IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 3, June, 2023, pp. 773-782
Journal homepage: http://iieta.org/journals/ria

Deep Learning Architectures for Abnormality Detection in Endoscopy Videos

Madhura Prakash Manjunath*, Krishnamurthy Ningappa Gorappa

Department of CSE, BNM Institute of Technology, Bangalore 560070, India

Corresponding Author Email: madhuraprakash18@gmail.com

https://doi.org/10.18280/ria.370326

Received: 20 April 2023
Accepted: 20 May 2023

Keywords:
endoscopy, classification, architecture pipeline, computer vision, deep architecture, medical procedure

ABSTRACT

Endoscopy is a widely employed technique for the diagnosis and treatment of various internal organs in the human body, including the gastrointestinal tract, lungs, bones, and abdominal region. During the procedure, an illuminated optical device records video data, which assists physicians during real-time analysis and post-procedure evaluations. Identifying areas of interest within the vast amount of recorded video data is critical for optimizing physicians' focus and time. A key task in this process involves classifying endoscopic frames as normal or abnormal. Current solutions for endoscopic frame classification either rely solely on handcrafted features or neural network features and lack efficient pre-processing techniques to eliminate irrelevant frame portions or enhance relevant region features. This study presents an innovative architecture pipeline for the efficient and robust detection of abnormal frames in endoscopic videos, combining effective pre-processing techniques with deep neural networks. A novel and customized pre-processing method has been integrated into three custom-tailored deep architectural pipelines, which are based on sequential convolutional networks, InceptionResNet, and EfficientNet. Models generated using these pipelines were trained and tested on custom-curated data from publicly available repositories. Among the three pipelines, the architecture based on EfficientNet outperformed current state-of-the-art approaches, achieving a sensitivity, specificity, and accuracy of 0.94, 0.91, and 0.93, respectively, for the classification of abnormal frames. This novel approach demonstrates the potential of leveraging advanced deep learning architectures to enhance abnormality detection in endoscopic videos.

1. INTRODUCTION

Convolutional Neural Networks (CNNs) have emerged as the state-of-the-art technique for extracting off-the-shelf features from images. These features encompass a wide range, from high-level attributes representing abstract and complex aspects, such as objects, scenes, and concepts that can be recognized and interpreted by humans, to nuances of the image, including color, texture, shape, and edge information. The comprehensive range of feature extraction facilitated by CNNs offers a distinct advantage compared to hand-crafted features obtained through traditional image processing techniques. Features extracted from CNNs can be employed in classification tasks to categorize images as belonging to specific classes, as well as in segmentation tasks to isolate and highlight relevant portions of an image. Over the past decade, the ImageNet challenge [1] has led to the development of numerous standard CNN architectures that have demonstrated remarkable results. These architectures comprise sequential blocks of convolution, pooling, and fully connected layers that can be combined in various sequences, functioning as plug-and-play units of a puzzle that can be connected in diverse configurations to generate customized architectures.

Training CNNs for classification and segmentation tasks necessitates a substantial amount of labeled data. The result of this training process is a trained neural network model with specific weights associated with each edge in the network. These neural network weights represent the learned feature matrix, which facilitates classification and segmentation tasks on test or unknown data sets. CNN architectures are employed to generate trained models with learned weights, either from scratch or by using pre-trained models for transfer learning. Incorporating transfer learning is advantageous as it promotes rapid convergence by offering a relevant starting point for training the model.

Abnormality detection in endoscopy video frames serves as an initial step in assisting medical professionals with analysis. Existing solutions for abnormality classification rely on either traditional image processing techniques or deep learning-based methods. However, these approaches often lack a focus on customizing the classification process for the endoscopy domain. Moreover, an ensemble of computer vision-based techniques and deep architectures could enhance the efficiency of abnormality classification. To address this, three deep architecture pipelines have been implemented and tested on curated endoscopy image data for the purpose of abnormality detection as a classification problem.

Data from several publicly available resources have been collected for this work. The collected data has been curated to include the categories of normal and abnormal endoscopy frames. The three pipelines implemented are CNN sequential architecture, Inception-Resnet architecture [2], and Efficientnet B0 architecture [3]. These three pipelines have been customized and experimented with several computer-vision-based pre-processing techniques like multi-channel mixing, histogram equalization, morphological operation, and

blurring in order to enhance the abnormality regions in the frame, like polyps, bleeding, tumor, or other lesions. The details of each of these experiments are presented in the further sections. The significance and the limitations of each of these techniques are also discussed in detail.

Endoscopy is utilized in both diagnosis and therapy in the various inner organs of the human body ranging from the gastrointestinal tract, lungs, bones, abdominal region, and further. The video recorded with the help of the illuminated optical device in the procedures aids the physicians both in real-time and post-procedure analysis. This video is revisited by the physicians to explain to the patients, to train young physicians and for detailed analysis. The duration of the endoscopy directly depends on the duration of the procedure which in turn depends on the specific endoscopy process. The recorded video data can be enormous. Pointing out the relevant areas in this video data is an important task to save the physicians precious focus time. A crucial task in this process is to be able to classify the endoscopy frames as normal or abnormal.

As compared to the classification process on regular images, the classification of endoscopy video frames presents several challenges because of the poor quality of the frames, the various artifacts that can be present as a part of the content of the frame like reflection, blood, smoke, motion blur or occlusions between tools used in therapy. In addition to these, the endoscopy frames have a black border which can have some textual information present. The textual content and the black border are irrelevant portions of the frame and training a neural network model based on these images requires a vigilant pre-processing technique that would aid the model training process to pick up relevant features of the frame and ignore all other points. With this aim, an algorithm for masking and cropping the irrelevant portion of the endoscopy frame has also been proposed in this work.

