorithm resulting from the statistical analysis (Section 5.4).

5.1. Mixed models in medical imaging

Mixed model analysis is a common approach in statistics. By analyzing more than 5000 papers published at the MICCAI conference, we found that only 1.7% of the papers have used mixed models for results analysis. In those papers, hierarchical models have primarily been proposed to discover variability at the group and population levels (e.g. by extracting patient-specific models (Swee and Grbić, 2014), for multi-group shape analysis (Shigwan and Awate, 2016) or to model solutions using decomposed tasks (e.g. in endovascular catheterization Rafii-Tari et al., 2014 and for generating patient specific segmentation Kutra et al., 2012). Furthermore, (linear) mixed effect models have mainly been used to account for the consistency of variables across repeated measures, to estimate distributions, or to explore individual and group differences (e.g. by learning spatiotemporal patterns on a network Koval et al., 2017, or by addressing inter/within-subject effects/changes between serial Magnetic Resonance Imaging (MRI) scans Kim et al., 2015). Other approaches used nonlinear mixed-effects to explore the patterns of early brain growth (Vardhan et al., 2017) or for the prediction of future observations based on past measurements and population statistics (Sadeghi et al., 2014). None of the papers used mixed model analysis to reveal the specific strengths and weaknesses of algorithms, as proposed in this work.

5.2. Framework for challenge analysis

In the field of biomedical image processing, failure analysis is often restricted to qualitative assessments (Wiesenfarth et al., 2021). Even quantitative analyses are typically restricted to single parameter assessments contrasting algorithm performance in the presence or absence of certain features (e.g. Soberanis-Mukul et al. (2020)). This

