Wireless Micro-Capsule Endoscopy Video Summarization,Classification And Segmentation Using Deep Neural Networks

Asst. Professor Aparna P, A Sai Sathwik, Om Anil Bhojraj, P Vishnu Sai
National Institute of Technology Karnataka

Abstract—**Automatic detection of diseases using Artificial Intelligence, specifically Deep Learning, is an active area of research. In particular, several studies have been done on Gastrointestinal (GI) tract disease diagnosis through endoscopic image and video processing and analysis. Here we report a model that uses Machine Learning techniques to summarize the microcapsule endoscopy video, Deep Neural Networks, to classify any diseases present in the GI tract and apply segmentation on one particular anomaly called a polyp, which is a mushroom-like appendage. The classification model combines three pre-trained Convolutional Neural Networks (CNN) architectures - MobileNet-V2, ResNet-50, and VGG-16. The extracted features are concatenated to create a single feature vector passed through a Global Average Pooling (GAP) layer to predict eight distinct GI anomalies. We used the Kvasir Dataset to train and test the classification model. The dataset consists of GI tract images classified into three critical anatomical landmarks and three clinically significant findings. Additionally, it contains two categories of images related to endoscopic polyp removal.**

The results were promising, with a training and testing accuracy of over 99%. The demand for Video processing is constantly growing in the medical electronics field. Despite the significant increase in modern workstations with good computational power, the amount of data in medical videos is snowballing, increasing the computational needs for processing and mining. Also, time is of the essence when diagnosing illnesses in the GI tract. This project discusses summarizing a microcapsule endoscopy video and implementation details and presents experimental results. The basics of deep learning methods in image classification and segmentation are successfully implemented for endoscopic images.

Index Terms—**Convolutional Neural Network, Deep Learning, Image Segmentation, Key Frame Extraction, Pretrained models, Transfer Learning, Video Summarization, Wireless Capsule Endoscopy.**

I. Introduction

Various gastrointestinal (GI) diseases, such as colorectal cancer and tumor, are the leading cause of death globally. The effective diagnosis of such GI diseases is a slow and time-consuming task. Moreover, most small GI lesions remain indistinguishable during the early stages, which ultimately evolves into fatal ailment. Therefore, the development of computerized approaches is essential to assist radiologists and physicians in effective diagnosis and treatment. Therefore, to make the real-time analysis of those images easy, significant efforts have been made over the last few decades to develop artificial intelligence (AI)-based computer-aided diagnosis (CAD) tools and applications in various medical fields. In the field of endoscopy, deep learning techniques were applied in the recent AI-based CAD tools (a set of advanced machine learning algorithms) to analyze various types of endoscopic scans. In general, deep learning algorithms extract the optimal representations of training data. Artificial neural networks (ANNs) that logically emulate the structure and activity of the brain neurons on a computer are the key components of such deep learning-based image analysis tools. Various types of ANNs, including convolutional neural networks (CNNs), were proposed for image recognition.

II. Motivation

A GI video taken during capsule endoscopy is 10 to 48 hours. For a human being to analyze the entire video and make a diagnosis takes much time.

Deep learning techniques can significantly reduce the time taken to study endoscopy videos. As a result, in times of emergency, a quick and accurate diagnosis is achievable and can save lives. All the previous studies proved very efficient in detecting or classifying a particular GI disease. In endoscopic image classification, a model with very high efficiency is respected because it will give almost perfect results compared to other models with less efficiency. So, we want to create a new model with better efficiency than the previously developed models.

III. Related work

In recent years, deep learning-based algorithms have been widely used in the field of endoscopy. Most of the previous studies have been carried out to perform the detection and classification of different types of GI polyps in the field of CE. The following table shows the comparison of our proposed method and the existing techniques for endoscopy disease classification (advantages and disadvantages of already existing methods and our proposed method).

Table I: Comparison of proposed method with existing methods for classifying anomaly in an infected frame.