Among the three pipelines experimented, the architecture pipeline based on EfficientNet has shown the highest performance compared to the current state-of-the-art approaches. A sensitivity, specificity, and accuracy of 0.94, 0.91, and 0.93 respectively have been achieved with the proposed novel approach to categorize the abnormal frames. The insights from the results based on the experiments conducted as a part of this work are multi-fold, (a) the requirement of customized, efficient, and effective pre-processing techniques and algorithms that can be applied to the images before feeding them to the network for training purposes (b) the need for designing a customized architecture and a pipeline to provide a solution for the task of classification (c) detailed analysis of results obtained from the three pipelines. The next section discusses the related work and gives a list of the publicly available dataset for endoscopy video analysis. The implementation details of the masking and cropping algorithm and the three solution pipelines for abnormality classification in the endoscopy frame are explained in section 3. The results obtained and the analysis is presented in section 4.

2. RELATED WORK

The presence of abnormalities in the endoscopy video frame is a piece of semantic information that aids in the selection of meaningful and relevant frames from the video. This section presents the study of existing works in the literature that have focused on the detection of abnormalities from endoscopy videos or the extraction of any other context-related information that can serve as a piece of semantically meaningful information to perform the classification of the frame.

The existing solutions reviewed can be categorized into (i) solutions focused on using hand-crafted feature extraction and classification [4-12] and (ii) solutions based on deep neural network features [13-24].

2.1 Based on hand-crafted features

Chen and Lee [4] have reviewed the current development of machine-vision-based analysis of endoscopy video, focusing on the research that identifies specific gastrointestinal (GI) pathology and methods of shot boundary detection based on traditional image processing techniques. Alexandre et al. [5] have examined the handcrafted color and position features versus texture features for polyp detection in endoscopic frames using the Support Vector Machine (SVM) classifier. Ghosh et al. [6] have proposed a block-based histogram feature extraction method based on traditional hand-crafted features for bleeding detection in Wireless Capsule Endoscopy (WCE) videos. To obtain local statistical features, the maximum pixel value of each spatial block was computed, and the global feature of an image was obtained considering the histogram bin frequency of block maxima. Sekuboyina et al. [7] transformed the color space of an image and classified a pixel in the frame as one belonging to an abnormality (malign pixel) or not (benign pixel) using the traditional image processing-based feature extraction technique.

Ghosh et al. [8] considered a block surrounding individual pixels for extracting local statistical features. By combining local block features of three different color planes of RGB color space, an index value was defined. A color histogram, extracted from those index values, provided a distinguishable color texture feature. A feature reduction technique utilizing color histogram patterns and principal component analysis was proposed for bleeding detection in WCE videos. Alexandre et al. [9] proposed a polyp detection technique based on the color and position information in the frames.

Bipin Dev [10] implemented a canny edge detector to detect the edge regions in the L channel of the image for identifying bleeding in WCE videos. Savazzi and Guarnaschelli [11] used an image patch-based feature extraction and classification. This method extracted features based on Local Binary Pattern (LBP), Gray Level Cooccurrence Matrix (GLCM), and CNN. Vasilakakis et al. [12] provided a study on the commercially available WCE platforms, as well as the advances made in optimizing the diagnostic capabilities of WCE. The study provides an overview of the traditional image processing techniques applied to aid in the diagnosis of WCE video frames.

2.2 Based on deep neural networks

Byrne et al. [13] have proposed a CNN classifier model based on exclusive off-the-shelf features for polyp differentiation on endoscopy frames. Yuan and Meng [14] proposed a deep feature learning method to recognize polyps in the WCE images. The proposed method used an image manifold constraint, which was constructed by the nearest neighbor graph that represented the intrinsic structures of images. The image manifold constraint enforced that images

within the same category share similar learned features and images in different categories should be kept far away. Thus, the learned features preserve large inter-variances and small intra-variances among images.

Bernal et al. [15] provided a comparative analysis of polyp detection methods in colonoscopy videos performed under various challenges like overlay information, specular highlights, and overexposed regions. Bouget et al. [16] provided an analysis of validation techniques employed to obtain detection performance results and establish comparisons between surgical tool detectors in endoscopy therapeutic videos. Law et al. [17] developed an instrument tracker based on the Hourglass neural networks and assessed the movement of the robotic instruments, and classified the technical level of surgeons with a linear classifier, using peer evaluations of skill as the reference standard.

Coelho et al. [18] proposed an evaluation of deep learning U-Net architecture, to detect and segment red lesions in the small bowel. Gong et al. [19] proposed a machine translation framework for automatic pathology annotation on medical images. Iakovidis et al. [20] proposed a solution based on three phases to locate gastrointestinal (GI) abnormalities. First, the method classified the video frames into abnormal or normal using Weakly Supervised Convolutional Neural Network architecture. Then salient points from deeper WCNN layers were detected, using a Deep Saliency Detection algorithm; and GI anomalies using an Iterative Cluster Unification algorithm.

Park and Lee [21] proposed a class-labeling method that can be used to design a learning model by constructing a knowledge base focused on main lesions defined in standard terminologies for capsule endoscopy. Cao et al. [22] obtained feature maps of the same resolution by performing a max pooling operation on different convolutional layers, and then quantifying the pooled feature maps for WCE frame classification. Yang et al. [23] merged multi-level features by explicitly modeling interdependencies between all feature maps of different convolution layers for lesion classification in WCE frames. Vasilakakis et al. [24] proposed an unsupervised color-based saliency detection scheme that combined both point and region-level saliency information and the estimation of local and global image color descriptors.

The literature survey conducted gave insight into the existing solutions for abnormality classification and detection in the endoscopy frames. Many of these solutions focus on detecting and localizing abnormalities like bleeding, lesions, and polyps. These solutions either exclusively extracted hand-crafted features followed by classification or used deep learning-based classification and segmentation. This work aimed at developing an ensemble solution based on hand-crafted feature extraction in the pre-processing stage followed by extracting the off-the-shelf deep features. This work presents the results of the experiments with the three ensemble pipelines and discusses their pros and cons. The resulting model based on the novel, innovative pipeline consisting of (i) border mask removal, (ii) feature enhancement for highlighting regions of abnormality in endoscopy frames, and (iii) training of curated data on customized EfficientNetB0 architecture, has been successful in classifying the endoscopy frames with an accuracy of 93%.