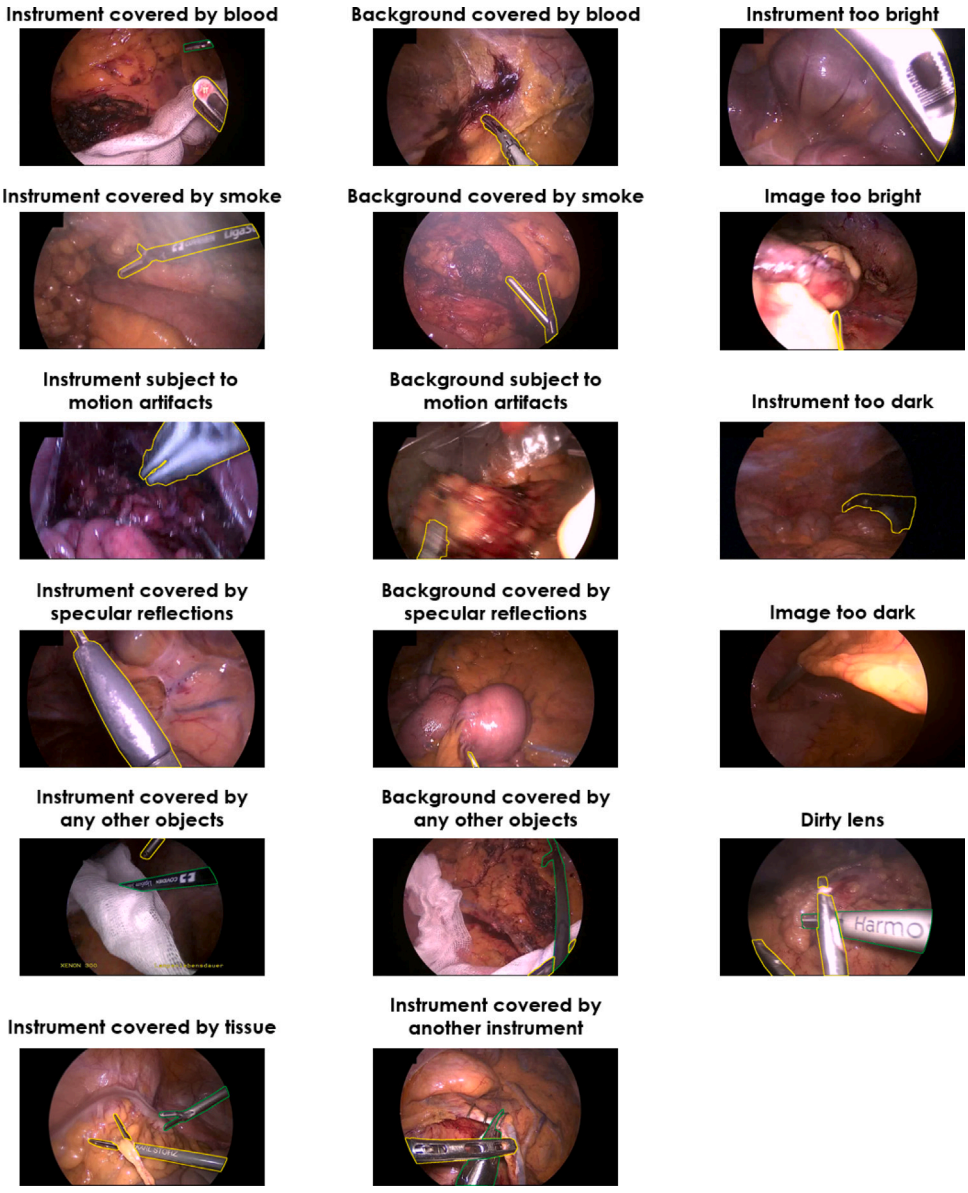

Fig. 8. Illustration of the 17 image characteristics that have a potential impact on algorithm performance.

approach should be used cautiously as it neglects possible strong interplay between features. We overcome these limitations by a statistically well-grounded approach. Leveraging the strengths of mixed models, we are able to disentangle the effects of different sources and clearly identify sources of algorithm failure. This enables the design of future algorithms dedicated to the actual needs.

A core strength of our approach can be seen in the fact that it is generically applicable regardless of domain (e.g. radiology, surgery, pathology) and algorithm category (e.g. model-based, machine learning-based). Furthermore, it is not restricted to challenge analysis but can also be used for the comprehensive performance assessment of an individual algorithm. In this case, the algorithm can be removed as fixed or random effect from Eq. (4) as detailed in Appendix C.

A limitation of our concept is the additional manual annotation effort involved. For comparison to the statistical approach, requiring this high amount of annotations, we performed an ablation study, in which five experts rated the influence of image characteristics of a subset of 100 images with poor performance (see Appendix D). The experiment revealed a high inter-rater variability with a diverse opinion on whether specific characteristics have a large negative influence or not. In addition, their ratings differ from the results of the statistical analysis. We therefore argue that the large amount of annotations and the statistical analysis will help to draw correct conclusions overcoming a confirmation bias related to the expert analysis. Depending on the specific application, the annotation effort could potentially be overcome by an automatic annotation of image characteristics, or by (quality-controlled) crowdsourcing (Maier-Hein et al., 2014b; Heim et al., 2018).

It should be noted that traditional feature importance measures, such as those applied to popular random forest-based methods, would merely provide a ranking of image characteristics. In contrast, we are interested in proper *statistical inference*. This approach comes with two key advantages:

1. **Awareness of uncertainties:** An image characteristic may have a large effect size but estimation uncertainty may likewise be large while another image characteristic may have a smaller effect with low estimation uncertainty, i.e. high precision. In this case, the latter image characteristic may be considered more relevant. Furthermore, in a small data set, estimates may solely

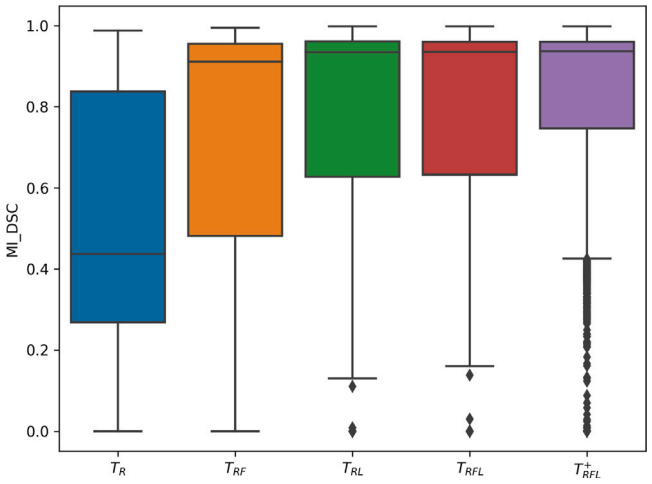

Fig. 9. Performance of the model for different input data. T_R: only (raw) input images; T_{RF} images and flow; T_{RL} images and instrument likelihood; T_{RFL} images, flow and likelihood. T^+_{RFL} indicates using an additional post-processing step after applying (T_{RFL}).

