Feature Mixing Approach for Detecting Intraoperative Adverse Events in Laparoscopic Roux-en-Y Gastric Bypass Surgery

Rupak Bose[1], Chinedu Innocent Nwoye[1,4,*[0000−0003−4777−0857]], Jorge F. Lazo[1[0000−0001−5508−0747]], Joël L. Lavanchy[2,3,+[0000−0003−0248−4996]], and Nicolas Padoy[1,4,+[0000−0002−5010−4137]]

[1] ICube, UMR7357, CNRS, INSERM, University of Strasbourg, France
[2] University Digestive Health Care Center, Clarunis, Basel, Switzerland
[3] Dept. of Biomedical Engineering, University of Basel, Switzerland
[4] IHU Strasbourg, France
+ Co-last authors * Corresponding author

Abstract. Intraoperative adverse events (IAEs), such as bleeding or thermal injury, can lead to severe postoperative complications if undetected. However, their rarity results in highly imbalanced datasets, posing challenges for AI-based detection and severity quantification. We propose BetaMixer, a novel deep learning model that addresses these challenges through a Beta distribution-based mixing approach, converting discrete IAE severity scores into continuous values for precise severity regression (0-5 scale). BetaMixer employs Beta distribution-based sampling to enhance underrepresented classes and regularizes intermediate embeddings to maintain a structured feature space. A generative approach aligns the feature space with sampled IAE severity, enabling robust classification and severity regression via a transformer. Evaluated on the MultiBypass140 dataset, which we extended with IAE labels, BetaMixer achieves a weighted F1 score of 0.76, recall of 0.81, PPV of 0.73, and NPV of 0.84, demonstrating strong performance on imbalanced data. By integrating Beta distribution-based sampling, feature mixing, and generative modeling, BetaMixer offers a robust solution for IAE detection and quantification in clinical settings.

Keywords: Intraoperative adverse events · bleeding detection · bleeding quantification · surgical injury detection · gastric bypass surgery

1 Introduction

Intraoperative adverse events (IAEs), such as bleeding, thermal injury, and mechanical injury, are rare but critical occurrences during surgery that can lead to severe postoperative complications, including infection, organ dysfunction, or even mortality. These events not only jeopardize patient safety but also increase healthcare costs due to prolonged recovery times. Timely detection and accurate IAE severity quantification are crucial for enabling prompt intervention

and improving surgical outcomes [17]. However, traditional manual monitoring by surgical teams is prone to human error, underscoring the need for automated, real-time detection systems [16]. Artificial Intelligence (AI) has emerged as a promising tool for automating IAE recognition and quantification, offering real-time feedback to surgeons [6]. Despite this potential, the rarity of IAEs results in highly imbalanced datasets, which pose significant challenges for training effective AI models. Standard detection techniques often struggle with such imbalances, and the inherent complexity of surgical procedures further complicates the identification of deviations from normal workflow. Additionally, quantifying the severity of IAEs—ranging from mild to critical—is essential for determining the appropriate surgical response, yet this aspect remains underexplored in existing literature. Current approaches to surgical anomaly detection often fail to address the diversity of IAEs or the importance of severity quantification, limiting their practical utility in clinical settings [2].

To bridge this gap, we propose BetaMixer, a deep learning-based framework designed for both IAE classification and their severity regression during laparoscopic Roux-en-Y gastric bypass surgery. BetaMixer leverages a Beta distribution to transform discrete IAE labels into continuous variables, enabling precise severity regression on a 0 to 5 scale. To address class imbalance, the model employs Beta distribution-based sampling and regularizes intermediate embeddings to maintain a structured feature space. A generative component aligns the feature space with continuous severity labels, while a transformer network classifies IAEs and regresses their severity using a mean squared error (MSE) loss on sampled predictions and ground truths.

We evaluate BetaMixer on the MultiBypass140 dataset [13], extended with IAE annotations. Our model achieves state-of-the-art performance, particularly for rare IAEs, with a weighted F1 score of 0.76, recall of 0.81, PPV of 0.73, and NPV of 0.84. These results highlight the importance of temporal context and continuous feature space modeling for accurate IAE detection and regression.