The implementation of any deep learning-based solution would require the presence of a large amount of data. In this work, an enormous labeled endoscopy frame data was required. The research work on the endoscopy video data starting from the classification of the frames to the segmentation of abnormality regions and further is feasible because of the widely available dataset that can be downloaded for research purposes. Some of these publicly available datasets for endoscopy video data analysis are listed below in Table 1.

Table 1. List of publicly available labelled datasets for endoscopy video frame

Sl. No.	Dataset	Details
1	WCE colon disease curated disease dataset	KVASIR Dataset images curated. 6000 images belonging to normal, ulcerative colitis, polyps, and esophagitis [25]
2	KVASIR	Images, annotated and verified by endoscopists, including 8 classes showing anatomical landmarks, pathological findings, or endoscopic procedures in the GI tract, i.e., hundreds of images for each class [26]
3	m2cai16-workflow Dataset	41 labelled cholecystectomy videos [27]
4	m2cai16-tool Dataset	15 labelled cholecystectomy videos [28]
5	Surgical160	160 short video clips showing typical surgical actions in gynaecologic laparoscopy. The dataset consists of 16 distinct classes [29]
6	CholecSeg8k	Labelled data for semantic segmentation [30]
7	HYPERKVASIR	Multi-class image and video dataset for gastrointestinal endoscopy, 110,079 images and 374 videos from various GI examinations [31]
8	KVASIR SEG	1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2 [32]
9	Cholec80 Dataset	80 labelled videos of cholecystectomy procedures [33]
10	EAD Dataset	Localization of bounding boxes and class labels for 8 artefact classes for given frames [34]
11	Kvasir-capsule	PillCAM dataset containing 47,238 labelled images and 117 videos [35]

3. IMPLEMENTATION

The implementation process for the classification of the endoscopy frames as normal or abnormal proceeded in several phases starting with collecting labeled images from various publicly available resources. The curated dataset comprised endoscopy frames grouped into two categories namely normal and abnormal. The abnormal images consisted of all the frames that depicted an abnormality ranging from bleeding, lesions, polyps, ulcer, esophagitis, and tumors. A total of 2000 images from both classes were initially considered. These images were augmented with a rotation range of 40, a shear range of 0.2, and a zoom range of 0.2 keeping the brightness range between 0.5 to 1.5 and the horizontal flipping to true. The image augmentation was performed to increase the generalizability of the trained model and to increase the sample set. The augmented images were further analyzed and any poor-quality frames with blur and high interpolation were discarded and a total of 8000 images in both classes were considered. 300 frames from the unaugmented set were reserved for testing the generated models.

The endoscopy curated dataset now consisted of images from several sources with varying content and resolution. A

sample set of images from both classes are shown in Figure 1(a) and Figure 1(b).

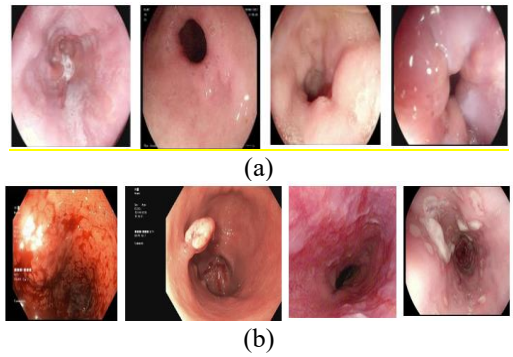


Figure 1. Sample frames from abnormal category

It can be observed from the frames in both normal and abnormal categories that the endoscopy frames might have oval or polygonal center information with a black border. In some of the endoscopy frames the black border might as well contain textual information pertaining to the procedure being performed. The black border is an irrelevant portion of the image. During training the model might try to look for features in this irrelevant zone. Hence it becomes important to design an algorithm to crop the irrelevant portion of the image. The masking and cropping algorithm designed for this purpose is explained in the next section.

3.1 Experimental setup

The deep-learning-based ensemble architectures was instantiated and the model generated in all the experimental pipelines was trained on an Azure NC-series NC6s_v3 Virtual Machine (VM) with an A100 Graphics processing unit (GPU). The training was performed with a batch size of 32. Adam and Stochastic gradient descent (SGD) optimization algorithms were considered for model training experiments. The adaptive learning rates were started with 0.01. Early stopping was incorporated in order to prevent overfitting and save computational resources. The patience level was set to 10 and the validation loss and accuracy were monitored during training.

3.2 Masking and cropping algorithm

Table 2. Phases in masking and cropping algorithm

Step 1: Load the original frame and the mask frame
Step 2: Set the tolerance value to 7, and convert the original frame to greyscale
Step 3: Apply a threshold and check the pixel values in the frame to the tolerance value set
Step 4: Consider the pixels that are greater than the tolerance value as they are not too dark as in the case with border pixel
Step 5: Pick the indices of the pixel that are not too dark based on the threshold on the coloured frame in all three channels
Step 6: Construct the cropped image by stacking indices picked in all three channels

The masking and cropping algorithm were designed to crop the black border region of the endoscopy frame so that the cropped frame would only have the targeted portion of the abnormality to a large extent. A mask frame was constructed to aid the cropping process. The algorithm steps are explained in Table 2. A few of the sample sequences of endoscopy frames after performing the masking and cropping are depicted in Figure 2.

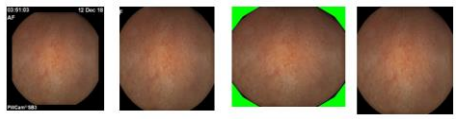


Figure 2. Sample frame in masking and cropping sequence

3.3 Architecture pipelines

Three pipeline architectures with an ensemble of hand-crafted features based on computer vision techniques and off-the-shelf features from deep learning techniques are experimented with. The input to these architectures was the cropped frames. The pre-processing algorithms used and the customized deep architectures in the three pipelines are explained further below.

3.3.1 Architectural pipeline-1

The first approach was based on the sequential convolutional neural network. The sequence of phases in the first solution pipeline is depicted in Figure 3.

