Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 477–484

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Polyps Detection in Colonoscopies

José Ribeiro[a], Sara Nóbrega[a], António Cunha[a,b,*]

[a]*Universidade de Trás-os-Montes e Alto Douro, Vila Real, 5000-801, Portugal*
[b]*INESC TEC – INESC Technology and Science, Porto, 4200-465, Portugal*

Abstract

A colonic polyp is a growth in the lining of the colon or rectum and can be detected through colonoscopies. The efficiency of colonoscopies depends on the number of polyps detected. However, detecting and classifying polyps is difficult, tedious, and prone to error. Knowing that this process's performance is far from perfect, the objective of this project is to help colonoscopists in the detection of polyps during the medical intervention, using Deep Learning (DL) alongside the image recognition capabilities of Convolutional Neural Networks (CNN) models that can process colonoscopy images at high speed in real-time.
In this paper, were tested different state-of-the-art CNNs using a transfer learning approach, achieving an average accuracy of 95,70% in the polyp detection task. Multiple public datasets were used in this study to train, test, and evaluate the classifiers. The negative class included images representative of healthy tissue as well as other pathologies, so the models would not mistake other diseases as polyps.

Keywords: Deep Learning; Convolutional Neural Networks; Colorectal Cancer Prevention; Colon polyps; Polyp detection.

* Corresponding author.
E-mail address: acunha@utad.pt

1877-0509
10.1016/j.procs.2021.12.039

1. Introduction

According to [8], colorectal cancer (CRC) is a major cause of mortality throughout the globe, accounting for over 9% of all cancer incidence. Additionally, it is the third most common cancer worldwide and the fourth most common cause of death, affecting both genders equally. Developed countries are the ones with the highest incidence rates [8].

CRC survival is highly dependent upon the stage of the disease at diagnosis and typically ranges from a 90% 5-year survival rate for cancers detected at the localized stage; 70% for regional cancer; to 10% for people diagnosed with distant metastatic cancer [8]. In general, the earlier the stage at diagnosis, the higher the chance of survival.

Polyp detection is vital for cancer prevention since early-stage detection significantly increases the chance of an effective treatment. Colorectal polyps are the growth of tissue from a mucous membrane, projecting itself into the intestine lumen (Fig. 1).

Fig. 1. Example of polyps taken during colonoscopies present in Kvasir-v2 dataset.

Initially, colonic polyps are benign, but over time, some of them can become malignant, and removing them is the most effective form of treatment [10]. Colonoscopy is usually the exam performed, and any polyps that are found during the examination are removed and posteriorly analyzed.

DL methods, such as CNNs, have been revolutionizing conventional techniques, presenting significantly better results than all the previous methods, extracting the characteristics automatically from the data [6,8].

Urban et al. in [6] built a polyp detection model using DL for real-time colonoscopies with the main goal of increasing the Adenoma Detection Rate (ADR), achieving an accuracy of 96,40% using a pre-trained CNN on the ImageNet.

Bardhi et. Al. in [9] presented DL techniques for polyp detection, classification, segmentation, and localization, using CNN-AutoEncoder, an algorithm considered by the authors promising since no image pre-processing was performed before training the model. The authors raised some concerns regarding the need for more images in the dataset and the need for a diverse range of polyp images.

In this work were trained models that use different kinds of pathologies on the non-polyp dataset, and we are going to evaluate the efficiency of this approach.

2. Methodology

The pipeline present in Fig. 2 shows the different phases executed during this work. Three public datasets were merged, aiming to evaluate different state-of-the-art architectures. The images were resized to 224x224 and suffered some pre-processing, where alphanumeric characters were removed. All the images were included into one of the possible classes (polyp and non-polyp), and for last, the data was split into train, validation, and test sets. Several models were trained on the train and validation sets and evaluated in the test set.