Our contribution is fourfold: (1) A unified framework for IAE classification and severity regression in Roux-en-Y gastric bypass surgery. (2) A Beta distribution-based method to model continuous severity from discrete annotations, addressing class imbalance via MSE loss. (3) A generative component to normalize feature space distributions, aligning with severity labels. (4) Superior results over baselines, with temporal models outperforming frame-based approaches, significantly improving IAE detection and regression across metrics.

2 Related Work

Detecting and mitigating Intraoperative Adverse Events (IAEs) is vital for improving surgical outcomes [6]. IAEs, arising from factors like human error, equipment malfunction, and patient responses, are rare in large video datasets, making their detection challenging. They are often annotated with class labels [20, 9], with their scale of severity often overlooked. The discrete nature of these annotations doesn't capture the continuous development of IAEs. Studies in other fields

suggest that using distributions like the Beta distribution better models severity uncertainty [15]. For their task proximity, anomaly detection methods, including supervised [11, 4], semi-supervised [18], and unsupervised [1] approaches, are employed for IAE detection. Traditional models combine CNNs and RNNs for spatial and temporal tasks, while Transformers [5] have recently outperformed them in capturing long-range dependencies. However, most existing work focuses on event classification [20, 7], leaving the quantification of IAE severity, especially for conditions like intraoperative bleeding and injuries, largely unexplored. While resampling and reweighting techniques such as SMOTE [3] and focal loss [14] address label imbalance, our regression-based smoothing offers an orthogonal alternative.

3 Methods

This section presents the methodology for developing *BetaMixer*, a deep learning model designed for the classification and regression of the severity of intraoperative adverse events (IAEs) in Roux-en-Y gastric bypass surgeries. The goal is to enhance timely surgical intervention and improve clinical outcomes by accurately detecting IAEs and assessing their severity.

3.1 Problem Definition

Given a surgical video $V = \{f_1, f_2, \ldots, f_n\}$ consisting of n sequential frames, the task is to predict the IAE class $\mathcal{C}$ and severity $\mathcal{S}$ for each frame $f_i \in V$. Here, $\mathcal{C} \in \{\text{BL}, \text{MI}, \text{TI}\}$ represents the IAE categories (bleeding, mechanical injury, and thermal injury), and $\mathcal{S} \in \{0, 1, \ldots, m\}$ denotes the severity level, where m varies by IAE type (e.g., $m = 5$ for bleeding). To predict $(\mathcal{C}_i, \mathcal{S}_i)$ for frame f_i, we utilize a sequence of contiguous frames from the past k time steps, i.e., $X_i = \{f_{i-k+1}, \ldots, f_i\}$. The model learns a function $\mathcal{F}(X_i) \rightarrow (\mathcal{C}_i, \mathcal{S}_i)$ that maps the current frame f_i to its IAE class and severity, leveraging temporal information from the preceding k frames.

3.2 Dataset

We utilize the MultiBypass140 dataset [13], which comprises 140 patient cases of laparoscopic Roux-en-Y gastric bypass surgery. This dataset has been extended with fine-grained annotations for IAEs, including their category and severity [12], see Table 1. The annotations were performed by a board-certified surgeon with over 10 years of visceral surgery experience, guided by the SEVERE index manual [10] to ensure consistency and accuracy. The dataset includes 782K frames extracted at 1 fps, with 780K frames labeled as normal and 1,594 frames annotated with IAEs (bleeding, mechanical injury, and thermal injury). Each event is labeled with start and end times, along with severity score ranging from 0 to 5, where higher values indicate greater severity. The dataset is split into 80 videos for training, 20 for validation, and 40 for testing. Table 2 provides the distribution of IAEs across clinical centers.

Table 1: Clinical definition of IAE severity scores.