Endoscopy Type	Method	Purpose	No. of classes	Advantages	Disadvantages
CE	Log Gabor filter,SUSAN edge detection and SVM classifier	Small bowel polyps and ulcers detection	2	Computationally efficient	Limited dataset and number of classes Low detection performance.
CE	Texture features (ULBP, wavelet)	Small bowel polyps and Polyp detection in GI tract	2	Robust to illumination change and scale invariant	Limited dataset and number of classes.
CE	Texture features (LBP, wavelet) + SVM	Tumor recognition in the digestive tract	2	Invariant to illumination change Extract multiresolution features	Limited dataset and number of classes.
CE	Texture features (SIFT, Saliency) + SVM	Polyp classification	2	Extract scale invariant features	Limited dataset and number of classes.
CE	Texture features(SIFT, HoG, LBP, CLBP, ULBP) + SVM, FLDA	Polyp Detection	2	Extract scale invariant features High classification performance	Limited dataset and number of classes.
CE	CNN	Small intestine movement characterization	6	High classification performance	Limited number of classes.
CE	CNN	Celiac disease classification	2	High sensitivity and specificity	Limited dataset and number of classes.
CE	CNN	Hookworm detection	2	Edge extraction network results in better performance	Limited number of classes.
EGD	CNN	H. pylori infection detection	9	Comparable performance of second CNN with the clinical diagnosis reference standard	CAD performance should be enhanced. A limited number of classes.
EGD	CNN	Anatomical classification of GI images	6	High classification performance Computationally efficient	Limited number of classes Only used for anatomical classification.
EGD	CNN-based SSD detector	nGastric cancer detection	2	High sensitivity Computationally efficient	Overall low positive prediction value Limited dataset and number of classes.
Colonoscopy	CNN	Colorectal polyp detection and classification	2	Computationally efficient	Limited dataset and number of classes Low classification performance.
Colonoscopy	CNN	Real-time colorectal polyp type analysis	2	High detection performance	Limited number of classes Low specificity.
Colonoscopy	Online and offline 3D-CNN	Detection of colorectal polyps 2	2	High accuracy and sensitivity	CAD performance should be enhanced.
CE	ResNet + Mo bileNetV2 + VGG (proposed one)	8 types of GI diseases(refer section IV)	8	Computationally efficient High classification performance	Parallel training of three pretrainied CNNs requires comparitively less time.

IV. Overview

A brief flowchart of our method for the classification of multiple GI diseases based on Deep Neural networks is shown in figure 2. For training and testing the designed model, the KVASIR Dataset was used. This dataset has endoscopic imagery (taken by a microcapsule) inside the GI tract with eight classes. This dataset has 1000 images belonging to each of 8 classes making up a total of 8000 images. The anomalies it has are; **z-lines, pylorus and cecum and pathological findings; esophagitis, polyps, and ulcerative colitis**. This dataset consists of two sets of images related to the removal of polyps called dyed and lifted polyps and the dyed resection margins. The pre-defined models are selected based on the accuracy they have in the predictions and the vector sizes they can have. The comparisons of the accuracy of some pre-defined models are given in table(II).

Classified video datasets of each anomaly, available in the Hyper KVASIR dataset, are used for video summarizing using key frame extraction method. And then extracted frames are passed into our model fig(2) for identification of the anomaly. The extracted frame will be passed for segmentation if any polyp was found in the frame.

V. Approach

GI videos are rife with redundant frames that delay the diagnosis. So, our first step is to extract all the frames with frame numbers and apply K-Means Clustering to extract the key frames. The frames are saved with frame numbers to ensure the diagnostician can quickly analyze the frames in the vicinity of the frame identified with an infected region. Since all extracted frames do not necessarily indicate an anomaly in the GI tract, the frames with abnormalities are identified and classified using the designed model. Among these classified frames, the frames with polyps, if found, are passed on for segmentation to highlight the polyp. Therefore, the approach can be divided into three parts: **Video Summarization, Classification, and Segmentation**, which are discussed in more detail in the following sections. A brief flowchart of the approach is given in fig(1)

A. *Video Summarization*

Video summarization plays a crucial role in Micro Capsule Endoscopy because one endoscopic video is a two to four-hour video (on average) and can sometimes run up to 6 hours long. This makes it very hard for a diagnostician to find the affected part in the GI tract. Endoscopy video summarization helps narrow the search and results in the physician treating patients faster. This can be done using two methods.