The architecture of the CNN consisted of a sequence of convolution and max pooling layers followed by flattened and dense layers. The curated and augmented dataset comprising 8k frames were initially used in the training based on this architecture both with and without applying the masking and cropping algorithm. The frames were resized to 254*254. The architecture had a total of 831,713 parameters and the summary of this sequential structure is shown in Figure 4.

The model when trained on the data after resizing, initially achieved a training accuracy of 85% but the validation accuracy was just 34%. In order to improve the validation accuracy, the trials of performing the first pre-processing technique namely channel mixing were conducted. The channel mixing helped in enhancing the details of the frame like bleeding, mucosa, fluids, etc. The enhanced frame helped the model to pick up better features. The details are mentioned in Table 3. The model was trained on several epochs ranging from 150 to 300 and the improvements in the model training accuracy with different weighted channel is shown in Figure 5.

Figure 3. Architectural pipeline 1 based on CNN sequential structure

The sensitivity, specificity, and accuracy were computed on the test set and the values ranged between 75% to 80% for different trials of the pre-processing. It can be observed from Table 3. that the pre-processing technique that included a channel mixing of red at 80% and green at 20% achieved the highest training accuracy of 94%. However, the model's performance on the test set was not very reliable as the specificity was close to 75%. In order to achieve better performance on the test set, a model based on Inception Resnet V2 was considered for experimentation next. This is explained in the next section.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 254, 254, 32)	896
max_pooling2d_1 (MaxPooling2	(None, 127, 127, 32)	0
conv2d_2 (Conv2D)	(None, 125, 125, 32)	9248
max_pooling2d_2 (MaxPooling2	(None, 62, 62, 32)	0
conv2d_3 (Conv2D)	(None, 60, 60, 32)	9248
max_pooling2d_3 (MaxPooling2	(None, 30, 30, 32)	0
conv2d_4 (Conv2D)	(None, 28, 28, 32)	9248
max_pooling2d_4 (MaxPooling2	(None, 14, 14, 32)	0
flatten_1 (Flatten)	(None, 6272)	0
dense_1 (Dense)	(None, 128)	802944
dense_2 (Dense)	(None, 1)	129

Total params: 831,713
Trainable params: 831,713
Non-trainable params: 0

Figure 4. Sequential model summary

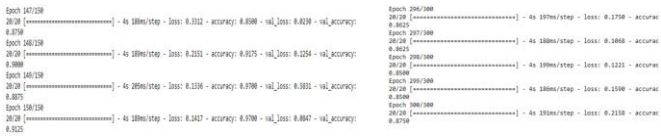


Figure 5. Sequential model training sequence

Table 3. Trails on model generation using sequential architecture

Sl. No.	Pre-processing Applied	Results
1	Weighted Channel: Red 0.1, Green 0.9	Training Accuracy–90%, Validation loss–0.63, Validation accuracy–87%
2	Weighted Channel: Red 0.5, Green 0.5	Training Accuracy–91%, Validation loss–0.152, Validation accuracy–90%
3	Weighted Channel: Red 0.8, Green 0.2	Training Accuracy–94%, Validation loss–0.0283, Validation accuracy–88.75%

3.3.2 Architectural pipeline 2

The Inception-ResNet v2 architecture is a deep neural network that combines the benefits of two popular architectures, Inception and ResNet. Inception-ResNet v2 has achieved state-of-the-art results on various benchmark datasets such as ImageNet. The architecture is designed to reduce computation costs while maintaining high accuracy by using techniques such as factorization and aggressive regularization. Also, the architecture is designed to be robust to variations in input data such as occlusions, translations, and rotations. The architecture was appended with additional layers namely convolution, flatten, dropout, and dense. The transfer learning strategy was applied for model training.

The dropout layer was added as it is a powerful regularization technique that can help to improve the generalization, robustness, and computational efficiency of deep learning neural networks. The input data frames were resized to 299*299 in the color channel and the model had a total of 58,728,813 parameters. The model summary and the architecture details are given in Figure 6.

The model was genrated considering the resized images from the colour channels and the weighted channels. As with the sequential model, the training accuracy that included data with a channel mixing of red with 80% and green with 20% achieved the highest training accuracy of 95%. However again the model's performance on the test set was not very reliable as the sensitivity was close to 79%. The sensitivity, specificity and accuracy computed on the test set resulted in values ranging 0.79, 0.82 and 0.84 respectively. In order to achieve better test performance, a model based on EfficientNet B0 with a custom tailored pre-processing technique was considered. This is explained in the next section.

Layer (type)	Output Shape	Param #
input_4 (InputLayer)	(None, 299, 299, 3)	0
batch_normalization_408 (Bat	(None, 299, 299, 3)	12
inception_resnet_v2 (Model)	(None, 8, 8, 1536)	54336736
conv2d_408 (Conv2D)	(None, 8, 8, 128)	196736
flatten_2 (Flatten)	(None, 8192)	0
dropout_3 (Dropout)	(None, 8192)	0
dense_3 (Dense)	(None, 512)	4194816
dropout_4 (Dropout)	(None, 512)	0
dense_4 (Dense)	(None, 1)	513

Total params: 58,728,813
Trainable params: 58,668,263
Non-trainable params: 60,550

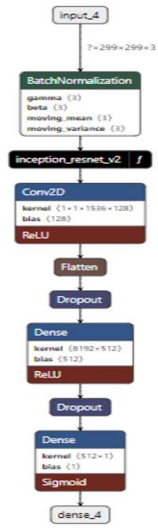


Figure 6. Model summary and architectural details for inception resnet v2 pipeline

3.3.3 Architectural pipeline-3

In the next sequence of trials to generate an efficient solution for the classification of abnormality in endoscopy frames, the EfficientNet B0 architecture was considered. EfficientNet B0 being a deep convolutional neural network architecture has been developed with the goal of achieving better accuracy and efficiency compared to previous models. It has several advantages including better accuracy on several benchmark datasets, such as ImageNet, with fewer parameters than other models. It can achieve high accuracy with fewer parameters, which translates to faster training times and lower memory requirements. Further, EfficientNet B0 has been shown to be highly effective at transfer learning, meaning it can be used as a base model with only minor modifications increasing the generalizability of the model.