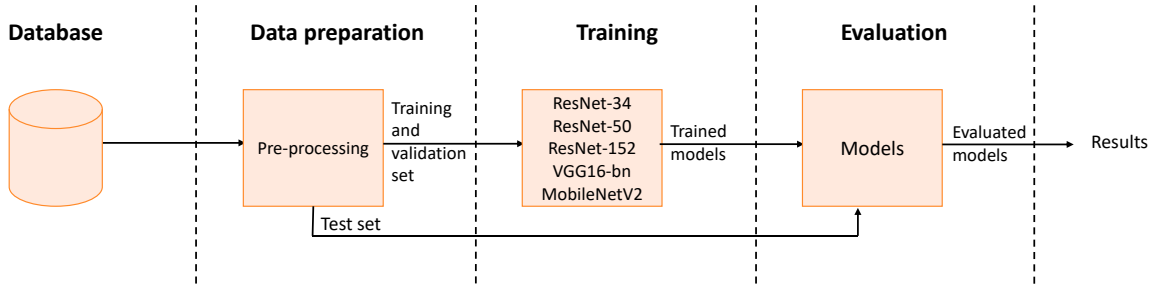

Fig. 2. Overview of the developed work.

2.1. Data preparation

The database used is a compilation of 4,808 images from three distinct datasets, that contains 2 classes, being them, polyps and non-polyps as shown in Table 1.

Table 1. Number of polyp and non-polyp images used for the train, validation, and test sets.

Dataset	Train	Validation	Test	Total
Polyp	1.208	300	300	1.808
Non-polyp	1.600	700	700	3.000
Total	2.808	1.000	1.000	4.808

The main dataset used is the Kvasir-v2 [1] which has images from different types of pathologies as can be seen in Fig. 3.

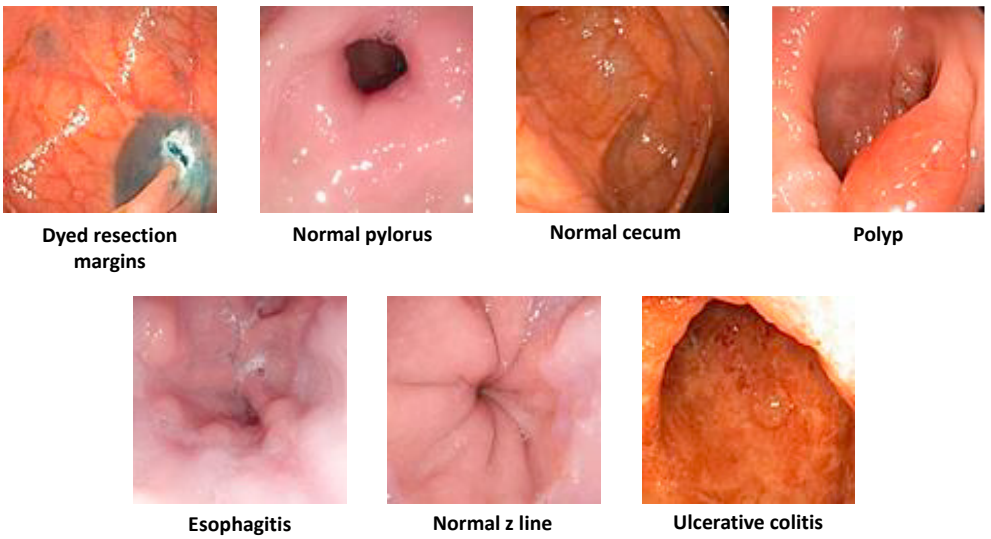

Fig. 3. Representative images were removed from Kvasir-v2 of the different classes included in the non-polyp and polyp class.

Kvasir-v2 dataset was split into two classes: polyps, and non-polyps, which contains normal mucosa images (normal pylorus, cecum, and z line) alongside with images from other pathologies (esophagitis and ulcerative colitis) and therapeutic intervention (dyed resection margins). Including other pathological findings and therapeutic

intervention into the non-polyp class will help the binary classification models to learn not to only differentiate polyps from normal mucosa, but also to distinguish polyps from other pathological findings.

Since the number of non-polyp images was superior to the number of polyp images, the dataset was imbalanced. Hence, polyp images from CVC-ClinicDB [2] and the ETIS-LaribPolypDB [3] datasets were included, aiming to balance the dataset.

In some images, it was present alphanumerical information as well as a black margin. Since this information could interfere with the learning process of the models, it was removed. All the images in the final dataset were resized to 224x224.

A bigger resolution could be used, but since Google Colababoratory was the development tool used for the development and it has limited resources, the images were scaled down to decrease the use of resources and improve run times because of that, 224x224 seemed a good speed/resources choice, the same approach was used in [6] when they tested a higher resolution, but it achieved near identical results to the 224x224 resolution. However, according to [6], the computational resources more than doubles when a higher resolution is considered.

