ORIGINAL ARTICLE

Endoscopic sleeve gastroplasty: stomach location and task classification for evaluation using artificial intelligence

James Dials[1] · Doga Demirel[1] · Reinaldo Sanchez-Arias[2] · Tansel Halic[3] · Suvranu De[4] · Mark A. Gromski[5]

Abstract

Purpose We have previously developed grading metrics to objectively measure endoscopist performance in endoscopic sleeve gastroplasty (ESG). One of our primary goals is to automate the process of measuring performance. To achieve this goal, the repeated task being performed (grasping or suturing) and the location of the endoscopic suturing device in the stomach (Incisura, Anterior Wall, Greater Curvature, or Posterior Wall) need to be accurately recorded.

Methods For this study, we populated our dataset using screenshots and video clips from experts carrying out the ESG procedure on ex vivo porcine specimens. Data augmentation was used to enlarge our dataset, and synthetic minority over-sampling (SMOTE) to balance it. We performed stomach localization for parts of the stomach and task classification using deep learning for images and computer vision for videos.

Results Classifying the stomach's location from the endoscope without SMOTE for images resulted in 89% and 84% testing and validation accuracy, respectively. For classifying the location of the stomach from the endoscope with SMOTE, the accuracies were 97% and 90% for images, while for videos, the accuracies were 99% and 98% for testing and validation, respectively. For task classification, the accuracies were 97% and 89% for images, while for videos, the accuracies were 100% for both testing and validation, respectively.

Conclusion We classified the four different stomach parts manipulated during the ESG procedure with 97% training accuracy and classified two repeated tasks with 99% training accuracy with images. We also classified the four parts of the stomach with a 99% training accuracy and two repeated tasks with a 100% training accuracy with video frames. This work will be essential in automating feedback mechanisms for learners in ESG.

Keywords Endoscopic simulator · Endoscopic sleeve gastroplasty · Anatomical localization · Data augmentation · Deep learning · Artificial intelligence

Introduction

The endoscopic sleeve gastroplasty (ESG) procedure lacks an objective automated training platform. Currently, the primary

✉ Doga Demirel
  ddemirel@floridapoly.edu

1 Department of Computer Science, Florida Polytechnic University, 4700 Research Way, Lakeland, FL 33805, USA

2 Department of Data Science and Business Analytics, Florida Polytechnic University, Lakeland, FL, USA

3 Intuitive Surgical, Peachtree Corners, GA, USA

4 FAMU-FSU College of Engineering, Tallahassee, FL, USA

5 Division of Gastroenterology and Hepatology, Indiana University School of Medicine, Indianapolis, IN, USA

training method is ad hoc training by physicians participating in industry-sponsored or society-sponsored courses that teach the steps of endoscopic suturing and ESG, usually initially in inanimate or ex vivo porcine specimens. Some learners undergo proctored fellowships in endoscopic bariatric endoscopy, but this is a rare exception. There is a need for more repetitive hands-on training with tailored feedback. Our long-term goal is to create a virtual reality (VR) simulator for training endoscopists to perform the ESG procedure with automated assessment to fulfill the need for a more effective training method.

Our previous studies were executed in the stride toward this goal, where we created a task analysis and objective performance metrics for the ESG procedure [1, 2]. As we are building our simulator, we noticed that an automated

approach to scoring the surgeons was missing from the literature, and the ESG procedure was currently being manually graded. We want to introduce an approach for automatic grading of the ESG procedure to the literature. Even though there are different approaches to carry out the ESG procedure, marking and suturing tasks are the two vital steps each surgeon must complete. It can be difficult during the procedure, with distortion/suturing of the stomach into a tubular organ, to keep orientation clear. To automate the VR trainer simulator grading, we create classifiers that can identify the different parts of the stomach and the two tasks in the ESG procedure. To grade the ESG procedure, four main parts of the stomach are necessary to identify: the angularis incisura, the starting point of the ESG procedure, the anterior wall, the greater curvature, and the posterior wall. During the procedure, the endoscopist performs two main tasks: grasping and suturing, in the correct order. Manual assessment is a labor-intensive and time-consuming process. Automatic grading provides instant and objective feedback to trainees. It eliminates the need for a manual scoring process, saving time and allowing for a seamless training experience. Automatic scoring can be more accurate and consistent, ensuring fair evaluation for all trainees.

Additionally, it enhances the learning process by offering immediate feedback, allowing trainees to understand their mistakes in real time. It also enables adaptive learning, where the difficulty level can be adjusted based on the trainee's performance, optimizing the training experience. Data shows the amount of stomach tubularized does correlate with weight loss in ESG procedures [3]. Thus, detecting the device's location in the stomach via this technique would provide immediate feedback to the learner. Also, it has been previously described that procedure time for ESG is shorter, and efficacy is improved by not suturing the fundus during ESG [4]. Thus, the analysis of location for this concept is also vital for the learner.