Severity	Bleeding	Thermal injury	Mechanical injury
1	Very low amount of blood lost	Superficial penetration to “less vital” tissue	Superficial penetration to "less vital" tissue, needle poke to tissue
2	Low amount of blood lost	Deep penetration to “less vital” tissue or any organ/tissue subjected to planned resection	Full-thickness injury
3	Intermediate amount of blood lost	Superficial penetration to “vital” tissue	Superficial penetration to "vital" tissue
4	High amount of blood lost	Deep penetration to “vital” tissue to the level of muscularis/parenchyma	Deep penetration to “vital” tissue
5	Very high amount of blood lost	Through and through injury to hollow organ or deeper parenchymal injury to solid organ	Through and through injury to "vital" tissue

Table 2: Distribution of the IAEs across clinical centers present in the dataset.

Center	Cases	# Frames	Normal	Bleeding	Mechanical Injury	Thermal Injury
Strasbourg	70	464,973	426983	33634	3674	682
Bern	70	316,646	282204	28068	5691	683

3.3 Discrete to Continuous Severity Distribution

The severity levels in the dataset are annotated as discrete values, which are inherently imbalanced, as seen in Table 2. However, in real-world scenarios, severity exists on a continuous spectrum. To better capture this variability and address annotation noise, we propose transforming the discrete ordinal numbers into a continuous distribution using the Beta distribution. The Beta distribution was chosen due to its mathematical properties and flexibility in modeling probabilistic severity scores on the normalized [0,1] interval, capturing a wide range of distributions and reflecting the inherent variability and uncertainty in clinical annotations. The Beta distribution, defined on the interval [0, 1], is parameterized by two shape parameters, α and β, computed as:

$$\alpha = \mu^2 \times \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu} \right), \quad \beta = \alpha \times \frac{1}{\mu} - 1, \tag{1}$$

where μ represents the mean and σ represents the standard deviation of the distribution. These parameters enable the Beta distribution to model smooth transitions between severity levels, making it suitable for handling sparse or noisy data. During training, we generate continuous severity values from the Beta distribution, providing a probabilistic representation that captures both the clinician's initial assessment and the inherent variability in quantification.

3.4 Model Architecture

The architecture of *BetaMixer* is presented in Fig. 1 and consists of a backbone, feature generator and discriminator, IAE encoder, classifier, and regressor.
Backbone Feature Extractor: For each frame f_i in the input sequence X, a backbone feature extractor $\mathcal{B}$ (MobileNetV2 initialized with ImageNet weights) processes the frame and outputs a feature vector $\hat{f}_i \in \mathbb{R}^d$, where d is the dimensionality of the feature space.

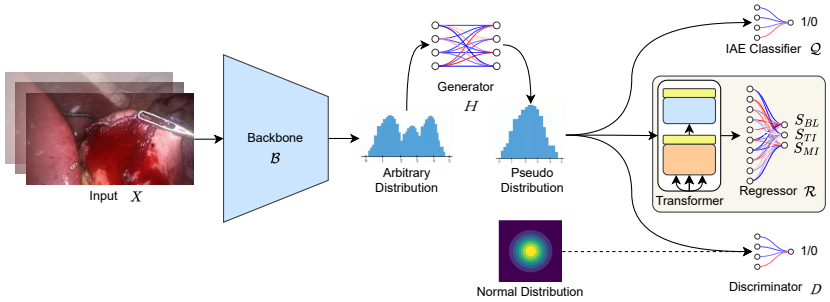

Fig. 1: Overview of *BetaMixer*: The backbone $\mathcal{B}$ extracts features, which are transformed into a normal distribution by the generator $\mathcal{H}$. A transformer with positional embeddings encodes, classifies, and regresses IAE severity, while the discriminator $\mathcal{D}$ ensures feature normalization.

