Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 469–476

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

Abnormality classification in small datasets of capsule endoscopy images

Filipe Fonseca[a,b], Beatriz Nunes[a], Marta Salgado[c], António Cunha[a,b]*

[a]*Universidade de Trás-os-Montes e Alto Douro, Vila Real Vila Real 5000-801, Portugal*
[b]*INESC TEC, Porto 4200-465, Portugal*
[c]*Centro Hospitalar Universitário do Porto, Porto 4099-001, Portugal*

Abstract

Capsule endoscopy made it possible to observe the inner lumen of the small bowel, but with the cost of a longer duration to process its resulting videos. Therefore, the scientific community has developed several machine learning strategies to help in detecting abnormalities in these videos. The published algorithms are typically trained and evaluated on small sets of images, ultimately not proving to be efficient when applied to full videos. In this experiment, we explored the problem of abnormality classification within an unbalanced dataset of images extracted from video capsule endoscopies, based on a vector feature extracted from the deepest layer of pre-trained Convolution Neural Networks to evaluate the impact of transfer learning with a small number of samples. The results showed that there is a reliable model on the classification task using small portions of data from video capsule endoscopies.

Peer-review under responsibility of the scientific committee of the 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

Keywords: Image classification;capsule endoscopy; medical imaging; deep learning; transfer learning.

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

Peer-review under responsibility of the scientific committee of the 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
10.1016/j.procs.2021.12.038

1. Introduction

Introduced in 2000, the wireless capsule endoscopy (WCE) is a pill-like, non-invasive camera that can be swallowed by the patient to be propelled through the digestive system taking advantage of its peristaltic movements, to provide an inner screening of the gastrointestinal tract [1]. It is considered the preferred method for the diagnosis of diseases of the small bowel, given its ability to cover the whole gastrointestinal tract, including this area of difficult access and its notable inner imaging results compared to other methods for abnormalities visualisation. Despite this, considering the resulting eight-hour-long recorded video, the main disadvantage with this procedure resides in a long (40 to 60 minutes), monotonous and susceptible to human error, reviewing of the screening by medical experts [2]. Thus, the interest in developing techniques for the automatic detection, classification and diagnosis of lesions to save time, and aid gastroenterologists in the examining of the video capsule endoscopies (VCE).

The computer-aided diagnosis for pathology is arguably the final frontier of vision-based abnormalities detection and disease diagnosis. However, like any other technology, digital pathology comes with its own challenges: whole-scan imaging generally generates large number of files that also require digital storage and are not easy to analyse via computer algorithms. Detection, segmentation and image classification, appears to be a quite daunting task for computer vision algorithms.

Looking at deep learning and its vast possibilities for classification seems to be a good path to address the above-mentioned obstacles of digital pathology. Diverse deep architectures have been trained with a large set of images, e.g., ImageNet project, to perform difficult tasks like object classification [3]-[6]. The results have been more than impressive. Accuracy numbers in the mid and high 90s have become quite common when deep networks, trained with millions of images, are tested to recognize unseen samples [7]. Despite all progress, one can observe that the applications of deep learning in digital pathology have still a lot of space to grow. The major obstacle appears to be the lack of large labelled datasets to properly train some type of multi-layer neural networks, a requirement that may still be missing for some years to come. Hence, start designing and training deep nets with the available datasets is the only way.

In this paper, we evaluate the performance of CNN (Convolutional Neural Networks) pre-trained models on non-medical imaging data [4]. Specifically, when used as feature extractors without fine tuning for an abnormality classification task, considering a small portion of images from video capsule endoscopies (VCE).

2. Methodology

The pipeline used for this experiment is described in Fig.1.

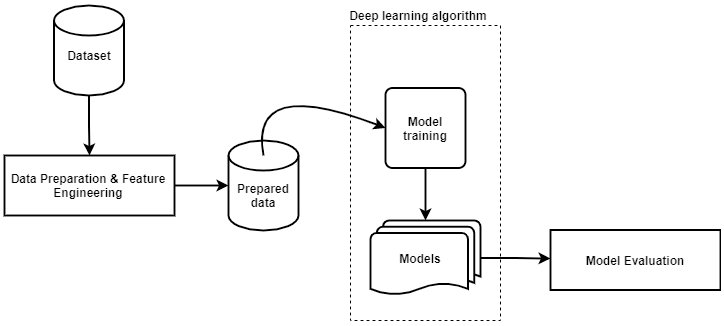


Fig. 1. The pipeline of the experiment.