The contribution of this study is threefold; data augmentation to create a larger dataset, synthetic minority oversampling (SMOTE) for balancing the dataset, and convolutional neural networks (CNN) for classification. The classification included tasks: (a) grasping and (b) suturing, and stomach locations: (a) angularis incisura, (b) anterior wall, (c) greater curvature, and (d) posterior wall for image and video datasets. Data augmentation is a process used to expand a dataset using techniques such as rotating, flipping, and cropping deep learning input images. Data augmentation is effective in image classification by increasing the dataset and reducing overfitting [5]. SMOTE is an oversampling technique used to balance the dataset. Having an imbalanced dataset causes bias within a model [6]. CNNs are neural networks where the layers are the convolutions of their previous layers [7]. The CNN will start by learning attributes such as

edges and hue, then slowly learning full objects over multiple layers/filters [8]. As mentioned above, we will use CNNs to identify specific parts of the stomach and which task is being performed to classify four parts of the stomach and differentiate between grasping or suturing tasks.

This work's literature spans multiple techniques, including synthetic data generation, classification, and localization. In a paper by Hussain et al. [9], they compared different data augmentation techniques on medical data to determine which strategy works the best on medical data. They determined that shear had the highest training accuracy (89%), while Gaussian filters (88%) and rotation (88%) had the highest validation accuracy. All the techniques ranged in the 80 s in accuracy except for noise and powers. Although they showed which techniques resulted in higher accuracy, none of the accuracies reached 90%, which shows that some other portion of their study is lacking. One potential cause is that the study does not specify the type of medical imaging used, which can result in a lack of focus and reduced classification accuracy.

In a study by Mikolajczyk et al. [10], data augmentation is used to enlarge a dataset of three medical case studies: skin melanoma diagnosis, histopathological images, and breast magnetic resonance imaging (MRI) scans. They used a new data augmentation method claiming that although standard data augmentation methods are proven to increase the training dataset, they are susceptible to adversarial attacks. In our study, we are not particularly concerned with this issue. Taylor et al. [11] performed a benchmark study on the different types of data augmentation. They found that geometric augmentation methods outperformed photometric methods when training on a coarse-grained data set. This study backs the use of geometric augmentation in our study.

Takiyama et al. [12] used CNNs to identify the anatomical location of esophagogastroduodenoscopy (EGD) images. This study also used four main categories for EGD imaging: the larynx, esophagus, stomach (upper, middle, lower), and duodenum. This study had 16,632 images and was classified with 97% accuracy. Although this study achieved high accuracy, it was explicitly written for EGD images, leaving it slightly generic to score the ESG procedure. This paper was written from the perspective of diagnosis rather than training. In a paper by Cetinsaya et al. [8], colorectal lesions are classified by comparing multiple transfer learning methods. The deep learning models they used are GoogLeNet, AlexNet, and InceptionV3, all deep learning models that use CNNs. They achieved the highest accuracy of their study with 92% by using inceptionV3.

In a study by Li et al. [13], CNN was used to classify lung image patches with interstitial lung disease. This study used what Li et al. considered a relatively small dataset (16,220 image patches) with non-distinct visual structures; therefore,

they designed their CNN to have dropout and single convolutional layers to avoid overfitting. One issue with this study is that they only recorded the recall and precision results but not the model's accuracy. Although the accuracy can be derived from these two measures, the precision and recall alone do not account for the correct classification of negative samples (true negatives); therefore, they should not be the only statistic used to measure the model's validity.

In a paper by Shaju et al. [14], the authors used SMOTE with decision trees to classify diabetes prognosis. The dataset they used contained 734 patient records from a diagnostics laboratory, where the data had 11 attributes (age, plasma glucose fasting, postprandial glucose level taken 2 h after a meal, BMI, systolic blood pressure, diastolic blood pressure, waist thickness, HbA1c value, family diabetic history, diabetic or non-diabetic). This was nine numerical attributes and two nominal attributes. The classifier obtained 92% accuracy before SMOTE while obtaining 94% accuracy after SMOTE. This study used SMOTE on numerical data rather than image data.

In a study by Bellinger et al. [15], they created a new version of SMOTE called manifold-based synthetic oversampling. They created this because they found that SMOTE is used on specific datasets, with a significant difference between group sizes, making SMOTE error-prone. They claim that SMOTE does not reach its full potential on these datasets. Since our datasets do not contain a significant imbalance, we do not need to adopt these practices, as this paper suggests.

Abeysinghe et al. [16] found that SMOTE with eight nearest neighbors resulted in the highest accuracy in combination with principal component analysis (PCA) for image segmentation. This was compared to under-sampling SMOTE with the four nearest neighbors, each with either FCM or PCA. Although this study provides insight into multiple sampling techniques, we did not use image segmentation in our study.

Methods

The Indiana University Human Research Protection Program has determined the project does not require an IRB review due to the project not involving human subjects.

In this section, we describe the methods used for classification using images and videos. These methods start with collecting the data and end with validating our model. Figure 1 shows the workflow of our system to classify images and videos. Our system has two main parts: (a) preprocessing and (b) classification.