Normalized Feature Generator and Discriminator: A generator $\mathcal{H}$ transforms the backbone features into normally distributed features $\tilde{f}_i \sim \mathcal{N}(0,1)$. The generator is implemented as a fully connected neural network (FCNN) with learnable parameters θ. A discriminator $\mathcal{D}$ enforces that the generated features follow a standard normal distribution by classifying them as real or fake.
Transformer Encoder with Regression Tokens: The generated features $\tilde{f}_i$ are passed through a transformer encoder, which incorporates positional embeddings and three regression tokens $T_{\mathrm{BL}}, T_{\mathrm{MI}}, T_{\mathrm{TI}}$ for bleeding, mechanical injury, and thermal injury, respectively. The transformer applies self-attention to capture temporal dependencies across frames.
IAE Classifier and Regressor: The pooled features are passed into a binary classifier $\mathcal{Q}$ for IAE detection. The regression tokens are used to predict continuous severity score for each IAE using the regression module. We use multiple regression heads to enable the detection of overlapping events with varying severity levels.

4 Experimental Setup

Implementation and Training. We use MobileNetV2 (feature level 5) as the backbone for feature extraction. The discriminator and generator in *BetaMixer* are implemented as fully convolutional networks, while the IAE classifier consists of a single convolutional layer, adaptive average pooling, and a linear layer. The transformer encoder has a projection depth of 128 with 4 transformation layers. The model is trained adversarially for the discriminator and generator, followed by freezing the generator to train the remaining components for 30 epochs. Three loss functions are used: the adversarial loss trains the discriminator to distinguish between real and generated features, while the generator aims to fool the discriminator. The classification loss trains the IAE classifier to detect adverse events using binary cross-entropy. The sampled regression loss trains the

Table 3: IAE classification results of our model in comparison of existing methods and baselines on 5 seconds clip over the whole testing set

Model	Bleeding			Mechanical injury			Thermal injury			Overall IAE		
	F1	PPV	NPV	F1	PPV	NPV	F1	PPV	NPV	F1	PPV	NPV
ResNet18 [8]	0.72	0.71	0.77	0.63	0.62	0.59	0.70	0.72	0.66	0.68	0.66	0.67
MobileNetV2 [19]	0.71	0.72	0.76	0.61	0.66	0.61	0.72	0.70	0.71	0.68	0.68	0.69
sMSTCN [20]	0.75	0.71	0.77	0.64	0.65	0.61	0.75	0.71	**0.79**	0.71	0.69	0.72
FRCNN [9]	0.76	0.73	0.78	0.68	0.67	0.60	0.74	0.72	0.77	0.72	0.70	0.71
Ours (Genless)	0.72	0.71	0.76	0.65	0.66	0.59	0.72	0.71	0.73	0.69	0.69	0.70
Ours	**0.81**	**0.76**	**0.82**	**0.70**	**0.69**	**0.63**	**0.77**	**0.74**	0.77	**0.76**	**0.73**	**0.84**

transformer to predict severity scores by minimizing mean squared error (MSE) on Beta-distributed ground truth labels, which are uniformly sampled across event classes and severity levels. We train our model on batches of 128×128 spatial resolution images (batch size = 32), which offers a practical balance between performance and computational cost. Optimization is performed using Adam with a learning rate $\lambda = 5 \times 10^{-5}$. All experiments are conducted on a single RTX3060 GPU using the MultiBypass140 dataset.

Baselines. We evaluate BetaMixer against four baselines: two frame-based models (ResNet18 [8] and MobileNetV2 [19]) and two temporal-based models (sMSTCN [20] and FRCNN [9]). ResNet18 is chosen for its deep residual structure, which effectively learns complex features, while MobileNetV2 is selected for its lightweight design, making it suitable for on-device inference. The temporal-based models have been previously explored for IAE detection tasks [9, 20]. All baselines are adapted to support severity regression using an extra linear layer.

Evaluation Metrics. We evaluate *BetaMixer* using standard metrics, including F1 score, recall, and Mean Squared Error (MSE), to assess performance on imbalanced datasets. The F1 score is weighted with severity-based weights $[0.02, 0.06, 0.12, 0.19, 0.26, 0.33]$ to balance sensitivity and specificity. For continuous severity prediction, thresholds of 0.5 (classification) and $(0.2, 0.4, 0.6, 0.8)$ (regression) are applied. Additionally, we use clinically relevant metrics: Classification Delay Time (CDT), Positive Predictive Value (PPV), and Negative Predictive Value (NPV). CDT measures the delay between the first occurrence of an event and its correct prediction. PPV evaluates the accuracy of predicting severe events (levels ≥ 3), while NPV assesses the prediction of non-severe events (levels ≤ 1), ensuring minimal unnecessary interventions.