- **Key Frame Extraction:** Endoscopy videos contain the frames of the entire GI tract, and this includes the healthy parts as well. The key frames are extracted from the video to avoid the redundant frames of healthy tissue.
- **Video Skimming:** The parts of the video which has the required frames are joined to form a short video which saves much time and can provide a summary of the original video.

Generally, an endoscopy video is skimmed first to get a short video containing the infected portion, and then frames are extracted from it to identify the anomaly. We summarize an endoscopy video using the key frame extraction method, which extracts critical frames from the video and then passes them to the model to identify the GI anomaly in the patient's tract. Kmeans Clustering and Gaussian clustering are popular methods used for summarizing videos by key frame extracting. We used Kmeans Clustering, which allows us to extract key frames from the video by rejecting some redundancy. After this step, we implemented an algorithm that helps in removing further redundancy on our wish. This algorithm compares frames extracted using the Kmeans algorithm and deletes frames with 90% similarities. This can be performed several times until there are no redundant frames left. Now, as only the key frames are extracted (most of the redundancy is removed), these are passed on to the classification section for identifying the anomaly if present in the frame.

B. *Classification*

The model had to be trained to identify the eight types of anomalies considered for classification. We use different pre-trained CNN feature extractors for our initial GI endoscopic imagery dataset experiments. This approach of using different pre-trained models of CNNs and using those results for feature extraction or any other purposes is called Transfer Learning (Discussed in detail in section(V-B1)). We feed these extracted feature vectors into an independent classifier to get the predicted class labels.

1) Transfer Learning: Transfer learning is a deep learning technique where one model which is trained on a particular task is re-purposed on another job that is related to the first one.
Basically, Transfer Learning has two approaches:

- Develop Model Approach
- Pre-trained Model Approach

1) **Develop Model Approach:** This approach follows the steps in the order shown below.
    - **Select Source Task**. First, we select a predictive modeling problem with an abundance of data. There should be some relationship between the input, output, and concepts learned while mapping from input to output data.
    - **Develop Source Model**. Now, we must develop the best model for this first task. This model must be better than a simple model to conclude that some features have been learned.
    - **Reuse Model**. The model designed for the source task can then be reused as a base for another model on the second task based on the application. Depending on the modeling technique, this may involve using parts or all of the primary model.
    - **Tune Model**. Optionally, the model may need to be adapted or refined on the input-output pair data available for the task of interest.
2) **Pre-trained Model Approach:** This approach has the following steps:

Table II: Comparisons Of Top-1 And Top-5 Accuracies Of Some Pre-defined Models

Model	Size	Top-1 accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88MB	0.790	0.945	22,910,480	126
VGG16	528MB	0.713	0.910	138,357,544	23
VGG19	549MB	0.713	0.900	143,667,240	26
ResNet50	98MB	0.749	0.921	25,636,712	-
ResNet101	171MB	0.764	0.928	44,707,176	-
ResNet152	232MB	0.766	0.931	60,419,944	-
InceptionV3	92MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215MB	0.903	0.953	55,873,736	572
MobileNet	16MB	0.704	0.895	4,253,864	88
MobileNetV2	14MB	0.713	0.901	3,538,984	88
DenseNet121	33MB	0.750	0.923	8,062,504	121
DenseNet169	57MB	0.762	0.932	14,307,880	169
DenseNet201	80MB	0.773	0.936	20,242,984	201
NASNetMobile	23MB	0.744	0.919	5,326,716	-
NASNetLarge	343MB	0.825	0.960	88,949,818	-

Note:. Depth includes activation layers and batch normalization layers(Values are obtained from Keras website), Online Link