The frames in the training set were initially masked and cropped to remove the irrelevant border and were resized to 224*224 in the color channel and the trained model did not show any significant improvement in the test accuracy. A customized pre-processing algorithm was defined to enhance the frames for aiding the model to pick better distinguishing features. The phases in this pipeline are depicted in Figure 7.

The model summary and architectural details are represented in Figure 8. Loss and accuracy plots captured during model training are depicted in Figure 9.

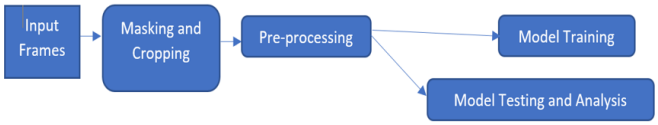


Figure 7. Architectural pipeline 3 based on efficientnet b0 structure

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 224, 224, 3)]	0
batch_normalization (BatchNormalization)	(None, 224, 224, 3)	12
efficientnetb0 (Functional)	(None, 7, 7, 1280)	4049571
conv2d (Conv2D)	(None, 7, 7, 128)	163968
flatten (Flatten)	(None, 6272)	0
dropout (Dropout)	(None, 6272)	0
dense (Dense)	(None, 512)	3211776
dropout_1 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 1)	513

Total params: 7,425,840
Trainable params: 7,383,811
Non-trainable params: 42,029

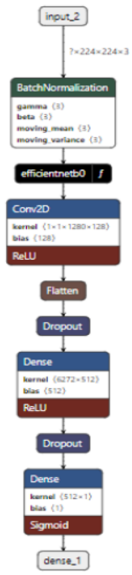


Figure 8. Model summary and architectural details for EfficientNet B0 pipeline

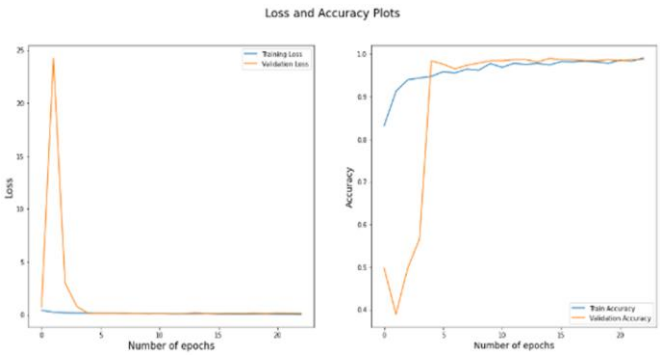


Figure 9. Loss and accuracy plots during model training

Table 4. Phases in pre-processing algorithm

Step 1: Apply Gaussian Blur on the greyscale version of the image with a kernel size of 3*3
Step 2: Find the contours in the given frame and determine the largest contours
Step 3: Crop and scale the image based on the centre and radius calculation
Step 4: Add padding to the image and get the border shape
Step 5: Define the border using border constant
Step 6: Process the constructed image using addWeighted function with a kernel size of 12*12 in the blur function

A customized pre-processing algorithm was designed and implemented for the frame pre-processing for model generation and testing. The pre-processing algorithm involves a six-step process, which is depicted in Table 4. This pre-processing technique enhances the features of the frame highlighting the regions of bleeding, polyps, or any other abnormality in the endoscopy frame. This enhancement helps the model to learn more relevant features and thus improves the accuracy. The sensitivity, specificity, and accuracy of the trained model on the test set after implementing this pre-processing algorithm rose to 0.94, 0.91 and 0.93 respectively. The frame representation after pre-processing is represented in Figure 10.

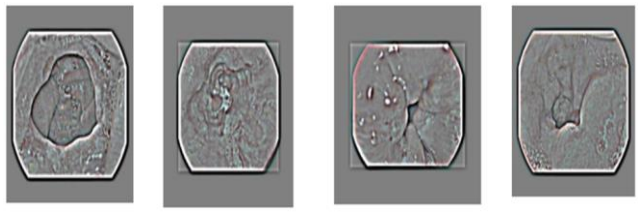


Figure 10. Sample pre-processed frames

3.3.4 Discussion on the pipelines

The masking and cropping procedure for clipping the irrelevant black border of the endoscopy frame was the preceding step for all the experimented pipelines. The first pipeline was based on the channel mixing pre-processing technique combined with the sequential CNN architecture. The weighted combination of the channels helped the model pick up relevant features. However, the trained model's performance on the test set was not reliable.

The second pipleine was based on the InceptionResNet V2 archtitecture. The weighted channel pre-processing technique was experimented with this architecture, and the model test results were not reliable. This seeded the process of a customized and innovative pre-processing technique to enhance the regions of abnormalities like lesions, tumors, and ulcers in the frames. This preprocessing technique used in the third pipeline based on EfficientNetB0 architecture helped in generating a model that performed better than the previous models. This ensemble technique with a customized pipeline has achieved an accuracy of 93% which is higher than the existing solutions as studied extensively in the literature.

4. RESULTS AND DISCUSSION

In this work, three ensemble architectures based on computer vision and deep-learning-based approaches are designed and implemented. These architectures were used to generate trained models for abnormality classification in endoscopy video frames. A masking and cropping algorithm has been proposed in this work to remove the irrelevant black border of the endoscopy frame. Initial experiments were conducted to include the border cropped and resized frame in the color mode as the starting point of model training. Further, to enhance the frame content details, trials of channel mixing were conducted. Upon observing the trained model behavior on the test set, a decision to design a customized pre-processing technique was taken. The customized pre-processing technique enhanced the presence of polyps, tumors, and lesions in the abnormal frames. These features highlighting the areas of abnormality helped in generating an efficient model.