In preprocessing, the first process for image classification is capturing images from videos and then resizing them to ensure they are all the same size—128 × 128 pixels. Once the

images are resized, they can be sent through the data augmentation process, and if the data set is imbalanced, the data set is balanced using SMOTE. Once the data has been through preprocessing, the data set can be used for classification. For classification, data goes through model training and model validation steps. The processes for stomach localization and task classification are slightly different. For localization, due to our dataset being imbalanced, the dataset goes through the entire process, including the path from data augmentation through balance data using SMOTE. Our task classification dataset is already balanced, which allows us to bypass the balance data step using the SMOTE process and continue from data augmentation to classification.

For video classification, the first process is to clip the expert video into labeled parts, explained in detail below. The second process is to extract the frames of the videos at a frame rate of five and then save those images. The third process is to resize the images, so they are all the same and run smoothly through the model. We chose to go with an image size of 128 × 128 pixels. Once the images are resized, SMOTE can be applied to balance the dataset. Once the data has been through this preprocessing, the data set can be used for classification.

Data collection

For our previous study [17], we collected seven videos (four experts and three novices) of the ESG procedure performed in ex vivo porcine specimens. We captured 62 screenshots of the four relevant stomach parts for the image classification necessary for ESG grading. Of these 62 screenshots, ten were of the angularis incisura (Fig. 2a), 20 were of the anterior wall (Fig. 2b), 17 were of the greater curvature (Fig. 2c), and 15 were of the posterior wall (Fig. 2d).

For video classification, we captured a single video for each of the four relevant parts of the stomach necessary for ESG grading. Of these four videos, the angularis video was 17 s, the anterior wall video was 13 s, the greater curvature video was 19 s, and the posterior wall video was 13 s. The videos were then converted into frame-by-frame images, resulting in 515 angularis, 390 anterior wall, 587 greater curvature, and 402 posterior wall images.

For the classification of the tasks, we captured 100 screenshots for image classification, which included 50 each of grasping (Fig. 3a) and suturing tasks (Fig. 3b). We captured two videos of grasping and suturing tasks for video classification. Both videos were 28 s in length. This resulted in 857 images for grasping and 853 images for suturing.

Data augmentation

For image classification, we used data augmentation, which includes rotations, flips, and zoom-ins, to create a more
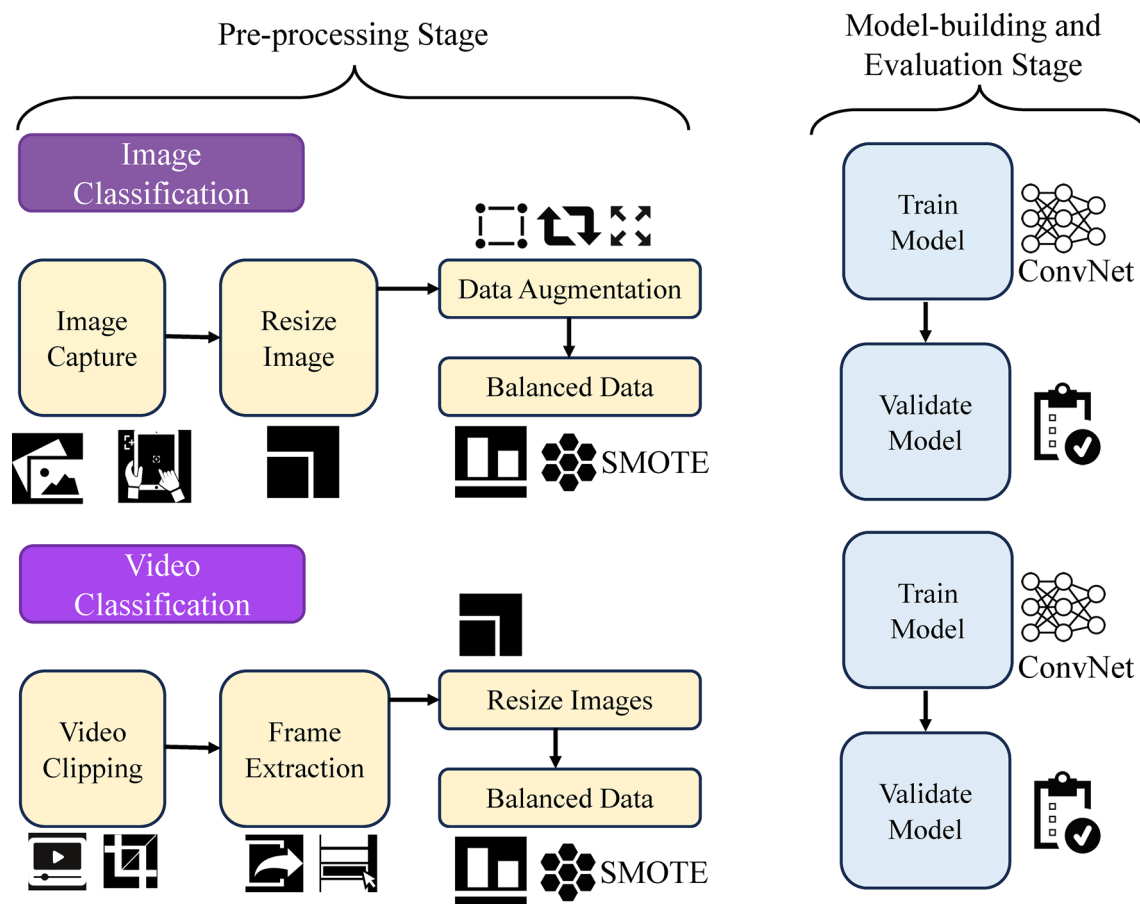
Fig. 1 Overall study design for image and video classification

extensive and well-versed dataset. When performing data augmentation on our 62 images of the stomach parts, we ended up with 914 images. These 914 images are broken down where 157 are of the angularis, 330 are of the anterior wall, 225 are of the greater curvature, and 202 are of the posterior wall. Figure 4 shows an example image after the augmentation of the angularis.