Data from a public dataset was pre-processed and organised into train and validation sets, for models training and test aiming a model final evaluation.

2.1. Dataset

The dataset used in this experiment was the Kvasir-Capsule, a video capsule endoscopy dataset collected from examinations at a Norwegian Hospital. In total, the dataset consists of 4,741,621 main data records, i.e., 47,238 images with labels, the 43 corresponding labelled videos (the videos from which the images are extracted) and 74 unlabelled videos. The 47,238 labelled images were stored using the PNG format and organized into two main categories: anatomical landmarks characterising the gastrointestinal tract and the content of the small bowel lumen (the aspect of the mucosa and mucosal lesions - pathological findings) [8].

For the experiment depicted in this paper, a small portion of images were used from the labelled classes: Normal Clean Mucosa, Angiectasia, Blood-Fresh and Polyp. Fig. 2 depicts example images from the Kvasir-Capsule dataset for each labelled class used in this experiment.

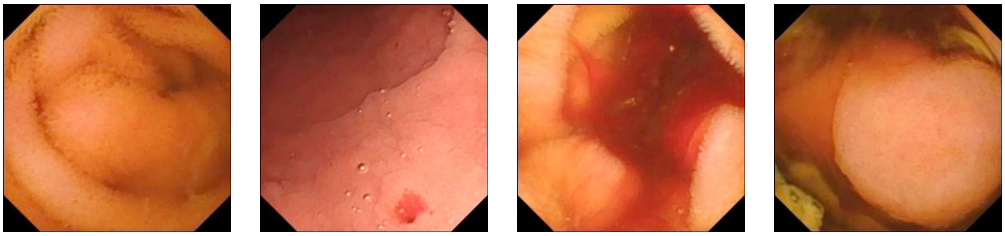


Fig. 2. (a) Normal Clean Mucosa; (b) Angiectasia; (c) Blood-Fresh; (d) Polyp.

2.2. Data Preparation & Feature Engineering

To evaluate the impact of transfer learning with a small set of samples, we have researched for datasets that, in addition to being properly structured and labelled, didn't require prior treatment of the images to be used in this experiment. In this research, we have come across the Kvasir-Capsule dataset which, besides being a relatively recent dataset, has a variety of differentiated and labelled data that deserves to be duly explored. From the Kvasir-Capsule dataset, 866 images were selected belonging to the class Angiectasia, 446 belonging to the class Blood-Fresh and belonging to the class Polyp, 55 images. For each selected image was identified the corresponding labelled video from which the images were extracted and then, distinctly selected the 11,677 images from the class Normal Clean Mucosa that had matched the identified labelled videos.

Afterwards, all selected images were divided into two folders: “normal” (containing the 11,677 images from the class Normal Clean Mucosa) and “not_normal” (contained the 1,367 images of the 3 previously mentioned classes). Keeping the same division of “normal” and “not_normal”, the data was randomly divided into three unbalanced sets: 70% for training, 15% for validation and 15% for test. Considering that images from a labelled video cannot belong to the train and test sets at the same time, it was required an adjustment in the division of percentage values.

All images were in a resolution of 336 x 336 pixels and for that reason, no adjustments were made in their dimension. On the other hand, given that the dataset was unbalanced and to contradict the negative effects that this would cause on training, some adjustments were performed, i.e. data augmentation [9].

The data augmentation allows to artificially expand the size of the training dataset by creating modified versions of images, that already belong to a dataset. The transformations can include shifts, flips, zoom, and other image manipulations [10]. To cope with the unbalanced sets, it was chosen Keras’ ImageDataGenerator library to conduct data augmentation. Apart from data augmentation, we have used the focal loss function for binary classification since it is relevant for problems with unbalanced data.