The architectural pipeline based on the EfficientNet B0 architecture achieved the best performance and the sensitivity, specificity, and accuracy of the model generated on the test set were 0.94, 0.91, and 0.93 respectively. The results establish the requirement for a customized and efficient computer

vision-based pre-processing technique that would enable the model to focus on the features from the relevant regions of the frame. Such enhancement techniques applied in the preprocessing stage would greatly contribute to the efficiency of the model generated from the deep architecture. Further, this work proposes a novel and customized, and optimized architecture pipeline for the efficient classification of abnormality in endoscopy frame.

A Grad-CAM (Gradient-weighted Class Activation Mapping) [36] based heatmap was generated on the test set to visualize the distinguishing areas in the frames for the model for abnormality classification. Grad-CAM is a technique for generating a heatmap that highlights the regions of an input image that are most important for a neural network's prediction. The Grad-CAM heatmap is generated by computing the gradient of the output class score with respect to the feature maps of the last convolutional layer of the network. This gradient is then used to weigh the feature maps, and the weighted feature maps are averaged to produce the final heatmap. The steps involved in generating a Grad-CAM heatmap are given in Table 5.

Table 5. Phases in Grad-CAM heatmap generation

Step 1: The neural network is trained to classify images into normal and abnormal classes.
Step 2: During inference, an input image is fed into the network, and the output class score is computed.
Step 3: The last convolutional layer of the network namely conv2d_17 is identified, and the gradient of the output class score with respect to the feature maps of this layer is computed.
Step 4: The gradient is then used to weigh the feature maps of the last convolutional layer, and the weighted feature maps are averaged to produce the final heatmap.
Step 5: The heatmap is then overlaid on top of the original input image, with the heatmap regions corresponding to the areas of the image that were most important for the network's prediction.

The Grad-CAM heatmap provides a useful visual explanation of how a neural network arrived at its prediction. The heatmap generated is represented in Figure 11.

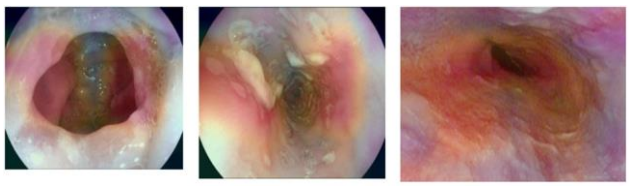


Figure 11. Sample Grad-CAM heatmap on test frames

The statistical parameters considered for evaluating the model performance on the test set in this study are sensitivity, specificity and accuracy. Sensitivity refers to the proportion of true positive cases that are correctly identified by a model or a test. Sensitivity is also called the true positive rate or recall. A high sensitivity indicates that the model is good at identifying abnormality frames, while a low sensitivity indicates that the model is missing abnormality frames. The sensitivity is defined as the number of true positive cases divided by the total number of positive cases. Sensitivity is represented as

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

where, True Positive is the number of abnormality frames that are correctly identified and False Negative is the number of abnormality frames that are incorrectly identified as negative.

Specificity refers to the proportion of normal frames that are correctly identified by the model. A high specificity indicates that the model is good at identifying normal frames, while a low specificity indicates that the model is incorrectly classifying normal frames as abnormal. Specificity is also called the true negative rate. The specificity is defined as the number of true negative cases divided by the total number of negative cases. Specificity is represented as

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \quad (2)$$

where, True Negative is the number of normal frames that are correctly identified and False Positive is the number of normal frames that are incorrectly identified as abnormal frames.

Accuracy is a metric that measures the overall correctness of a model or a test. It is defined as the proportion of correctly classified cases to the total number of cases. Accuracy is represented as

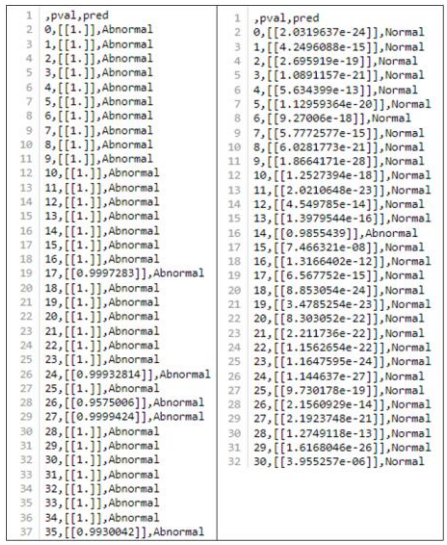
$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (3)$$

where, True Positive is the number of abnormal frames that are correctly identified, False Positive is the number of normal frames that are incorrectly identified as abnormal frames, True Negative is the number of abnormal frames that are correctly identified, and False Negative is the number of abnormal frames that are incorrectly identified as normal frames.

Accuracy is a metric for evaluating the performance of binary classification models. It measures the proportion of correct predictions made by the model on the test set. Accuracy is defined as

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ classifications} \quad (4)$$