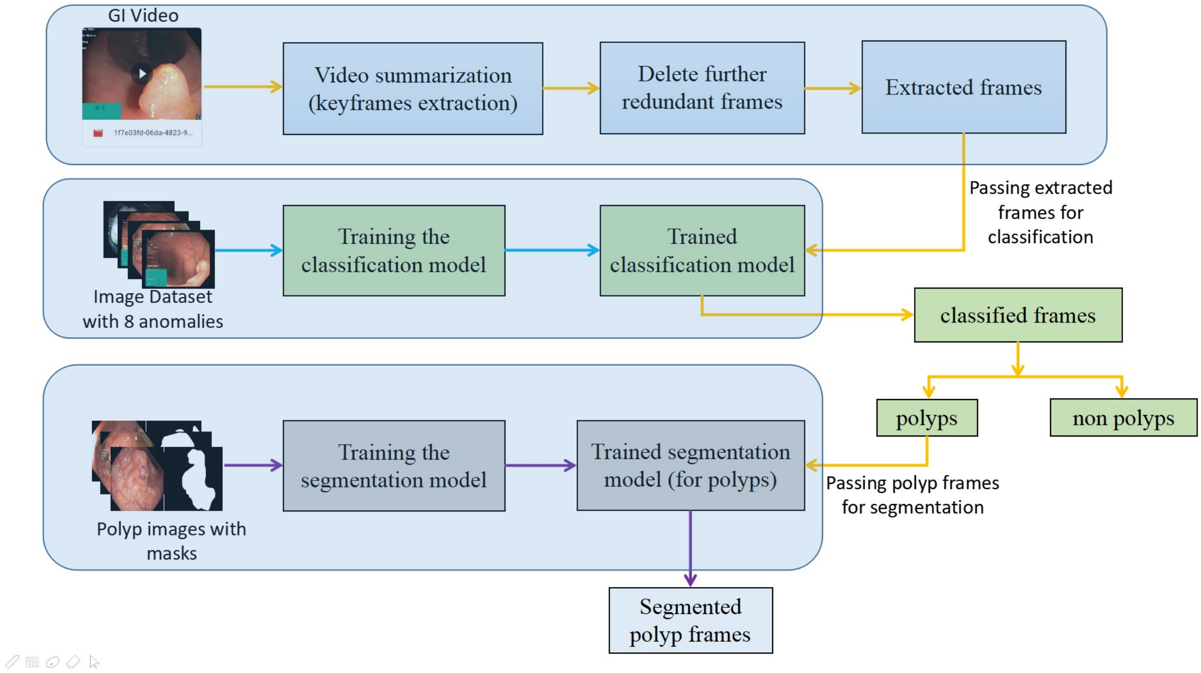

GI Video
Video summarization (keyframes extraction)
Delete further redundant frames
Extracted frames
Passing extracted frames for classification
Image Dataset with 8 anomalies
Training the classification model
Trained classification model
classified frames
polyps
non polyps
Polyp images with masks
Training the segmentation model
Trained segmentation model (for polyps)
Passing polyp frames for segmentation
Segmented polyp frames

Figure 1: Flowchart of the Approach

- **Select Source Model**. A pre-trained source model is chosen from available models. It can be selected from models released by many research institutes on large and challenging datasets.
- **Reuse Model**. The pre-trained model can then be used as the starting step for a model on the second task of interest. This approach may involve using all or parts of the model depending on the modeling technique.
- **Tune Model**. Optionally, the model may need to be adapted or refined on the input-output pair data available for the task of interest.

The ”pre-trained model” approach has the advantage of reusing an already developed source model and tuning it for other related problems. The main reason behind choosing the transfer learning technique is the advantage of pre-trained models used in our model and the ease of computation. The comparison of some pre-trained models is shown in the tableII.

The proposed approach for anomaly classification is based on the steps described below.

During the first step in classification, we preprocess our dataset, as shown below.

a) ***Preprocessing:*** The imagery in the dataset came with varied resolutions, from 720 x 576 to 1920 x 1072 pixels. Therefore, we downsample the images into 224 x 224 pixels.

Next, we use a set of prominent CNNs with a global average pooling layer to obtain feature vectors. Then the ultimate feature vector for the classification task is gained by appending vectors from the previous step. Subsequently, we feed this feature vector into a three-hidden-layer Artificial Neural Network (ANN), which contains 128 hidden units with a ReLU activation function.