2.3. Deep Learning Algorithm

The CNNs models show excellent performance on machine learning problems, especially when these problems include the classification of images' dataset [4]. Models need to be deep enough to be able to capture resemblance like texture, colour and shape features between image samples: while initial layers of the model gather those high-level features, later ones capture information that relates those features with the outputs, learning to discriminate between them. Unfortunately, in many cases, the amount of available data is not sufficient to adopt an approach like this. Transfer learning solves this problem by applying models previously trained on a large available dataset (usually trained for a completely different task, using the same input but returning different output), in a novel task. This is achieved by a previous capture of the relations in the features of that data that can then be reused for other problems.

Due to the aforementioned success of transfer learning in image classification, namely in the medical area, an approach using this technique was followed for the abnormality's classification task in images from VCE. In transfer learning approaches, reusable features returned by some layers will be used as input features that allow the training of a new model that only needs to learn the relations of those features for the new problem, given that it has already learnt about data patterns in the data that was used to pre-train the model. Besides requiring much fewer parameters to learn, transfer learning also has the advantage of better generalisation of the models, given that they underlay the phenomenon more than they model the data, owing to the fact that the model has access to different types of data.

For this experiment, within state-of-the-art CNNs models we have chosen the EfficientNetB3, Xception, ResNet50. The Xception and the ResNet50 since they were on the top of Keras' ranking[1], taking into consideration the model's performance on the ImageNet validation dataset. These networks have the following dimensions: EfficientNetB3 has 12,320,535 parameters; Xception has 22,910,480; and ResNet50 has 25,636,712.

Each pre-trained model was trained using two different approaches: only the last layer was updated, i.e. the pre-trained model worked as a feature extractor and only the weights of the classification layers were changed. Each model was trained for 100 epochs, with a batch size of 200, and using the Adam optimization algorithm which is an extension to the stochastic gradient descendent method that is based on adaptive estimation of the first order and second-order moments. Momentum was set to 0.9 and learning rate to 0.00001.

2.4. Model Evaluation

We have used the standard classification metrics like Precision equation (1), Recall or Sensitivity equation (2) and Specificity equation (3), commonly used for classification problems and can then be computed with a resort to the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). The area under the Receiver Operating Characteristic (ROC) curve is another evaluation metric commonly used to depict the performance of each classification model for all possible classification thresholds, and will also be computed in this work.

$$Precision = \frac{TP}{TP + FP} \quad (1) \qquad Recall = \frac{TP}{TP + FN} \quad (2) \qquad Specificity = \frac{TN}{TN + FP} \quad (3)$$

Additionally, we have also used the F1-score equation (4) to better compare recall and precision between different models. The metric F1 is especially important in this experiment given the unbalanced data used.

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

[1] More information: https://keras.io/api/applications/.

3. Results and discussion

Independently, 3 versions of pre-trained CNNs model: EfficientNetB3, ResNet50 and Xception on ImageNet datasets were trained on our prepared dataset of VCE images from the Kvasir-Capsule dataset, with the goal of transferring the information into abnormalities classification task that had limited training data. The results of transfer learning using the 3 mentioned models are illustrated in Table 1 by common classification metrics, precision, recall, and F1-score. Taking into account that for a model to be considered as a satisfactory model, precision and recall should be as high as possible. The F1-score helps to better compare recall and precision and on a perfect model is close to 1. With that in mind, ResNet50 model in this experiment was the most reliable among the other pre-trained CNNs model, to perform the given task within limited data.

Table 1. Precision, recall and F1-score obtained by the trained models.

Model	Class	Precision	Recall	F1-score
ResNet50	0	0.97	0.99	0.98
	1	0.99	0.69	0.81
Xception	0	0.97	0.95	0.96
	1	0.61	0.74	0.67
EfficientNetB3	0	0.93	0.50	0.66
	1	0.13	0.75	0.23

A careful analysis of Table 1 shows that models with the highest number of layers do not always achieve the best performance. The model with the best performance in classifying images as being without abnormalities, meaning belonging to the class Normal Clean Mucosa was the ResNet50, with a precision of 0.97 and recall of 0.99, in contrast to the model with the worst performance, the EfficientNetB3 which had a precision of 0.93 and recall of 0.50. Considering the classification of images with abnormalities - either belonging to the classes Angiectasia, Blood-Fresh or Polyp (True Negatives), the ResNet50 model was again the best performing model given the F1-score of 0.81. It is debatable that given the recall values of 0.74 and 0.75, respectively for Xception and EfficientNetB3, that these models are better than ResNet50 when classifying abnormalities, but considering the dimension of the data in this experiment, the margin in between models is not that high.