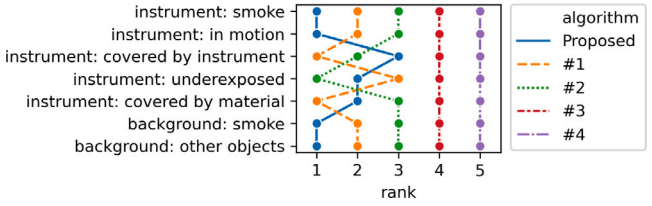

Fig. 10. Comparison of our method (T^+_{RFL}) with the four best-performing methods of the ROBUST-MIS challenge (#1-#4) for the seven characteristics that have a significant negative impact on algorithm performance according to our effect analysis. It can be seen that our proposed method (T^+_{RFL}) outperforms the competitors in most categories.

be due to chance. To address such issues, statistical tests and confidence intervals in the statistical model take the uncertainty and in particular the sample size into account in decision making.

2. **Awareness of data dependencies:** We are interested in the significance of the effects of image characteristics, which are entered into the model as fixed effects. However, random effects are used mainly to account for correlations in the data and avoid biases and violations of distributional assumptions when performing statistical tests.

It should further be noted that we instantiated the concept only for a single challenge, with specific focus on multi-instance medical instrument segmentation. However, as segmentation is by far the most widely used task in challenges (Maier-Hein et al., 2018), our concept can be transferred to a majority of studies and also create awareness of the problem that arises when using statistical tests for bounded metrics. Finally, it must be noted that statistical modeling is a complex process of balancing adherence to specific assumptions and model complexity. Delivering a recipe applicable to every challenge is not possible within the scope of a single paper but we believe that this work could trigger more sophisticated challenge analyses following the general approach presented here.

5.3. Instantiation for ROBUST-MIS challenge

In contrast to the literature (e.g. Maier-Hein et al. (2014a), Ali et al. (2021b, 2020, 2021a), Funke et al. (2018), Soberanis-Mukul et al. (2020), Bodenstedt et al. (2018), Allan et al. (2020) and Roß et al. (2020)), our study did not reveal a harmful effect of reflections, blood, and smoke on the algorithm performances. Instead, our analysis showed that the main limiting factors are when an instrument is in motion, underexposed, or covered by another instrument. Interestingly, other characteristics such as an instrument being overexposed or covered by tissue, reflections or blood, seem to even support the algorithm performances. That artifacts/characteristics are not always harmful is in line with a recent work from Kayser et al. (2020) who use reflections to improve the segmentation of polyps. The fact that the characteristic "instrument covered by tissue" has a slight positive effect on algorithm performance is most likely due to the fact that the visible tissue overlay mainly occurs when the instrument is clearly visible and distinguishable from the background.

While most of the characteristics either harm or benefit both metrics investigated, we identified one characteristic that yielded different results for precision and recall. Specifically, when the background contains other objects, we observe an increased recall but a decreased precision. This indicates that an oversegmentation occurred typically occurred in these cases.

The strongest negative impact by far was found when the instrument is being covered by another instrument. This can be attributed to the architecture of Mask R-CNNs and their way of processing images. A Mask R-CNN relies on a region proposal network that provides bounding boxes around regions of particular interest. Especially in regions where instruments overlap, those bounding boxes might contain parts of multiple instrument instances, which then leads to poor segmentations. This finding is in line with work on the Mask R-CNN problem of overlapping instances (Xu et al., 2020).