Initially, we designed a model similar to the one dis-

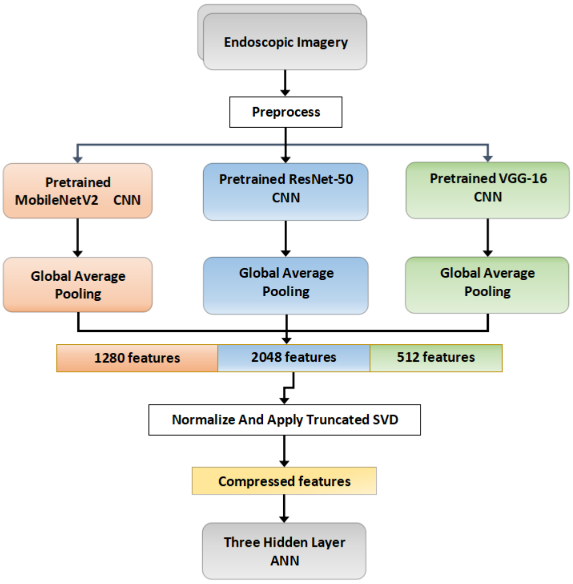
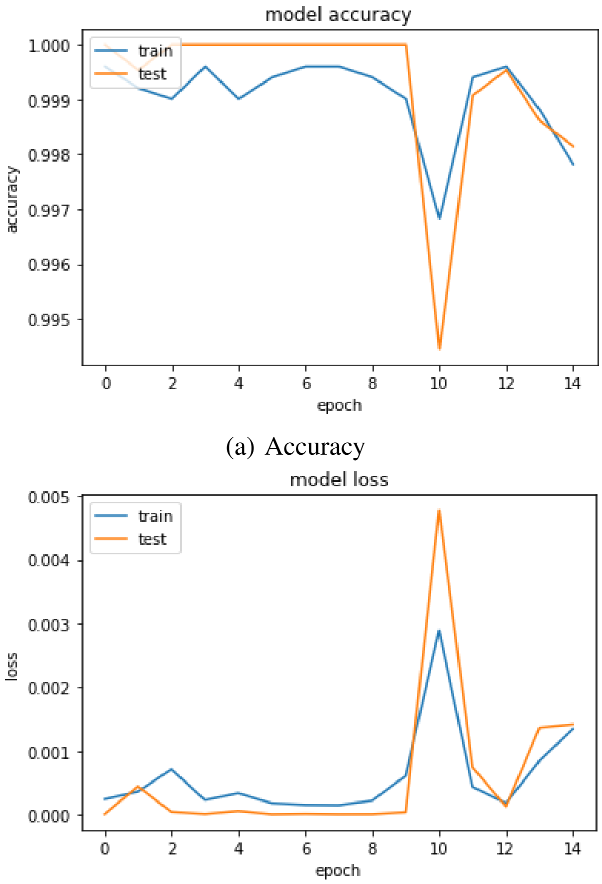


Figure 2: Flowchart of the classification model



(a) Accuracy

(b) Loss

Figure 3: Train metrics vs Validation(Test) metrics of the model

cussed in [6] and then redesigned it by making essential changes that resulted in much better results. We use a combination of three CNN feature extractors, namely MobileNetV2, ResNet-50, and VGG-16. Each image in the dataset is fed to the MobileNetV2 CNN model and extracted 1280 features by combining the GAP layer after the last pooling layer. Correspondingly, our dataset is processed over ResNet-50 (2048 features) and VGG-16 (512 features), similar to the MobileNetV2. We obtain a feature vector with 3840 (1280 + 512 + 512) features for each image. Subsequently, we normalize our feature vectors to remove redundant and noisy features and pass them to the truncated SVD (Singular Value Decomposition) to remove further redundancies. After normalization and SVD, we compress all the features and add an ANN to classify abnormality in an infected GI image (among the eight classes considered).

The ANN we have designed has three dense layers connected in series, which generate 1024, 512, and 128 outputs, respectively. This ANN is connected to the output of the SVD. The designed model is shown in figure 2. Now we train the model by using 90% of the images (i.e., 900 images from each of 8 classes = 7200 images) from each of 8 categories, and the remaining 10% (i.e., 100 images from each of 8 classes = 800 images) are used for testing purpose.