5 Results and Discussion

Our proposed *BetaMixer* model demonstrates superior performance in IAE classification and severity regression compared to existing baselines, as shown in Tables 3 and 4. Overall, *BetaMixer* achieves a +4% improvement in F1 score,

Table 4: Result of IAE classification (F1/recall) and regression (MSE)

Model	F1 ↑	Recall ↑	MSE ↓
ResNet18 [8]	0.68±0.12	0.76±0.15	0.30±0.14
MobileNetV2 [19]	0.68±0.18	0.74±0.13	0.32±0.20
sMSTCN [20]	0.71±0.14	0.79±0.12	0.28±0.16
FRCNN [9]	0.72±0.15	0.78±0.16	0.26±0.19
Ours	**0.76±0.12**	**0.81±0.14**	**0.23±0.15**

Table 5: IAE regression results of our model for k = 5 for different severity levels in terms of mean squared error on the testing set

Model	Bleeding					Mechanical injury			Thermal injury			
	0	1	2	3	4	0	1	3	0	1	3	5
ResNet18 [8]	0.47	0.26	0.31	0.28	0.29	0.27	0.19	0.27	0.25	0.20	0.20	0.69
MobileNetV2 [19]	0.51	0.27	0.31	0.29	0.30	0.29	0.23	0.29	0.27	0.23	0.21	0.68
sMSTCN [20]	0.35	0.25	0.32	0.30	0.29	0.21	0.22	0.20	0.23	0.17	0.23	0.63
FRCNN [9]	0.35	0.22	0.30	0.28	0.27	0.19	0.15	0.21	0.19	0.18	0.23	0.65
Ours (Genless)	0.34	0.23	0.30	0.29	0.28	0.22	0.14	0.21	0.19	0.17	0.24	0.61
Ours	0.30	0.21	0.26	0.24	0.25	0.19	0.10	0.18	0.17	0.14	0.19	0.57

+3% in recall, +3% in PPV, and +13% in NPV. Notably, it excels in detecting and quantifying bleeding and mechanical injury, achieving the best scores across all metrics. For thermal injury, while the NPV is slightly lower (−2%) compared to FRCNN, *BetaMixer* outperforms in PPV and F1 score, likely due to the presence of smoke from coagulation tools, which serves as a strong visual indicator for this IAE. In terms of severity regression (Table 5), *BetaMixer* consistently performs well across all IAE categories. It achieves the lowest mean squared error (MSE) of 0.2 for level 1 bleeding and 0.1 for level 1 mechanical injury. However, predicting higher-severity thermal injury (level 5) remains challenging due to the absence of such samples in the training set. This highlights the need for more diverse training data to improve performance on rare, high-severity events.

The Classification Delay Time (CDT) metric, evaluated in Table 6, further underscores the effectiveness of *BetaMixer*. It achieves the lowest CDT for bleeding (1.31) and mechanical injury (1.12), outperforming sMSTCN and FRCNN. For thermal injury, the CDT is slightly higher, indicating room for improvement in detecting this specific IAE. These results demonstrate the model's ability to provide timely predictions, which is critical for intraoperative decision-making.

An ablation study on the impact of input sequence length (Table 7) reveals that a 5-frame input yields the best performance across all IAE categories, with an F1 score of 0.76, PPV of 0.73, and NPV of 0.84. This suggests that IAEs are temporal events best captured within short intervals, as performance degrades with longer or shorter sequences. Evaluating *BetaMixer* without the generator component (*Ours (Genless)* in Table 5) reveals a performance drop, underscoring the generator's role in feature normalization and overall accuracy improvement.

Table 6: Classification delay time (in secs) of IAE using 5 seconds clip window.