For the grasping and suturing tasks, we performed the same data augmentation (rotations, flips, zoom-in, etc.) on the 100 images of suturing and grasping. This process provided us with 632 images, including an even 316 in each category. Figure 5 shows an example image after the augmentation of the grasping task.

Convolutional neural networks

For stomach localization using images, we used 914 images; for the video data, we used a balanced dataset of 2,348 images. For task classification using images, we used 632 images, and for video classification, a balanced dataset of 1714 images. We ran them through the same sequential convolutional neural network. The model we used had three 2D

convolutional layers, three max-pooling layers, three dropout layers, a flattened layer, and two dense layers. We used a learning rate of 0.001, allowing for a steady increase in accuracy. The CNN architecture used in the study is depicted in Fig. 6.
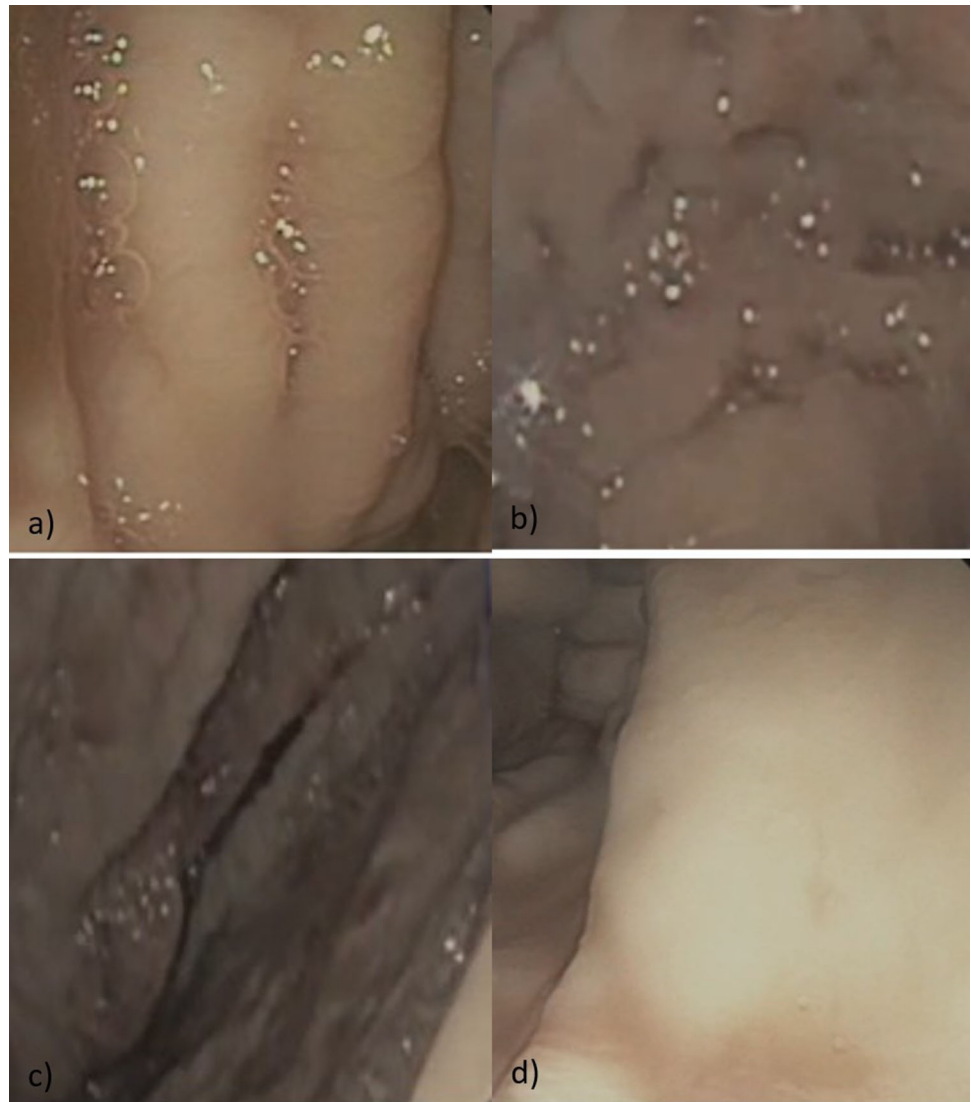
SMOTE