Now we have designed a model that helps segment polyps if found in the GI video. The design of the model is discussed in the section(V-C1)

The proposed approach resulted a guaranteed accuracy of 99% as shown fig(3(a)). The fig(3(b)) shows the train loss vs test loss for 15 epochs.

C. *Segmentation*

U-Net was originally developed and first used for biomedical image segmentation. Its architecture can be broadly thought of as an encoder network followed by a decoder network. Unlike classification, where the result of the deep network is the only important thing, semantic segmentation requires discrimination at the pixel level and a mechanism to project the discriminative features learned at different stages of the encoder onto the pixel space.

- The encoder is the first half of the architecture diagram (Figure (4)). It is usually a pre-trained classification network like VGG/ResNet, where you apply convolution blocks followed by a max pool layer that downsamples to encode the input image into feature representations at multiple levels.
- The decoder is the second half of the architecture. The goal is to semantically project the discriminative features (lower resolution) learned by the encoder onto the pixel space (higher resolution) to get a dense classification. The decoder consists of upsampling and concatenation followed by regular convolution operations.

*1) **U-Net with Densenet169 backbone**:* U-Net is an extension of an encoder-decoder fully convolutional network. The convolutional layers perform feature extraction by employing the filters in these layers to learn low and high-dimensional features as they iteratively get trained. The intuition behind U-Net is to encode the image passing it through a CNN as it gets downsampled, and then decode it back or upsample it to obtain the segmentation mask. The features to be detected in the mask depend on the learned weight filters, upsampling downsampling blocks

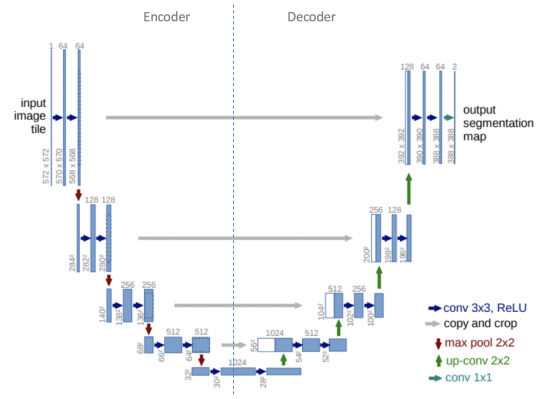


Figure 4: U-Net Architecture

(which can also be made learnable), and the concatenations skip connections. The backbone is the architectural element that defines how these layers are arranged in the encoder network and determine how the decoder network should be built.

The backbones used are often Vanilla CNNs such as VGG, ResNet, Inception, EfficientNet, etc. which perform encoding and downsampling by themselves. These networks are taken and their counterparts are built to perform decoding and upsampling to form the final U-Net. There has been much research on making these networks better by using backbones that are capable of extracting features better.
The backbone for our segmentation model is Densenet169.

*2) **Training Dataset**:* The dataset used for training our model is again from the Hyper-Kvasir dataset. It consists of 1000 polyp images and their corresponding masks, all in JPEG format. In the mask, the pixels depicting polyp tissue, the region of interest, are represented by the foreground (white), while the background (black) does not contain polyp pixels. The bounding box is defined as the outermost pixels of the found polyp.
The dataset size was increased to 6000 using different augmentation techniques. The augmentations used were Vertical Flip, Horizontal Flip, Random Brightness, Random Contrast and Blur. The model was trained for 20 epochs on a Nvidia RTX 2060 graphics driver.

VI. Classification of images from Image Dataset

The images used for testing the neural network are successfully classified (only ten images from each of the eight classes are shown in the following figures for the report purpose). The following figures show the images captioned with the group to which they belong. The test data has 800 images. One hundred images belong to each of the eight classes in the order 1-dyed polyps, 2-dyed margin, 3-esophagitis, 4-normal cecum, 5-normal pylorus, 6-zline, 7-polyps, 8-ulcerative colitis. The 50th image from the test images of each class is predicted, and the results are shown in fig(5).