2.2. Training

In this study were trained different CNNs for the binary classification task and a transfer learning approach was applied. The architectures selected were: ResNet-34 [14], ResNet-50 [14], ResNet-152 [14], MobileNetV2 [15], and VGG16_bn [16]. All the models were pre-trained on the ImageNet dataset [13], which were posteriorly fine-tuned in this study to distinguish polyp from non-polyp images.

These different transfer learning models were selected based on their performance and popularity. The selected architectures have a broad range of layers which allowed to test the impact of different architecture depths in this task. Multiple variants of the ResNet architectures were selected aiming to evaluate the same architectural strategy with 34, 50, and 152 layers of depth. In addition, were also tested MobileNetV2 [15] (88 layers of depth) and VGG16-bn [16] (23 layers of depth) to observe the performance of different architectural strategies. Note also that ResNet [14] and VGG [16] models were designed for the resolution used in the dataset (224x224).

All experiments were implemented inside Google Colaboratory using the fast.ai [11] and TensorFlow [12] software libraries.

2.3. Evaluation

Four metrics were used to evaluate the models based on the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN):

- Accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$), which is the percentage of correct predictions.
- Recall ($\frac{TP}{TP+FN}$), which captures the ability of the classifier to find all the positive samples.
- Precision ($\frac{TP}{TP+FP}$), which is the ability of the classifier not to label a negative sample positive.
- The F_1 score ($\frac{2*(precision*recall)}{precision+recall}$), which is the harmonic mean of precision and recall, computes values in the range [0,1].

3. Results and Discussion

To train the models was used the Smith's One Cycle Policy [17] which uses large, cyclical learning rates to train models quicker, with higher accuracy and less overfitting. On average for each model, it only needed around 6 to 8 epochs of training to achieve good results. In the rare cases of overfitting, some minor fine-tuning was performed.

The confusion matrix for each one of the developed models can be consulted in Table 2.

Table 2. Confusion matrix for all the developed models. "NP" and "P" refer to "Non-polyp" and "Polyp", respectively.

Models		**Predicted class**									
		ResNet-34		*ResNet-50*		*ResNet-152*		*MobileNetV2*		*VGG16_bn*	
		NP	P	NP	P	NP	P	NP	P	NP	P
True class	NP	679	21	684	16	683	17	680	20	691	9
	P	30	270	21	279	17	283	22	278	43	257

The metrics referred previously were calculated using the number of TP, TN, FP, and FN, present in Table 2, and obtained by the models in the test set. The performance of the models in the test set can be observed in Table 3.

Table 3. Results obtained by evaluated models.

Model	**TP**	**TN**	**FP**	**FN**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F_1 Score**
ResNet-34 [14]	270	679	21	30	94,90	92,80	90,00	0,90
ResNet-50 [14]	279	684	16	21	96,30	94,60	93,00	0,90
ResNet-152 [14]	**283**	683	17	**17**	**96,60**	94,30	**94,30**	0,90
MobileNetV2 [15]	278	680	20	22	95,80	93,30	92,70	0,90
VGG16_bn [16]	257	**691**	**9**	43	94,80	**96,70**	85,70	0,90

Observing the performance of the models it is possible to see all had a similar performance. Since the main goal is to detect colonic polyps, it was selected ResNet-152 [14], the architecture with the higher accuracy and recall on the test set, for a deeper analysis. Comparing the performance of ResNet-50 [14] with the performance of ResNet-152 [14], it is possible to observe through the metrics that ResNet-152 [14] only had a slightly better performance. Hence, a more complete analysis of the performance of ResNet-50 [14] will also be made.

In Fig. 4, it is possible to observe some of the prediction results made by ResNet-152 [14]. The selected results represented the TP, TN, FP, and FN where the model had high confidence in the prediction. Besides, Fig.4 also shows some predictions where the model had some difficulty in deciding if in the images was present a polyp.

For each image, Fig.4 contains the original label, the predicted label, and the probability outputted by the model that translates the confidence of the model in the prediction made.