For better visualisation of the analysis above, Fig. 3 shows two examples of images correctly classified with the corresponding predict values from each model.

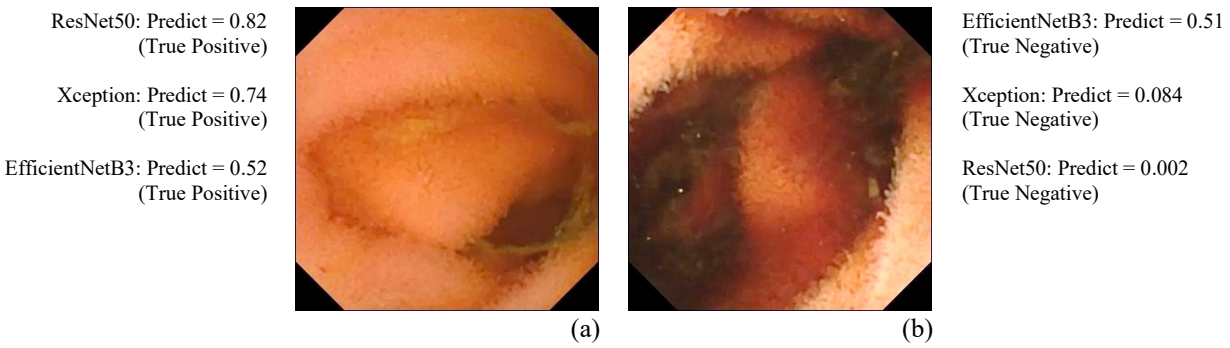


Fig. 3. (a) Normal Clean Mucosa; (b) Blood-Fresh. Images correctly classified with predict values for each model.

A close analysis of the following confusion matrix shown in Table 2, we confirmed that for binary classification of images into normal and abnormality classes, the performance of the model ResNet50 was, by far, the best one regarding the classification of the images to belong to the normal class. Regarding the classification of images as

abnormalities and given the specificity values of 0.74 and 0.75, respectively for Xception and EfficientNetB3, they had a slight better performance when compared to the ResNet50.

Table 2. The confusion matrix (TP, true positives; FN, false negatives; FP, false positives; TN, true negatives), the sensitivity and the specificity are provided along with the AUC for each model.

Model	TP	FN	FP	TN	Sensitivity	Specificity	AUC
ResNet50	963	1	30	67	0.99	0.69	0.95
Xception	917	47	25	72	0.95	0.74	0.90
EfficientNetB3	486	478	24	73	0.50	0.75	0.65

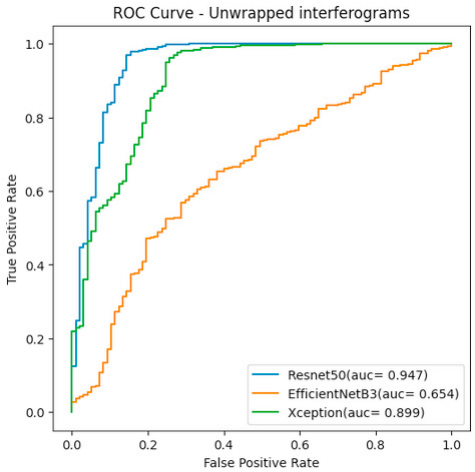
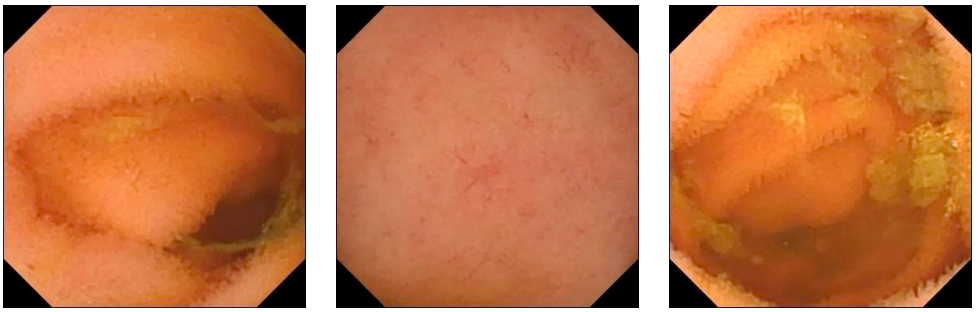


Fig. 4. ROC curve for each classification model.