For the image classification, due to our dataset for the localization portion of this study being imbalanced, we decided to use SMOTE. We ran our newly created 914 images through SMOTE, which gave us 1,320 total images, where each part of the stomach contains 330 images. Since the largest original group was the anterior wall with 330 samples, all the other groups were balanced to the same number of samples in this dataset to 330 images. Using SMOTE was unnecessary for this study's task classification since the two groups were balanced.
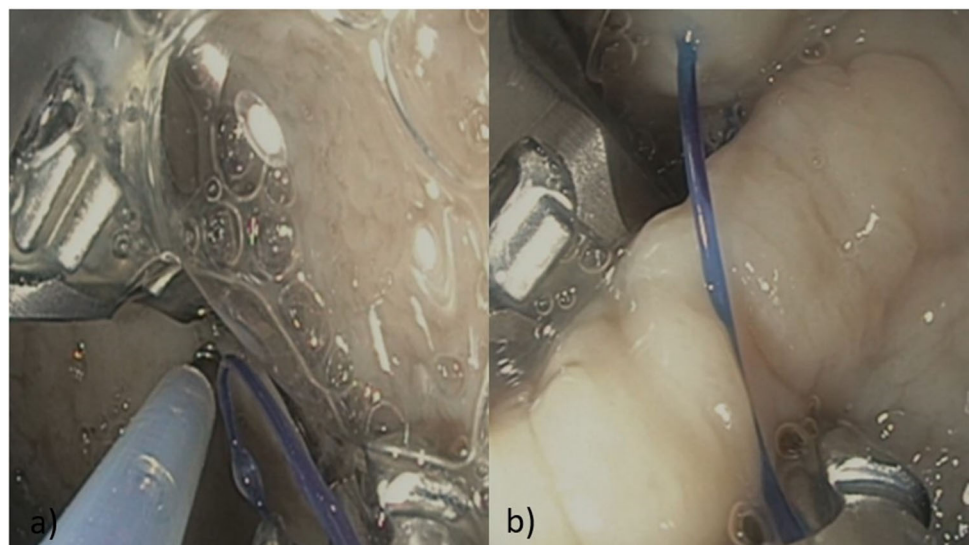
For video classification, we ran our newly created 1894 frames through SMOTE, which gave us 2,348 total images, where each part of the stomach contains 587 images. Since the largest original group was the greater curvature with 587

Fig. 2 **a** Angularis incisura, **b** anterior wall, **c** greater curvature, and **d** posterior wall

Fig. 3 **a** Grasping task, **b** suturing task

Fig. 4 Image after the augmentation of the angularis

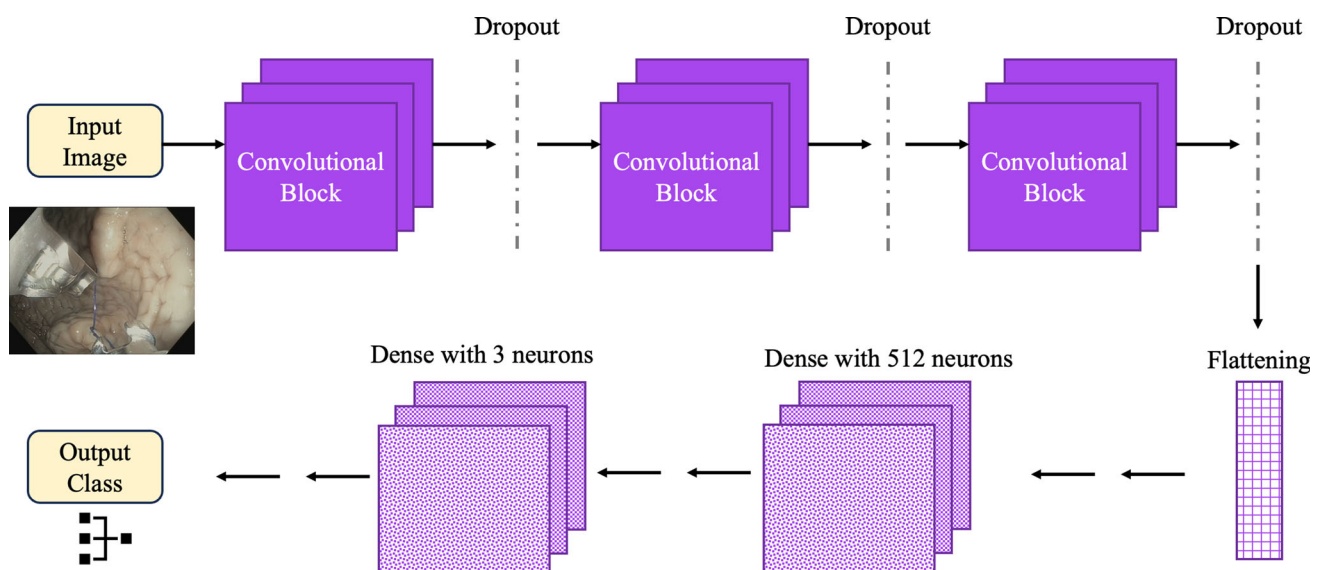**Fig. 5** Image after the augmentation of the grasping task

samples, all the other groups were balanced to the same number of samples in this dataset to 587 images. We also used SMOTE for task classification, which balanced our dataset from 857 grasping and 853 suturing to 857 grasping and 857 suturing images.