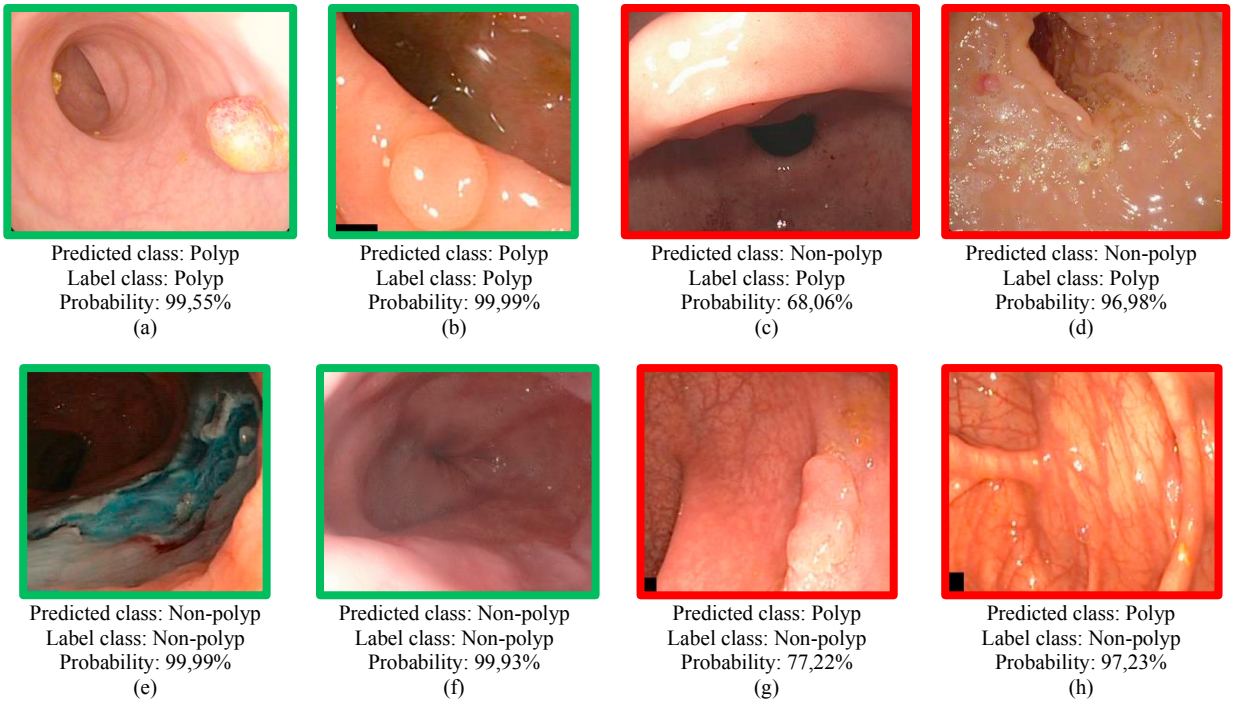

Fig. 4. Images and respective predictions made by the ResNet-152 [14] model.

Observing the two examples of TP with a high confidence score, present in Fig.4 (a) and Fig.4 (b), it is possible to observe that the polyp present in each one of the images has a considerable size, occupying a large area in the image totality. Hence, the ResNet-152 [14] could classify with a high probability both cases as polyp representative.

In Fig.4 (d), the polyp occupies a reduced area since its small (reddish area present on the top left side of the image), which induced the network in error, classifying the image as non-polyp representative.

Regarding the two TN, present in Fig.4 (e) and Fig. 4 (f), it is possible to observe in the first one that is an image representative of a dyed resection margin. The coloration associated with the image, produced by the dyer, could lead the network to affirm with certainty that the image was not a polyp. Regarding Fig. 4 (f) TN, it is possible to see that this image shows a normal gastrointestinal mucosa, without any texture that could induce the model in error.

For last, observing the FN present in Fig.4 (g), it is possible to observe that although the polyp has a considerable size (voluminous area present in the lower right corner), it is inserted into an area with intestinal folds. Hence, the model may have associated the polyp with this structure that is present in a normal healthy intestine.

The opposite occurred in the FP present in Fig.4 (c), where the normal anatomical intestine structure was classified as polyp representative. Nevertheless, in both situations, despite the model has failed, the confidence in those predictions was close to the threshold confidence used in this study (50%), indicating that the model had some difficulty in classifying the images present in both cases. In the same line of thinking and observing Fig.4 (h), a considerable area of the image includes folds, and the model may have classified them as polyp representative.