The following sequence of images intended to help visualise the results described in Table 1. Fig. 5 and Fig. 6 show the result of the ResNet50 classification for normal and abnormal images, respectively, with the predict values. In contrast, Fig. 7 and Fig. 8 show the results of EfficientNetB3 classification for normal and abnormal images, respectively, with the predict values.



ResNet50: Predict = 0.82
(True Positive)
(a)

ResNet50: Predict = 0.82
(True Positive)
(b)

ResNet50: Predict = 0.47
(False Negative)
(c)

Fig. 5. ResNet50 classification of normal images: (a), (b) and (c) Normal Clean Mucosa.

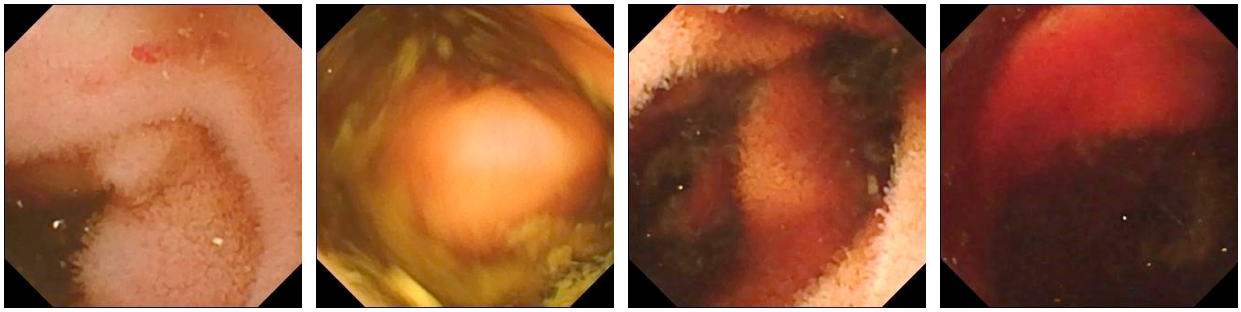


Fig. 6. ResNet50 classification of images not normal: (a) Angiectasia; (b) Polyp; (c) Blood-Fresh; (d) Blood-Fresh.

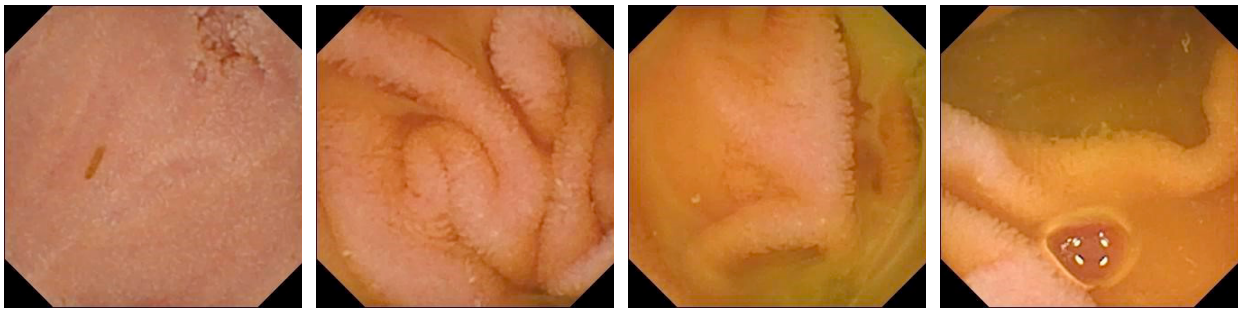


Fig. 7. EfficientNetB3 classification of normal images: (a), (b), (c) and (d) Normal Clean Mucosa.