Model	Bleeding	Mechanical injury	Thermal injury	Overall IAE
ResNet18	1.53	1.51	1.23	1.42
MobileNet	1.40	1.41	1.21	1.34
sMSTCN	1.43	1.42	1.10	1.31
FRCNN	1.41	1.23	**0.91**	1.27
Ours	**1.31**	**1.12**	1.13	**1.23**

Table 7: Performance of our model on the length of clips

Frames	Bleeding			Mechanical injury			Thermal injury			Overall IAE		
	F1	PPV	NPV	F1	PPV	NPV	F1	PPV	NPV	F1	PPV	NPV
1	0.78	0.72	0.79	0.65	0.67	0.60	0.71	0.71	0.70	0.71	0.70	0.69
5	**0.81**	**0.76**	**0.82**	**0.70**	**0.69**	**0.63**	**0.77**	**0.74**	**0.77**	**0.76**	**0.73**	**0.84**
10	0.77	0.74	0.80	0.68	0.66	0.61	0.73	0.70	0.74	0.73	0.70	0.74
25	0.75	0.75	0.80	0.67	0.66	0.60	0.74	0.69	0.73	0.72	0.70	0.71

Qualitative results in Fig. 2 further validate that *BetaMixer* more accurately approximates ground truth in both classification and severity regression compared to baselines. From these observations, *BetaMixer* sets a new benchmark for IAE detection and severity regression, demonstrating robustness in handling imbalanced datasets and providing timely, accurate predictions. Its performance underscores the importance of temporal modeling and continuous severity representation in surgical AI systems.

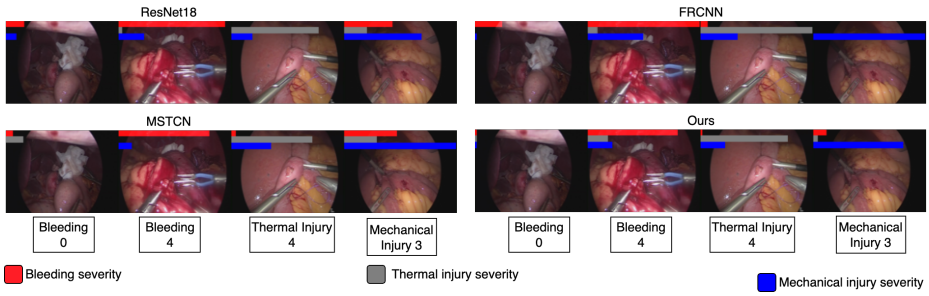

Fig. 2: Comparison of model predictions with the baselines. Bar length indicates severity. Groundtruth in box.

6 Conclusions

This paper addresses the challenge of classifying and quantifying intraoperative adverse events (IAEs), such as bleeding, thermal injury, and mechanical injury, during laparoscopic gastric bypass surgery. To tackle data imbalance caused by

the rarity of these events, we propose *BetaMixer*, a novel approach integrating normalized feature mixing, Beta distribution-based sampling, and continuous feature space regularization. Our method outperforms baselines in IAE classification and severity regression, achieving superior performance across automated and clinical metrics. An ablation study reveals optimal performance with 5-frame inputs, emphasizing the importance of temporal modeling for short-duration events. *BetaMixer* quantifies adverse events on a 0 to 5 scale with high accuracy, mitigating data imbalance and providing a robust solution for real-time IAE detection and severity assessment. While focused on Roux-en-Y gastric bypass due to data availability, future work will explore generalizability to other surgical domains and incorporate additional data sources to enhance performance. To foster research in this direction, we released our newly generated IAE annotations as part of the public MultiBypass140 dataset.