VII. Classification of frames extracted from Video Dataset

The first step in GI video summarization is to extract all the key frames using K-Means clustering. As there will be many duplicate frames (similar frames), it becomes tough to check for anomalies in each frame. Hence, we delete the identical frames with 90% similarities using a simple algorithm to calculate the percentage difference between two images. Duplicate frame deletion can be done multiple times until no duplicates are left. This results in a comparatively significantly fewer number of frames to be checked for the presence of anomalies. Now, as the video dataset is already classified, it becomes easy to verify the results when the extracted frames are classified using our model (shown in fig 2). Each frame classified is printed with the probability of that anomaly. If the probability is less than 0.8, it is classified as ”Not identified”.

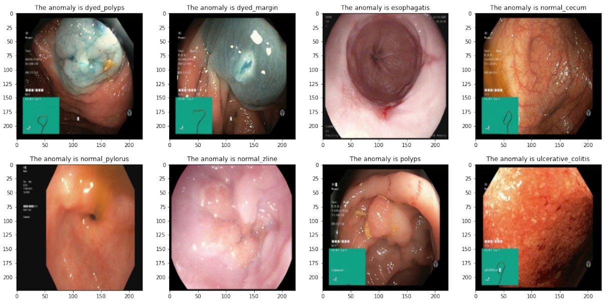

The anomaly is dyed_polyps
The anomaly is dyed_margin
The anomaly is esophagatis
The anomaly is normal_cecum
The anomaly is normal_pylorus
The anomaly is normal_zline
The anomaly is polyps
The anomaly is ulcerative_colitis
0
25
50
75
100
125
150
175
200
0
50
100
150
200

Figure 5: All Anomalies(one image from each class)

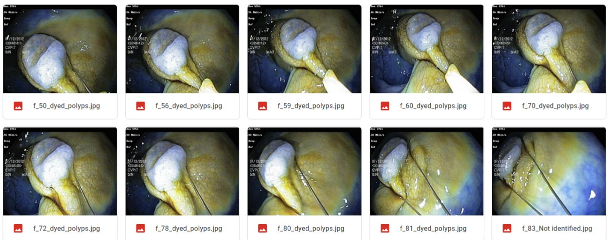


Figure 6: Prediction of extracted key frames(only 10 are shown)

VIII. Polyps segmentation

The Polyps segmentation dataset consists of 6000 polyp images and their corresponding masks, all in JPEG format. In the mask, the pixels depicting polyp tissue, the region of interest, are represented by the foreground (white mask), while the background (in black) does not contain polyp pixels. The bounding box is defined as the outermost pixels of the found polyp. Only ten extracted polyp images from the test dataset are segmented for report purposes.

A. Segmentation of polyps using Polyps Dataset

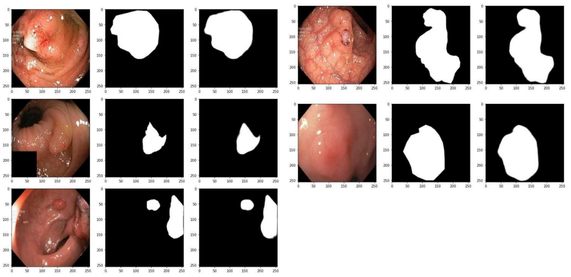


Figure 7: Polyps Segmentation using test dataset

B. Segmentation using polyps frames extracted from video dataset

The key frames extracted from the polyp video dataset are considered for segmentation and the results corresponding to frames classified as polyps are shown below.