5.4. Algorithm development tailored to failure cases

With a few exceptions (Hasan and Linte, 2019; Isensee and Maier-Hein, 2020), the few methods published on multi-instance segmentation to date use a Mask R-CNN (González et al., 2020; Kletz et al., 2019) as core component. To our knowledge, Kletz et al. (2019) were the first to use a Mask R-CNN for surgical instrument segmentation. They developed their method for laparoscopic gynecology videos but reported limitations under conditions of occlusion and overlapping instruments. In the context of binary segmentation, several authors have aimed to tackle this issue by including temporal information in order to improve their segmentation, e.g., in situations where instruments are only partially visible, due to overlapping tissue. Jin et al. (2019) were the first who estimated the optical flow by using a CNN, namely UnFlow (Meister et al., 2018), for the segmentation of instruments. Instead of using the optical flow as an additional feature, they used it as a prior for initializing the attention of a temporal attention pyramid network to learn to focus on moving objects. However, their approach was for binary segmentation, type classification and instrument parts segmentation and classification, thus requiring a much simpler network architecture and no management of pixels that could possibly belong to different object instances.

The first to combine a Mask R-CNN and the optical flow were González et al. (2020), who proposed the use of a Mask-RCNN to ensure a segmentation and classification of instruments as, e.g., grasper or scissors. The authors computed the optical flow of previous frames to include a temporal consistency module and consider an instance's predictions across the frames in a sequence. The approach outperformed the state-of-the-art methods on the Endoscopic Vision 2017 and 2018 Robotic Instrument Segmentation data sets (Allan et al., 2019). However, this work was also not used for multi-instance segmentation.

Up to this point, and to the best of our knowledge, none of the approaches for multi-instance segmentation have successfully incorporated temporal information in the algorithm, as further reported by Roß et al. (2020). To address this gap in the literature, a new algorithm was developed that generated a benefit using optical flow in combination with instrument probability in order to explicitly address the previously mentioned weaknesses (see Appendix A). While we found a huge performance boost when integrating the flow as an additional input of

a native Mask R-CNN, the effect was negligible when also incorporating instrument likelikehood. In contrast, we achieved a high performance boost with an additional post-processing step dedicated to resolving ambiguities in the presence of overlapping instruments, as shown in Table 3.

The presented results suggest that typical challenges of laparoscopic videos, such as reflection, blood, and lighting variations, are already well manageable by state-of-the-art methods. However, difficulties with tube-like structures that are misclassified as instruments or transparent objects such as trocars persist. Furthermore, images with crossing or close instruments remain difficult to separate for both state-of-the-art methods and the presented approach, even though the latter was specially designed to manage such difficulties. One limiting factor may be seen in restrictions in the training and test data set, where only 8% of the images contain more than two instrument instances. Furthermore, only in rare cases do those instances overlap or intersect, thus resulting in only limited opportunities for training and evaluating the algorithm's separation capabilities.

It should be mentioned that real-time capability is an essential prerequisite for successful translation to a clinical setting. Currently, the Conditional Random Field (CRF) Appendix A.4 and the estimation of the optical flow would already approximately take more than 2 s per image. However, the method presented was merely an attempt to solve the multi-instance segmentation problem. The next step could be rendering the algorithm real-time capable.

We would further like to highlight that our main goal was not to produce a new state-of-the-art algorithm, instead, we wanted to show that the systematic analysis of challenge results potentially leads to actionable insights and thus to a significant improvement in the algorithm performances.

Overall, our methodology achieved a new best score on the ROBUST-MIS challenge data set. While we did not use the challenge test data to tune hyperparameters, it should be mentioned that we had access to the other participants' performance results to inform the strength–weakness driven algorithm development. This could still be seen as a competitive advantage. However, the primary aim of this study was not to present a new state-of-the-art method for instrument segmentation but a novel concept for learning from challenges. With this work, we have not only identified typical failure cases for the task of medical instrument segmentation but also showcased an entirely new way of problem-driven algorithm development based on insights gained through challenge results.

6. Conclusion

In conclusion, the proposed approach to leveraging meta data annotations for a mixed model-based analysis of challenge results opens up entirely new opportunities for systematically learning from challenges. By identification of characteristics that lead to algorithm failure it not only provides a much deeper understanding of the state-of-the-art for a given application but also enables tailoring future algorithm development to the actual remaining needs.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Part of this work was funded by the Helmholtz Imaging Platform (HIP), a platform of the Helmholtz Incubator on Information and Data Science and by the Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg.