Results

CNN localization classification before SMOTE

For image classification, the CNN model for classifying the location of the stomach from the endoscope without SMOTE resulted in an 89% testing accuracy with a validation accuracy of 84%. This model's accuracy and loss before SMOTE can be seen in Fig. 7a, b, respectively. In Fig. 7a, the model's training accuracy stays linear after a significant spike from 0 to 65%. There are minimal spikes with a learning rate of 0.001 and dropout layers. The validation accuracy had a few more spikes, especially around the 35th epoch, but remained linear. We determined that we would use 40 epochs due to the plateau in validation accuracy after about 36 epochs. After determining this from the first model, we kept 40 epochs as a constant.
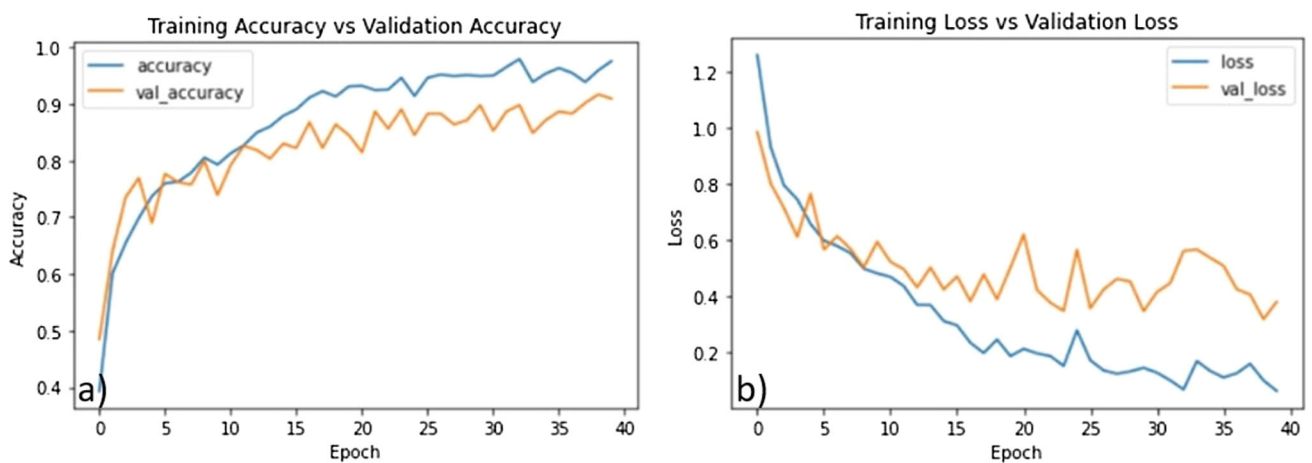
The loss is shown below in Fig. 7b, where the loss rapidly decreased until about 0.7 and gradually decreased. We saw the same spike in Fig. 7a but for loss around epoch 35. The spikes can be flattened through a lower learning rate but would hinder the model's performance, failing to train.
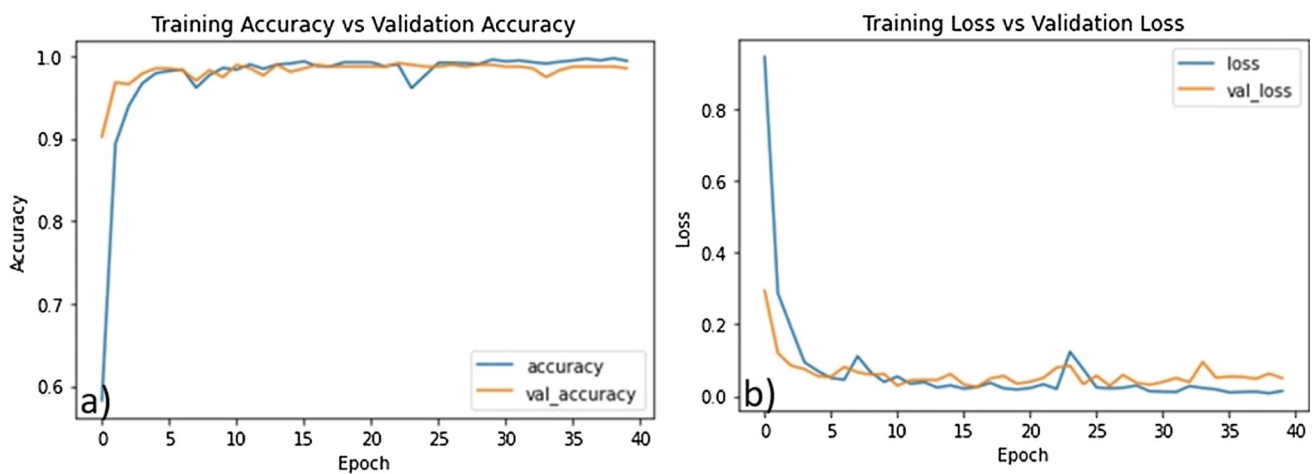