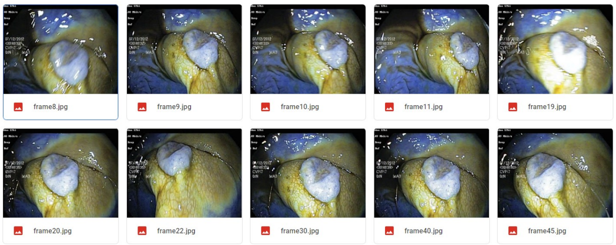


Figure 8: Polyp frames extracted from video

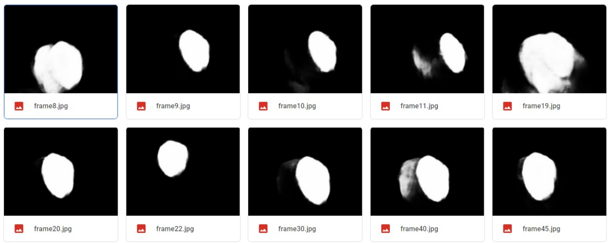


Figure 9: Segmented polyp frames extracted from video

NOTE : Only the ten polyp frames from the extracted key frames are shown in the above figures.The frame numbers of each frame can be seen in the frame label.

```
25/25 [==============================]
- 468s 19s/step - loss: 0.0984 - accuracy:
0.9225 - precision_1: 0.9279 - recall_1: 0.9175
[0.09840794652700424, 0.9225000143051147,
0.9279392957687378, 0.9175000190734863]
```

References

[1] **"Wireless capsule endoscopy "**, *Link to website*

[2] **"The Kvasir Dataset "**, *Link to the website*

[3] Hanna Borgli, VajiraThambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, CarstenGriwodz,Håkon K. Stensland, EnriqueGarcia-Ceja , Peter T. Schmidt,Hugo L. Hammer,Michael A.Riegler ,Pål Halvorsen Thomas de Lange, **"HyperKvasir, A Comprehensive Multi-Class Image And Video Dataset For Gastrointestinal Endoscopy"**,2020 , *Link to the research paper*

[4] Konstantin Pogorelov,Kristin Ranheim Randel,Carsten Griwodz,Sigrun Losada Eskeland,Thomas de Lange,Dag Johansen,Concetto Spampinato,Duc-Tien Dang-Nguyen,Mathias Lux,Peter Thelin Schmidt,Michael Riegler And Pål Halvorsen **"A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection"**, 2017 , *Link to the research paper*

[5] Anjany Kumar Sekuboyina, Surya Teja Devarakonda, and Chandra Sekhar Seelamantula,**"A Convolutional Neural Network Approach For Abnormality Detection In Wireless Capsule Endoscopy"**,2017 , *Link to the research paper*

[6] Chathurika Gamage,Isuru Wijesinghe, Charith Chitraranjan, Indika Perera, **"GI-Net: Anomalies Classification in Gastrointestinal Tract through Endoscopic Imagery with Deep Learning"**,2019 IEEE , *Link to the research paper*

[7] Husanbir Singh Pannu, Sahil Ahuja ,Nitin Dang, Sahil Soni,Avleen Kaur Malhi, **"Deep learning based image classification for intestinal hemorrhage"**,2020 , *Link to the research paper*

[8] Jamil Ahmad , Khan Muhammad , Mi Young Lee , Sung Wook Baik, **"Endoscopic Image Classification and Retrieval using Clustered Convolutional Features"**,2017 , *Link to the research paper*

[9] Muhammad Owais, Muhammad Arsalan, Jiho Choi, Tahir Mahmood and Kang Ryoung Park, **"Artificial Intelligence-Based Classification of Multiple Gastrointestinal Diseases Using Endoscopy Videos for Clinical Diagnosis"**,2019 , *Link to the research paper*

[10] Jin Chen, Yuexian Zou, Yi Wang, **"Wireless Capsule Endoscopy Video Summarization:A Learning Approach Based on Siamese Neural Network and Support Vector Machine "**,2016 , *Link to the research paper*

[11] Shees Nadeem, Muhammad Atif Tahir, Syed Sadiq Ali Naqvi, and Muhammad Zaid, **"Ensemble of Texture and Deep Learning Features for finding abnormalities in the Gastro-Intestinal Tract"**,2018, *Link to the research paper*

[12] V. B. Surya Prasath, **"Polyp Detection and Segmentation from Video Capsule Endoscopy: A Review"**,2016 , *Link to the research paper*

[13] Sruthi Jadon, Mahmood Jasim, **"Unsupervised video summarization framework using keyframe extraction and video skimming Convolutional"**,June 2020 ,

*[14] Surya Remanan, **"Deep learning-based video summarization — A detailed exploration"**,2020 , Link to the article*

*[15] Alok, **"Video Key Frame Extraction With katna"**,2019 , Link to the article*

*[16] **"Katna documentation"** Link to the documentation*