Acknowledgments and conflicts of interest

The authors would like to thank Minu Dietlinde Tizabi and Alexander Seitel for proofreading the paper. Part of this work was funded by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Incubator (HI) and by the Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg. Finally, thanks to all authors of the ROBUST-MIS challenge manuscript (Roß et al., 2019), which served as foundation of this work.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.102765.

References

Al Hajj, H., Lamard, M., Conze, P.-H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., et al., 2019. CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. Med. Image Anal. 52, 24–41.

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al., 2021a. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. 102002.

Ali, S., Zhou, F., Bailey, A., Braden, B., East, J.E., Lu, X., Rittscher, J., 2021b. A deep learning framework for quality assessment and restoration in video endoscopy. Med. Image Anal. 68, 101900.

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., et al., 2020. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. 10 (1), 1–15.

Allan, M., Chang, P.-L., Ourselin, S., Hawkes, D.J., Sridhar, A., Kelly, J., Stoyanov, D., 2015. Image based surgical instrument pose estimation with multi-class labelling and optical flow. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 331–338.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al., 2020. 2018 Robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190.

Allan, M., Mcleod, J., Wang, C.C., Rosenthal, J.C., Fu, K.X., Zeffiro, T., Xia, W., Zhanshi, Z., Luo, H., Zhang, X., et al., 2021. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al., 2019. 2017 Robotic instrument segmentation challenge. Available Online at: arXiv preprint arXiv:1902.06426.

Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., et al., 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv preprint arXiv:1805.02475.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Funke, I., Bodenstedt, S., Riediger, C., Weitz, J., Speidel, S., 2018. Generative adversarial networks for specular highlight removal in endoscopic images. In: Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling. 10576, International Society for Optics and Photonics, 1057604.

García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: International Workshop on Computer-Assisted and Robotic Endoscopy. Springer, pp. 84–95.

González, C., Bravo-Sánchez, L., Arbelaez, P., 2020. Isinet: An instance-based approach for surgical instrument segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 595–605.

Hasan, S.K., Linte, C.A., 2019. U-NetPlus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 7205–7211.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: IEEE International Conference on Computer Vision. pp. 2961–2969.

Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A.W., et al., 2018. Large-scale medical image annotation with crowd-powered algorithms. J. Med. Imaging 5 (3), 034002.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks. pp. 2462–2470.

Isensee, F., Maier-Hein, K.H., 2020. OR-UNet: an optimized robust residual U-Net for instrument segmentation in endoscopic images. arXiv preprint arXiv:2004.12668.

Jin, Y., Cheng, K., Dou, Q., Heng, P.-A., 2019. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 440–448.

Kayser, M., Soberanis-Mukul, R.D., Zvereva, A.-M., Klare, P., Navab, N., Albarqouni, S., 2020. Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting. arXiv preprint arXiv:2002.02883.

Kim, H., Lepage, C., Evans, A.C., Barkovich, A.J., Xu, D., 2015. NEOCIVET: Extraction of cortical surface and analysis of neonatal gyrification using a modified CIVET pipeline. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 571–579.

Kletz, S., Schoeffmann, K., Benois-Pineau, J., Husslein, H., 2019. Identifying surgical instruments in laparoscopy using deep learning instance segmentation. In: IEEE International Conference on Content-Based Multimedia Indexing. IEEE, pp. 1–6.

Koval, I., Schiratti, J.-B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., Durleman, S., Initiative, A.D.N., et al., 2017. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 451–459.

Kutra, D., Saalbach, A., Lehmann, H., Groth, A., Dries, S.P., Krueger, M.W., Dössel, O., Weese, J., 2012. Automatic multi-model-based segmentation of the left atrium in cardiac MRI scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 1–8.

Lee, E.-J., Plishker, W., Liu, X., Kane, T., Bhattacharyya, S.S., Shekhar, R., 2019. Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework. In: Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling. International Society for Optics and Photonics, p. 109511T.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Commun. 9 (1), 5217.