Acknowledgments. This work was supported by French state funds managed within the Plan Investissements d'Avenir by the ANR under references: National AI Chair AI4ORSafety [ANR-20-CHIA-0029-01], IHU Strasbourg [ANR-10-IAHU-02] and by BPI France [Project 5G-OR]. Lavanchy J.L. acknowledges funding by the Swiss National Science Foundation [P500PM206724, P5R5PM217663]. This work has also received funding from the European Union (ERC, CompSURG, 101088553). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Audibert, J., Michiardi, P., Guyard, F., Marti, S., Zuluaga, M.A.: Usad: Unsupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 3395–3404 (2020)
2. Beyersdorffer, P., Kunert, W., Jansen, K., Miller, J., Wilhelm, P., Burgert, O., Kirschniak, A., Rolinger, J.: Detection of adverse events leading to inadvertent injury during laparoscopic cholecystectomy using convolutional neural networks. Biomedical Engineering/Biomedizinische Technik **66**(4), 413–421 (2021)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
4. Checcucci, E., Piazzolla, P., Marullo, G., Innocente, C., Salerno, F., Ulrich, L., Moos, S., Quarà, A., Volpi, G., Amparore, D., et al.: Development of bleeding artificial intelligence detector (blair) system for robotic radical prostatectomy. Journal of Clinical Medicine **12**(23), 7355 (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

6. Eppler, M.B., Sayegh, A.S., Maas, M., Venkat, A., Hemal, S., Desai, M.M., Hung, A.J., Grantcharov, T., Cacciamani, G.E., Goldenberg, M.G.: Automated capture of intraoperative adverse events using artificial intelligence: A systematic review and meta-analysis. Journal of Clinical Medicine **12**(4), 1687 (2023)
7. Gawria, L., Rosenthal, R., van Goor, H., Dell-Kuster, S., ten Broek, R., Rosman, C., Aduse-Poku, M., Aghlamandi, S., Bissett, I., Blanc, C., et al.: Classification of intraoperative adverse events in visceral surgery. Surgery **171**(6), 1570–1579 (2022)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
9. Hua, S., Gao, J., Wang, Z., Yeerkenbieke, P., Li, J., Wang, J., He, G., Jiang, J., Lu, Y., Yu, Q., et al.: Automatic bleeding detection in laparoscopic surgery based on a faster region-based convolutional neural network. Annals of Translational Medicine **10**(10) (2022)
10. Jung, J.J., Jüni, P., Gee, D.W., Zak, Y., Cheverie, J., Yoo, J.S., Morton, J.M., Grantcharov, T.: Development and evaluation of a novel instrument to measure severity of intraoperative events using video data. Annals of surgery **272**(2), 220–226 (2020)
11. Kawachi, Y., Koizumi, Y., Harada, N.: Complementary set variational autoencoder for supervised anomaly detection. In: 2018 ICASSP. pp. 2366–2370. IEEE (2018)
12. Lavanchy, J.L., Alapatt, D., Sestini, L., Kraljević, M., Nett, P.C., Mutter, D., Müller-Stich, B.P., Padoy, N.: Analyzing the impact of surgical technique on intraoperative adverse events in laparoscopic Roux-en-Y gastric bypass surgery by video-based assessment. Surgical Endoscopy (2025). https://doi.org/10.1007/s00464-025-11557-z
13. Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. International journal of computer assisted radiology and surgery pp. 1–9 (2024)
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
15. Mariooryad, S., Busso, C.: The cost of dichotomizing continuous labels for binary classification problems: Deriving a bayesian-optimal classifier. IEEE Transactions on Affective Computing **8**(1), 119–130 (2015)
16. Mitchell, I., Schuster, A., Smith, K., Pronovost, P., Wu, A.: Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after 'to err is human'. BMJ quality & safety **25**(2), 92–99 (2016)
17. Nwoye, C.I.: Deep learning methods for the detection and recognition of surgical tools and activities in laparoscopic videos. Ph.D. thesis, Université de Strasbourg (2021)
18. Ruff, L., Vandermeulen, R.A., Görnitz, N., Binder, A., Müller, E., Müller, K.R., Kloft, M.: Deep semi-supervised anomaly detection. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=HkgH0TEYwH
19. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)
20. Wei, H., Rudzicz, F., Fleet, D., Grantcharov, T., Taati, B.: Intraoperative adverse event detection in laparoscopic surgery: stabilized multi-stage temporal convolutional network with focal-uncertainty loss. In: Machine Learning for Healthcare Conference. pp. 283–307. PMLR (2021)