The same situations analyzed for ResNet-152 [14], were analyzed for ResNet-50. Fig.5 shows these situations in a summarized way.

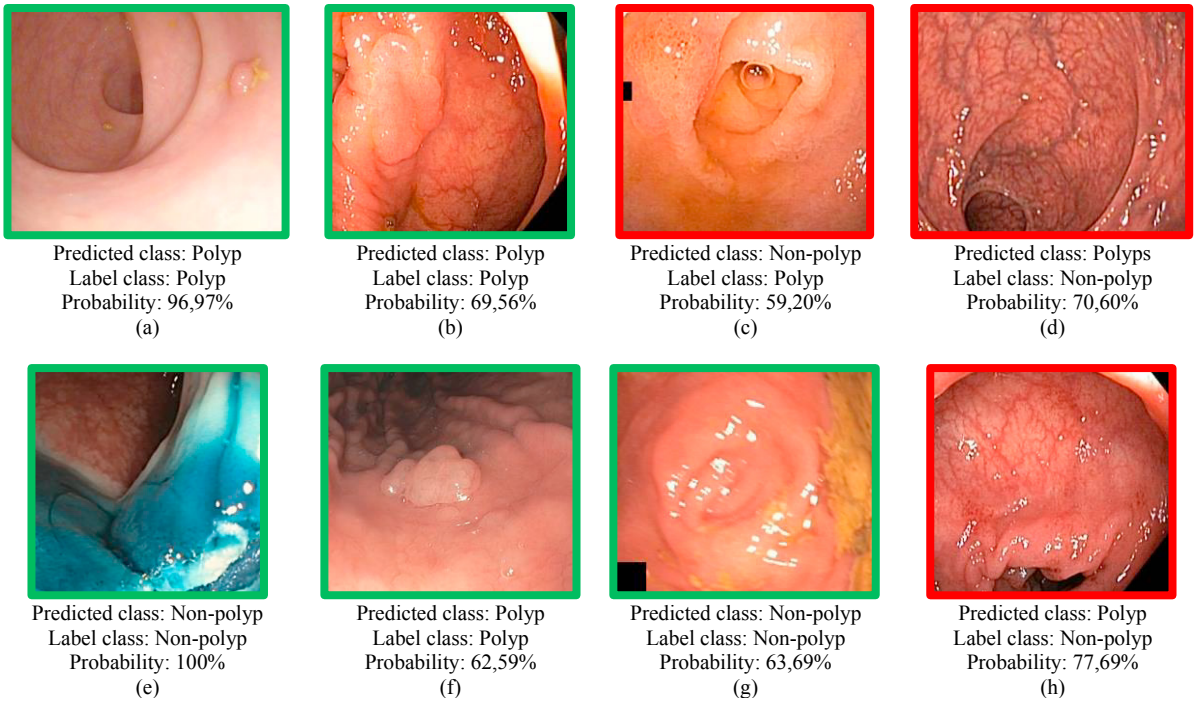

Fig. 5. Images and respective predictions made by the ResNet-50 [14] model.

In Fig.5, observing the two FP (Fig.5 (d) and (h)) produced by the network, it is possible to see some normal mucosa textures similar to polyps in both images. Hence, this situation could lead the model to fail these predictions.

Regarding the TP present in Fig.5 (a), the polyp stands out in the image, and despite the polyp size, the model could accurately detect it. Fig.5 (b) and (f) are also TP, but the model did not have much confidence in the prediction, which could be justified by the fact that both polyps are inserted in normal intestinal structures with some volume that could resemble polyps.

Observing the TN in Fig.5 (e), it is clear that the image does not present any polyp and the normal colonic mucosa does not present any structure, such as folds that could induce the model in error, which justifies the confidence of the model in this prediction. Besides, the dyer present in it could also be associated with the confidence in the prediction since the dyer is present in images representative of dyed dissection margins included in the non-polyp class.

For last, two situations where the ResNet-50 [14] did not have much confidence in the predictions made, will be analyzed. Just like most FP produced by both models, in Fig.5 (c) it is possible to observe an intestinal fold that presents volume, which could induce the model in error. Regarding Fig. 5 (g), although the model correctly classified the image as non-polyp, the motion blur could justify the low confidence in the prediction.