Maier-Hein, L., Groch, A., Bartoli, A., Bodenstedt, S., Boissonnat, G., Chang, P.-L., Clancy, N., Elson, D.S., Haase, S., Heim, E., et al., 2014a. Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction. IEEE Trans. Med. Imaging 33 (10), 1913–1930.

Maier-Hein, L., Mersmann, S., Kondermann, D., Stock, C., Kenngott, H.G., Sanchez, A., Wagner, M., Preukschas, A., Wekerle, A.-L., Helfert, S., et al., 2014b. Crowdsourcing for reference correspondence generation in endoscopic images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 349–356.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al., 2020. BIAS: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Kisilenko, A., Müller, B., Davitashvili, T., Capek, M., Tizabi, M.D., Eisenmann, M., Adler, T.J., Gröhl, J., Schellenberg, M., Seidlitz, S., Lai, T.Y.E., Pekdemir, B., Roethlingshoefer, V., Both, F., Bittel, S., Mengler, M., Mündermann, L., Apitz, M., Kopp-Schneider, A., Speidel, S., Nickel, F., Probst, P., Kenngott, H.G., Müller-Stich, B.P., 2021. Heidelberg colorectal data set for surgical data science in the sensor operating room. Sci. Data 8 (1), 101. http://dx.doi.org/10.1038/s41597-021-00882-2.

McCulloch, C., Searle, S., Neuhaus, J., 2011. Generalized, Linear, and Mixed Models. In: Wiley Series in Probability and Statistics, Wiley.

Meister, S., Hur, J., Roth, S., 2018. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In: Thirty-Second AAAI Conference on Artificial Intelligence.

Mendrik, A.M., Aylward, S.R., 2019. A framework for challenge design: Insight and deployment challenges to address medical image analysis problems. arXiv preprint arXiv:1911.08531.

Rafii-Tari, H., Liu, J., Payne, C.J., Bicknell, C., Yang, G.-Z., 2014. Hierarchical HMM based learning of navigation primitives for cooperative robotic endovascular catheterization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 496–503.

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P.M., Bogunovic, H., Landman, B.A., Maier, O., Menze, B., et al., 2018. How to exploit weaknesses in biomedical challenge design and organization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 388–395.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G.-B., Bodenstedt, S., Bolmgren, J., Bravo-Sánchez, L., Chen, H., González, C., Guo, D., Halvorsen, P., Heng, P., Hosgor, E., Hou, Z., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K., Ni, Z., Riegler, M., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, B., Maier-Hein, L., 2020. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. Med. Image Anal. 101920. http://dx.doi.org/10.1016/j.media.2020.101920.

Roß, T., Reinke, A., Maier-Hein, L., 2019. Robust medical instrument segmentation (ROBUST-MIS) challenge (grand-challenge.org). https://robustmis2019.grand-challenge.org/. Accessed: 2019-10-29.

Sadeghi, N., Fletcher, P.T., Prastawa, M., Gilmore, J.H., Gerig, G., 2014. Subject-specific prediction using nonlinear population modeling: application to early brain maturation from DTI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 33–40.

Shigwan, S.J., Awate, S.P., 2016. Hierarchical generative modeling and Monte-Carlo EM in Riemannian shape space for hypothesis testing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 191–200.

Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I., 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning. In: IEEE International Conference on Machine Learning and Applications). IEEE, pp. 624–628.

Soberanis-Mukul, R.D., Albarqouni, S., Navab, N., 2020. Polyp-artifact relationship analysis using graph inductive learned representations. arXiv preprint arXiv:2009.07109.

Swee, J.K., Grbić, S., 2014. Advanced transcatheter aortic valve implantation (TAVI) planning from CT with ShapeForest. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 17–24.

Thode, H.C., 2002. Testing for Normality, Vol. 164. CRC Press.

Vardhan, A., Fishbaugh, J., Vachet, C., Gerig, G., 2017. Longitudinal modeling of multimodal image contrast reveals patterns of early brain growth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 75–83.

West, B.T., Welch, K.B., Galecki, A.T., 2014. Linear Mixed Models: A Practical Guide using Statistical Software. Crc Press.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11 (1), 1–15.

Xu, S., Lan, S., Qi, Z., 2020. MaskPlus: Improving mask generation for instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2030–2038.