Fig. 6 Convolutional neural network architecture

Fig. 7 Training versus validation **a** accuracy for ESG localization before SMOTE and **b** loss for ESG localization before SMOTE
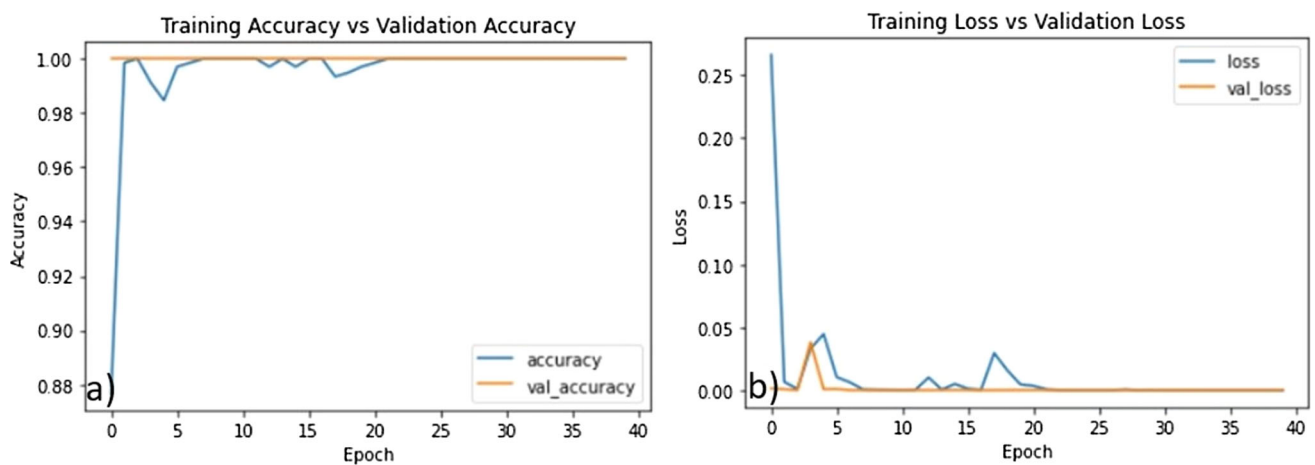

Fig. 8 Training versus validation **a** accuracy for ESG localization after SMOTE and **b** loss for ESG localization after SMOTE

Fig. 9 Training versus validation **a** accuracy for ESG localization and **b** loss for ESG localization

Fig. 10 Image classification—training versus validation **a** accuracy for grasping versus suturing and **b** loss for grasping versus suturing

Fig. 11 Video classification—training versus validation **a** accuracy for grasping versus suturing and **b** loss for grasping versus suturing

CNN localization classification after SMOTE

For image classification, after balancing the dataset with SMOTE and running the data through the same model, the accuracy was 97%, with a validation accuracy of 90%. The accuracy and loss for this model with SMOTE can be seen in Fig. 8a, b, respectively. In Fig. 8a, the training and validation accuracies are much more linear, resulting in fewer spikes than the model without SMOTE. Not only does the accuracy reflect the significance of SMOTE, but having a balanced dataset also results in more occasional spikes during learning.

In Fig. 8b, the loss is shown. There are more spikes in the validation loss compared to the validation accuracy. The issue is due to possible overfitting caused by the number of epochs, which is 35. Although this may be the case, the model corrects itself after epoch 35, resulting in higher accuracy. Since the model continued to increase in accuracy, we neglected this idea rather than cutting it off at 35 epochs.

For video classification, when balancing the dataset with SMOTE and running the data through the CNN model, the training accuracy was 99%, with a validation accuracy of 98%. Figure 9a, b shows the accuracy and loss with SMOTE. As seen in Fig. 9a, the training and validation accuracies have linear shapes, resulting in a few spikes, indicating a robust model. The loss for training and validation is a smooth drop throughout the model. This shows that the model did not overfit, as seen in Fig. 9b.

CNN grasping and suturing task classification

For image classification of grasping and suturing, the model's accuracy was 97%, with a validation accuracy of 89%. This model was already balanced; therefore, there was no need to compare before and after SMOTE. The training accuracy is much more gradual than the validation accuracy, as seen in Fig. 10a. The validation accuracy began to plateau around 15 epochs. In this part of the study, we used 33 epochs due to this validation plateau.

The training loss gradually went to zero, while the validation loss rose, as seen in Fig. 10b. This validation loss shows some overfitting. Around epoch 15, the validation loss began to rise, indicating an overfitting problem. In our initial analysis, an overfitting challenge was encountered. To address this issue, we augmented the dataset size from 632 to 1131 instances. Additionally, we added regularization terms (L1 and L2), introduced dropout layers, and adjusted the learning rate to a lower value of 0.001.

For video classification of grasping and suturing use, the model's accuracy was 100%, with a validation accuracy of 100%. This model was nearly balanced, but we still decided to use SMOTE moving up from 1710 (grasping (857), suturing (853)) images to 1714 (grasping (857), suturing (857)). Figure 11a shows a strong accuracy throughout the whole model, especially during the validation phase.