The most challenging aspect faced during the development of this work was the number of polyp representative images. However, the ResNet-50 [14] and ResNet-152 [14] models achieved results similar to the state-of-the-art [6,9]. These results could be improved by using a bigger and more diverse dataset that could show the different models, polyps of different sizes and shapes.

Additionally, by analyzing the obtained metrics (Table 2 and Table 3) it is possible to observe that no matter the depth of the ResNet architecture selected, the results are nearly the same but with slightly better performance in the ones with more depth: ResNet-50 [14] and ResNet-152 [14]. When all the models are compared, the model with the smallest depth, VGG16-bn achieved similar results when compared to ResNet-152 that was the model with the highest depth. Hence, the computational cost required by the bigger models does not justify their use since these architectures only produce a minimal increase in the results.

Observing the polyp detection articles analyzed in [18], the results achieved by all the models in this work are similar to the ones produced by the state-of-the-art models considered by the review article.

4. Conclusion

In this study, are presented different models to detect polyps in colonoscopies and an average accuracy of 95,70% was achieved by all the models. This study tested different layer depths and architectures, aiming to observe the effect produced in the model's performance and polyp detection capabilities. Observing the results, it is possible to see that in the different state-of-the-art architectures tested, the results were similar. Hence, architectures with a smaller depth, that used smaller computational resources should be used.

Regarding ResNet-50 [14] and ResNet-152 [14], after a more detailed analysis of the results in the test images, it was possible to observe that both models mistook some normal intestinal structures, such as folds, as polyps since these structures also have a volume that stands out in of remaining parts of the image. Hence, a strategy to reduce the FP would be to include in the dataset more images of this situation, allowing the model to better learn how to distinguish polyps from folds. In future work, post-hoc explainable artificial intelligence techniques can be used in these models to observe which image information is being considered to make predictions.

The use of these DL techniques in medical exams is one of the best ways to improve the adenoma detection rate (ADR) [6] since it has become more accessible through the years thanks to investments in frameworks like TensorFlow and fast.ai.

References

[1] Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., ... & Halvorsen, P. (2017, June). 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).

[2] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43, 99-111.

[3] Angermann, Q., Histace, A., Romain, O., Dray, X., Pinna, A., & Granado, B. (2015). Smart videocapsule for early diagnosis of colorectal cancer: toward embedded image analysis. In *Computational Intelligence in Digital and Network Designs and Applications* (pp. 325-350). Springer, Cham.

[4] Angermann, Q., Bernal, J., Sánchez-Montes, C., Hammami, M., Fernández-Esparrach, G., Dray, X., ... & Histace, A. (2017). Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures* (pp. 29-41). Springer, Cham.

[5] Bernal, J., Histace, A., Masana, M., Angermann, Q., Sánchez-Montes, C., Rodriguez, C., ... & Sanchez, J. (2018, June). Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In *Proceedings of 32nd CARS conference.*

[6] Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., & Baldi, P. (2018). Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4), 1069-1078.

[7] Ahmed, F. E. (2005). Artificial neural networks for diagnosis and survival prediction in colon cancer. *Molecular cancer*, 4(1), 1-12.

[8] Haggar, F. A., & Boushey, R. P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4), 191.

[9] Bardhi, O., Sierra-Sosa, D., Garcia-Zapirain, B., & Bujanda, L. (2021). Deep Learning Models for Colorectal Polyps. *Information*, 12(6), 245.

[10] Marks, J. W. (2019, December 17). *8 Colon Polyps Symptoms, Pictures, Types, Causes, Treatment.* MedicineNet. https://www.medicinenet.com/colon_polyps/article.htm

[11] Howard, J., & Thomas, R. (n.d.). fast.ai · Making neural nets uncool again. https://www.fast.ai/

[12] (n.d.). TensorFlow. https://www.tensorflow.org/

[13] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 *IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the *IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[15] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861.*

[16] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

[17] Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820.*

[18] Sánchez-Peralta, L. F., Bote-Curiel, L., Picón, A., Sánchez-Margallo, F. M., & Pagador, J. B. (2020). Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Artificial intelligence in medicine, 101923.