The prediction values of a few of the test frames under both categories are depicted in Figure 12.



```
,pval,pred
0,[[1.]],Abnormal
1,[[1.]],Abnormal
2,[[1.]],Abnormal
3,[[1.]],Abnormal
4,[[1.]],Abnormal
5,[[1.]],Abnormal
6,[[1.]],Abnormal
7,[[1.]],Abnormal
8,[[1.]],Abnormal
9,[[1.]],Abnormal
10,[[1.]],Abnormal
11,[[1.]],Abnormal
12,[[1.]],Abnormal
13,[[1.]],Abnormal
14,[[1.]],Abnormal
15,[[1.]],Abnormal
16,[[1.]],Abnormal
17,[[0.9997283]],Abnormal
18,[[1.]],Abnormal
19,[[1.]],Abnormal
20,[[1.]],Abnormal
21,[[1.]],Abnormal
22,[[1.]],Abnormal
23,[[1.]],Abnormal
24,[[0.99932814]],Abnormal
25,[[1.]],Abnormal
26,[[0.9575006]],Abnormal
27,[[0.9999424]],Abnormal
28,[[1.]],Abnormal
29,[[1.]],Abnormal
30,[[1.]],Abnormal
31,[[1.]],Abnormal
32,[[1.]],Abnormal
33,[[1.]],Abnormal
34,[[1.]],Abnormal
35,[[0.9930042]],Abnormal
```

```
,pval,pred
0,[[2.0319637e-24]],Normal
1,[[4.2496088e-15]],Normal
2,[[2.695919e-19]],Normal
3,[[1.0891157e-21]],Normal
4,[[5.634399e-13]],Normal
5,[[1.12959364e-20]],Normal
6,[[9.27006e-18]],Normal
7,[[5.7772577e-15]],Normal
8,[[6.0281773e-21]],Normal
9,[[1.8664171e-28]],Normal
10,[[1.2527394e-18]],Normal
11,[[2.0210648e-23]],Normal
12,[[4.549785e-14]],Normal
13,[[1.3979544e-16]],Normal
14,[[0.9855439]],Abnormal
15,[[7.466321e-08]],Normal
16,[[1.3166402e-12]],Normal
17,[[6.567752e-15]],Normal
18,[[8.853054e-24]],Normal
19,[[3.4785254e-23]],Normal
20,[[8.303052e-22]],Normal
21,[[2.211736e-22]],Normal
22,[[1.1562654e-22]],Normal
23,[[1.1647595e-24]],Normal
24,[[1.144637e-27]],Normal
25,[[9.730178e-19]],Normal
26,[[2.1560929e-14]],Normal
27,[[2.1923748e-21]],Normal
28,[[1.2749118e-13]],Normal
29,[[1.6168046e-26]],Normal
30,[[3.955257e-06]],Normal
```

Figure 12. Prediction values on a few test-frames

5. CONCLUSION

Abnormality classification in the endoscopy frame is one of the first steps in assisting physicians in the diagnosis and analysis of the video content. It would save an enormous amount of the physician's precious time if the irrelevant contents are filtered out. This work focused on abnormality classification in endoscopy frames. For this, the trials of model generation based on three possible ensemble architecture pipelines have been conducted. The ensemble architectures are based on the combination of computer vision techniques and deep learning techniques. As compared to the existing solutions, this work proposes a unique pre-processing technique that aids to enhance the regions of abnormality in the endoscopy frame. This pre-processing technique helped in representing frames with more discriminative features, enabling the model to make better-informed classification decisions.

The analysis of the outcomes of this study has resulted in the understanding that an efficient and customized pre-processing technique is necessary for generating an efficient model. The model generated based on the EfficientNet B0 architecture has achieved significant performance with sensitivity, specificity, and accuracy of 0.94, 0.91, and 0.93 respectively. This work has resulted in a conclusion that it is possible to achieve significantly improved results and outperform existing solutions by combining efficient pre-processing with deep learning pipelines. This work focused on classification of gastro-intestinal endoscopy videos. However, this solution can also be extended for endoscopic procedures like colonoscopy, arthroscopy, bronchoscopy, laparoscopy, and further.

6. FUTURE ENHANCEMENT

The normal and abnormal classification in the context of endoscopy videos would help in filtering out the extraneous frames and reduce the number of frames requiring the physician's attention to a great extent. In addition to classifying the frames as normal or abnormal, the regions of abnormality can be segmented to depict the areas of concern. Further, to generalize the model for unseen instances diverse and representative dataset that covers a wide range of examples and variations must be curated. Approaches for detecting, locating, and tracking abnormalities can be considered. The hybrid approach proposed in this work targetted at classification can also be extended for segmentation of regions of interest in endoscopy videos.

CONFLICT OF INTEREST STATEMENT

"The authors have no conflicts of interest to declare. Both the authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work."

REFERENCES

[1] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[2] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). https://doi.org/10.1609/aaai.v31i1.11231

[3] Tan, M., Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning. PMLR, pp. 6105-6114.

[4] Chen, Y., Lee, J. (2012). A review of machine-vision-based analysis of wireless capsule endoscopy video. Diagnostic and Therapeutic Endoscopy, 2012. https://doi.org/10.1155/2012/418037

[5] Alexandre, L.A., Nobre, N., Casteleiro, J. (2008). Color and position versus texture features for endoscopic polyp detection. In 2008 International Conference on BioMedical Engineering and Informatics. IEEE, 2: 38-42. https://doi.org/10.1109/BMEI.2008.246

[6] Ghosh, T., Fattah, S.A., Shahnaz, C., Kundu, A.K., Rizve, M.N. (2015). Block based histogram feature extraction method for bleeding detection in wireless capsule endoscopy. In TENCON 2015-2015 IEEE Region 10 Conference. IEEE, pp. 1-4. https://doi.org/10.1109/TENCON.2015.7373186

[7] Sekuboyina, A.K., Devarakonda, S.T., Seelamantula, C.S. (2017). A convolutional neural network approach for abnormality detection in wireless capsule endoscopy. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 1057-1060. https://doi.org/10.1109/ISBI.2017.7950698

[8] Ghosh, T., Fattah, S.A., Wahid, K.A. (2018). CHOBS: Color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video. IEEE Journal of Translational Engineering in Health and Medicine, 6: 1-12. https://doi.org/10.1109/JTEHM.2017.2756034

[9] Alexandre, L.A., Casteleiro, J., Nobreinst, N. (2007). Polyp detection in endoscopic video using svms. In Knowledge Discovery in Databases: PKDD 2007: 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, Springer Berlin Heidelberg. September 17-21, 2007. Proceedings, 11: 358-365. https://doi.org/10.1007/978-3-540-74976-9_34

[10] Bipin Dev, S.S. (2016). Gastro intestinal bleeding identification in endoscopy videos using canny edge detection and removal algorithm. International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064.

[11] Savazzi, M.L., Guarnaschelli, M. (2017). Machine learning for tissue classification in laryngeal endoscopic videos.