Figure 11b shows the loss for both training and validation. The model has very few spikes, with only two significant spikes in training loss and one significant spike in validation loss. With a learning rate of 0.001 and 3 dropout layers, this model does not overfit or underfit and is ready for real-world data.

Conclusion

In this study, we collected four different datasets. The first two datasets were of four different stomach parts (incisura angularis, anterior wall, greater curvature, and posterior wall) to determine the endoscope's location in the stomach during the ESG procedure. The third and fourth data sets were collected to determine whether the endoscopist was grasping or suturing. A potential limitation in our study is due to the restricted number of videos utilized for the deep learning algorithm. To address this issue, we implemented data augmentation techniques. We applied two methods, the first being the use of images and the second being the use of video frames. Using video frames with computer vision provided a better dataset, resulting in higher accuracy for both stomach location (8% higher) and task classification (11% higher). SMOTE increased the testing accuracy for image classification by 6%, going from 84 to 90%. All classification algorithms were written as a preliminary study toward building an ESG virtual reality training simulator. This simulator will objectively give feedback to the learner and provide an automated training score that can be tracked over time and archived for review of mentors. In conclusion, we demonstrated that combining data augmentation and SMOTE could be used to enlarge and balance a dataset to classify ESG data with a CNN.

Declarations

Conflict of interest Mr. Dials and Drs. Demirel, Sanchez-Arias, Halic, De, and Gromski declare that they have no competing interests in regard to this study.

Human and animal studies The Indiana University Human Research Protection Program has determined the project does not require an IRB review due to the project not involving human subjects.

Informed consent None.

References

1. Halic T, De S, Dials J, Gromski MA, Demirel D, Ryason A, Gilmore AC, Al-Haddad MA, Kundumadam S (2020) S1191 task analysis and performance metrics of endoscopic sleeve gastroplasty: preparation for virtual simulation development. Off J Am Coll Gastroenterol ACG 115:S595. https://doi.org/10.14309/01.ajg.0000706812.30100.05

2. Dials J, Demirel D, Halic T, De S, Ryason A, Kundumadam S, Al-Haddad M, Gromski MA (2021) Hierarchical task analysis of endoscopic sleeve gastroplasty. Surg Endosc 1–16

3. Polese L, Prevedello L, Belluzzi A, Giugliano E, Albanese A, Foletto M (2022) Endoscopic sleeve gastroplasty: results from a single surgical bariatric centre. Updat Surg 74:1971–1975. https://doi.org/10.1007/s13304-022-01385-4

4. Farha J, McGowan C, Hedjoudje A, Itani MI, Abbarh S, Simsek C, Ichkhanian Y, Vulpis T, James TW, Fayad L, Khashab MA, Oberbach A, Badurdeen D, Kumbhari V (2021) Endoscopic sleeve gastroplasty: suturing the gastric fundus does not confer benefit. Endoscopy 53:727–731. https://doi.org/10.1055/a-1236-9347

5. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. Convolut Neural Netw Vis Recognit 11:1–8

6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953

7. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: 2017 international conference on engineering and technology (ICET), pp 1–6

8. Cetinsaya B, Dials J, Demirel D, Halic T, De S, Gromski M, Rex D (2020) Comparison study of deep learning models for colorectal lesions classification. In: Proceedings of the 2020 the 4th international conference on information system and data mining. Association for Computing Machinery, New York, pp 84–88

9. Hussain Z, Gimenez F, Yi D, Rubin D (2018) Differential data augmentation techniques for medical imaging classification tasks. AMIA Annu Symp Proc 2017:979–984

10. Mikolajczyk A, Grochowski M (2018) Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE, Swinoujście, pp 117–122

11. Taylor L, Nitschke G (2018) Improving deep learning with generic data augmentation. In: 2018 IEEE symposium series on computational intelligence (SSCI). IEEE, pp 1542–1547

12. Takiyama H, Ozawa T, Ishihara S, Fujishiro M, Shichijo S, Nomura S, Miura M, Tada T (2018) Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Sci Rep 8:7497. https://doi.org/10.1038/s41598-018-25842-6

13. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M (2014) Medical image classification with convolutional neural network. In: 2014 13th international conference on control automation robotics vision (ICARCV), pp 844–848

14. Mirza S, Mittal S, Zaman M (2018) Decision support predictive model for prognosis of diabetes using SMOTE and decision tree. Int J Appl Eng Res 13:9277–9282

15. Bellinger C, Drummond C, Japkowicz N (2016) Beyond the boundaries of SMOTE. In: Frasconi P, Landwehr N, Manco G, Vreeken J (eds) Machine learning and knowledge discovery in databases. Springer, Cham, pp 248–263

16. Abeysinghe W, Hung C-C, Bechikh S, Wang X, Rattani A (2018) Clustering algorithms on imbalanced data using the SMOTE technique for image segmentation. In: Proceedings of the 2018 conference on research in adaptive and convergent systems. ACM, Honolulu Hawaii, pp 17–22

17. Dials J, Demirel D, Sanchez-Arias R, Halic T, Kruger U, De S, Gromski MA (2023) Skill-level classification and performance evaluation for endoscopic sleeve gastroplasty. Surg Endosc 1–12