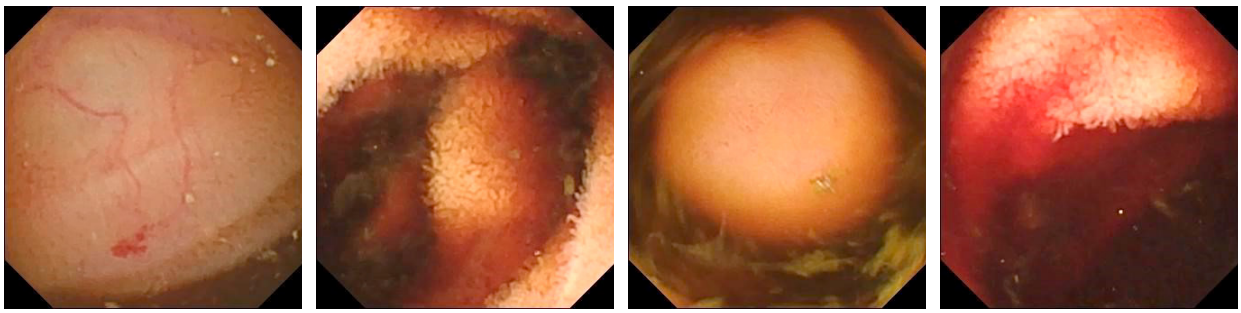


Fig. 8. EfficientNetB3 classification of images not normal: (a) Angiectasia; (b) Blood-Fresh; (c) Polyp; (d) Blood-Fresh.

4. Conclusions and future work

We have performed an experiment to evaluate the effects of transfer learning structures with 3 pre-trained CNNs models, which synergistically integrates transfer learning strategies for VCE images classification. It was observed that deep transfer learning architecture like ResNet50 presented satisfactory results when classifying VCE images into normal (Normal Clean Mucosa) or not normal, among 3 different abnormalities (Angiectasia, Blood-Fresh and Polyp) within small and unbalanced dataset.

The result of this experiment provided insightful indications regarding the usage of which model to be considered, in future work, when understanding the major differences between computer-aided diagnosis systems that use and don't use active learning within a small and evolving labelled dataset.

Acknowledgements

This work is financed by national funds through the portuguese funding agency, *FCT - Fundação para a Ciência e a Tecnologia*, within project UIDB/50014/2020.

References

[1] Iddan G., Meron G., Glukhovsky A. and Swain P, (2000). "Wireless capsule endoscopy". Nature, vol. 405, no. 6785, p. 417.
[2] Koulaouzidis A., Iakovidis D. K., Karargyris A. and Plevris J. N., (2015). "Optimizing lesion detection in small-bowel capsule endoscopy: from present problems to future solutions". Expert Review of Gastroenterology and Hepatology, 9:2, 217–235.
[3] Pan S. J. and Yang Q. (2010). "A survey on transfer learning". IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345– 1359.
[4] Shin H. C., Roth H. R., Gao M., et al (2016). "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning". IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285–1298.
[5] Deng J., Dong W., *et al* (2009), "Imagenet: A large-scale hierarchical image database". *Computer Vision and Pattern Recognition*, *CVPR 2009. IEEE Conference on. IEEE*, pp. 248–255.
[6] Girshick R., Donahue J., Darrell T. and Malik J. (2016). "Region-based convolutional networks for accurate object detection and segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158.
[7] Yi D., Sawyer R., *et al* (2017). "Optimizing and visualizing deep learning for benign/malignant classification in breast tumors." *CoRR*, vol. abs/1705.06362.
[8] Smedsrud P.H., Thambawita V., Hicks S.A., *et al* (2021). "Kvasir-Capsule, a video capsule endoscopy dataset". *Scientific Data*, 8:142. https://doi.org/10.1038/s41597-021-00920-z.
[9] Esteva A., Robicquet A., Ramsundar B., *et al* (2019). "A guide to deep learning in healthcare". *Nat Med*, 25:24.
[10] Soffer S., Ben-Cohen A., Shimon O., *et al* (2019). "Convolutional neural networks for radiologic images: a radiologist's guide". Radiology, 290: 590-606.