[12] Vasilakakis, M., Koulaouzidis, A., Yung, D.E., Plevris, J.N., Toth, E., Iakovidis, D.K. (2019). Follow-up on: Optimizing lesion detection in small bowel capsule endoscopy and beyond from present problems to future solutions. Expert Review of Gastroenterology & Hepatology, 13(2): 129-141. https://doi.org/10.1080/17474124.2019.1553616

[13] Byrne, M.F., Chapados, N., Soudan, F., Oertel, C., Pérez, M.L., Kelly, R., Iqbal, N., Chandelier, F., Rex, D.K. (2019). Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis

of unaltered videos of standard colonoscopy using a deep learning model. Gut, 68(1): 94-100. http://dx.doi.org/10.1136/gutjnl-2017-314547

[14] Yuan, Y., Meng, M.Q.H. (2017). Deep learning for polyp recognition in wireless capsule endoscopy images. Medical Physics, 44(4): 1379-1389. https://doi.org/10.1002/mp.12147

[15] Bernal, J., Tajkbaksh, N., Sanchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S.R., Fernández-Esparrach, G., Dray, X., Liang, J., Histace, A. (2017). Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. IEEE Transactions on Medical Imaging, 36(6): 1231-1249. https://doi.org/10.1109/TMI.2017.2664042

[16] Bouget, D., Allan, M., Stoyanov, D., Jannin, P. (2017). Vision-based and marker-less surgical tool detection and tracking: A review of the literature. Medical Image Analysis, 35: 633-654. https://doi.org/10.1016/j.media.2016.09.003

[17] Law, H., Ghani, K., Deng, J. (2017). Surgeon technical skill assessment using computer vision based analysis. In 2nd Machine Learning for Healthcare Conference. PMLR, 68: 88-99.

[18] Coelho, P., Pereira, A., Leite, A., Salgado, M., Cunha, A. (2018). A deep learning approach for red lesions detection in video capsule endoscopies. In Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27-29, 2018, Springer International Publishing. Proceedings, 15: 553-561. https://doi.org/10.1007/978-3-319-93000-8_63

[19] Gong, T., Li, S., Tan, C.L., Pang, B.C., Lim, C.T., Lee, C.K., Tian, Q., Zhang, Z. (2010). Automatic pathology annotation on medical images: A statistical machine translation framework. In 2010 20th International Conference on Pattern Recognition. IEEE, pp. 2504-2507. https://doi.org/10.1109/ICPR.2010.613

[20] Iakovidis, D.K., Georgakopoulos, S.V., Vasilakakis, M., Koulaouzidis, A., Plagianakos, V.P. (2018). Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. IEEE Transactions on Medical Imaging, 37(10): 2196-2210. https://doi.org/10.1109/TMI.2018.2837002

[21] Park, Y.S., Lee, J.W. (2020). Class-labeling method for designing a deep neural network of capsule endoscopic images using a lesion-focused knowledge model. Journal of Information Processing Systems, 16(1): 171-183. https://doi.org/10.3745/JIPS.02.0127

[22] Cao, Y., Yang, W., Chen, K., Ren, Y., Liao, Q. (2018). Capsule endoscopy image classification with deep convolutional neural networks. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC). IEEE, pp. 1584-1588. https://doi.org/10.1109/CompComm.2018.8780859

[23] Yang, W., Cao, Y., Zhao, Q., Ren, Y., Liao, Q. (2019). Lesion classification of wireless capsule endoscopy images. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 1238-1242. https://doi.org/10.1109/ISBI.2019.8759577

[24] Vasilakakis, M.D., Iakovidis, D.K., Spyrou, E., Koulaouzidis, A. (2018). DINOSARC: Color features based on selective aggregation of chromatic image components for wireless capsule endoscopy. Computational and Mathematical Methods in Medicine, 2018. https://doi.org/10.1155/2018/2026962

[25] Silva, J., Histace, A., Romain, O., Dray, X., Granado, B. (2014). Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery, 9: 283-293. https://doi.org/10.1007/s11548-013-0926-3

[26] Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., Riegler, M.A., Halvorsen, P. (2017). Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 164-169. https://doi.org/10.1145/3083187.3083212

[27] Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feußner, H., Navab, N. (2016). The TUM LapChole dataset for the M2CAI 2016 workflow challenge. arXiv Preprint arXiv: 1610.09278. https://doi.org/10.48550/arXiv.1610.09278

[28] Raju, A., Wang, S., Huang, J. (2016). M2cai surgical tool detection challenge report. In Workshop and Challenges on Modeling and Monitoring of Computer Assisted Intervention (M2CAI), Athens, Greece, Technical Report, pp. 1-4.

[29] Schoeffmann, K., Husslein, H., Kletz, S., Petscharnig, S., Muenzer, B., Beecks, C. (2018). Video retrieval in laparoscopic video recordings with dynamic content descriptors. Multimedia Tools and Applications, 77: 16813-16832. https://doi.org/10.1007/s11042-017-5252-2

[30] Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S. (2020). Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv Preprint arXiv, 2012.12453. https://doi.org/10.48550/arXiv.2012.12453

[31] Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data, 7(1): 283. https://doi.org/10.1038/s41597-020-00622-y

[32] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D. (2020). Kvasir-seg: A segmented polyp dataset. In MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Springer International Publishing. Proceedings, Part II 26: 451-462. https://doi.org/10.1007/978-3-030-37734-2_37

[33] Yu, T., Mutter, D., Marescaux, J., Padoy, N. (2018). Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition. arXiv Preprint arXiv, 1812.00033. https://doi.org/10.48550/arXiv.1812.00033

[34] Ali, S., Zhou, F., Daul, C., Braden, B., Bailey, A., Realdon, S., East, J., Wagnières, G., Loschenov, V., Grisan, E., Blondel, W., Rittscher, J. (2019). Endoscopy artifact detection (EAD 2019) challenge dataset. arXiv

Preprint arXiv, 1905.03209. https://doi.org/10.17632/C7FJBXCGJ9.1

[35] Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., Lux, M., Espeland, H., Petlund, A., Nguyen, D.T.D., Garcia-Ceja, E., Johansen, D., Schmidt, P.T., Toth, E., Hammer, H.L., de Lange, T., Riegler, M.A., Halvorsen, P. (2021). Kvasir-capsule, a video capsule endoscopy dataset. Scientific Data, 8(1): 142. https://doi.org/10.1038/s41597-021-00920-z

[36] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626.