Automatic image and text-based description for colorectal polyps using BASIC classification

Artificial Intelligence In Medicine 121 (2021) 102178

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Automatic image and text-based description for colorectal polyps using BASIC classification

Roger Fonollà [a,*], Quirine E.W. van der Zander [b,c], Ramon M. Schreuder [d], Sharmila Subramaniam [f], Pradeep Bhandari [f], Ad A.M. Masclee [b,e], Erik J. Schoon [d], Fons van der Sommen [a], Peter H.N. de With [a]

[a] Department of Electrical Engineering, Video Coding and Architectures (VCA), Eindhoven University of Technology, Eindhoven, Noord-Brabant, the Netherlands
[b] Division of Gastroenterology and Hepatology, Maastricht University Medical Center, Maastricht, the Netherlands
[c] GROW, School for Oncology and Developmental Biology, Maastricht University, Maastricht, the Netherlands
[d] Department of Gastroenterology and Hepatology, Catharina Hospital, Eindhoven, Noord-Brabant, the Netherlands
[e] NUTRIM, School of Nutrition & Translational Research in Metabolism, Maastricht University, Maastricht, the Netherlands
[f] Department of Gastroenterology, Portsmouth Hospitals University NHS Trust, Portsmouth, United Kingdom

ARTICLE INFO

Keywords:
Blue light imaging
Linked color imaging
BASIC
Image captioning
Artificial intelligence
Deep learning
CADx

ABSTRACT

Colorectal polyps (CRP) are precursor lesions of colorectal cancer (CRC). Correct identification of CRPs during in-vivo colonoscopy is supported by the endoscopist's expertise and medical classification models. A recent developed classification model is the Blue light imaging Adenoma Serrated International Classification (BASIC) which describes the differences between non-neoplastic and neoplastic lesions acquired with blue light imaging (BLI). Computer-aided detection (CADe) and diagnosis (CADx) systems are efficient at visually assisting with medical decisions but fall short at translating decisions into relevant clinical information. The communication between machine and medical expert is of crucial importance to improve diagnosis of CRP during in-vivo procedures. In this work, the combination of a polyp image classification model and a language model is proposed to develop a CADx system that automatically generates text comparable to the human language employed by endoscopists. The developed system generates equivalent sentences as the human-reference and describes CRP images acquired with white light (WL), blue light imaging (BLI) and linked color imaging (LCI). An image feature encoder and a BERT module are employed to build the AI model and an external test set is used to evaluate the results and compute the linguistic metrics. The experimental results show the construction of complete sentences with an established metric scores of BLEU-1 = 0.67, ROUGE-L = 0.83 and METEOR = 0.50. The developed CADx system for automatic CRP image captioning facilitates future advances towards automatic reporting and may help reduce time-consuming histology assessment.

1. Introduction

Colorectal polyps (CRP) are precursor lesions and indicators of colorectal cancer (CRC). CRPs are roughly divided between benign CRPs, which include the hyperplastic polyps (HPs), and pre-malignant CRPs, comprising adenomas (ADs) and the sessile serrated adenomas (SSAs). HPs are the most common polyp type found during colonoscopy and are usually considered benign. In contrast, ADs and SSAs are capable of developing into CRC when kept untreated [1]. Current medical protocols dictate that all detected CRPs should be resected to undergo histological evaluation, but this protocol has two considerable drawbacks, since (1) unnecessary removal of benign polyps exposes the patient to additional risks of polypectomy-related complications, and (2) histological examination of all resected polyps leads to significantly increased costs. To minimize both the cost and risk, the strategy of *resect-and-discard* has been proposed for diminutive (≤5 mm) adenomatous polyps [2–4] and the *diagnose-and-leave* strategy for diminutive hyperplastic polyps [5,6] in the left colon.

Optical diagnosis of CRPs is often supported by clinical classification models. A commonly used scheme is the Paris classification [7,8] based on endoscopic appearance and morphology of observed CRPs during colonoscopy. White light endoscopy (WL) is the most common technique

[*] Corresponding author.
E-mail address: r.fonolla.navarro@tue.nl (R. Fonollà).

https://doi.org/10.1016/j.artmed.2021.102178
Received 29 April 2021; Received in revised form 1 September 2021; Accepted 21 September 2021
Available online 28 September 2021

to visually assess lesions in the colon, but it less capable of enhancing the visualization of surface and vessel patterns. The injection of chemical dyes into the colon, also referred to as chromoendoscopy [9], achieves higher contrast results than WL imaging. The Kudo pit pattern [10] classification takes advantage of improved visualization through magnification and staining, to classify polyps according to the pit appearance, structure, and staining patterns. Despite the benefits over WL, the use of stains requires the injection of chemicals inside the intestinal tissue, which is often time-consuming. Similar visual effects can be achieved with the use of in-vivo optical filters, like Narrow-Band Imaging (NBI) (Olympus) [11,12]. The incorporation of NBI to colonoscopy procedures has introduced the need of clinical models for in-vivo characterization, such as the NICE classification [13,14]. This clinical model has been developed for differentiation of HPs, ADs and deep submucosal invasive cancer on non-magnifying images and, the JNET classification [15] for magnifying images. The Workgroup serrAted polypS and Polyposis proposed the WASP [16] classification as a modification of the NICE criteria to facilitate the differentiation of SSAs. Alternatively, enhancement of endoscopic images can be achieved by means of post-processing technologies. Flexible spectral imaging color enhancement (FICE) (Pentax) and I-Scan digital contrast are two existing technologies that improve the contrast, sharpening and spectrum of the images. For the latter imaging mode, the ICE classification [17] was proposed for diagnosis of non-adenomatous and adenomatous CRPs. Lastly, laser technology, powered by the four-LED Multi Light Technology (Fujifilm Co.) provides an innovative approach to visualize the intestinal tissue. This technology is based on the combination of four types of light as source emitters: blue-violet, blue, green and red. Blue Light Imaging (BLI) and Linked Color Imaging (LCI) are two of the observation modes of the four-LED Multi Light Technology, that allow enhanced visualization of haemoglobin. BLI intensifies the blue-light emission in the range of 410 nm, which enhances the visualization of the vessels and the mucosa. Alternatively, LCI accentuates color contrast by decreasing the blue-light intensity while emphasizing the red-light signal, providing better delineation and detection of lesions and inflammations [18]. To take advantage of the visual improvements of BLI, the Blue light imaging Adenoma Serrated International Classification (BASIC) [19] was created to help clinicians in visually differentiating non-neoplastic and neoplastic lesions.

The broad spectrum of new acquisition systems coupled with the wide range of clinical models for in-vivo classification of CRPs, introduces new challenges for image and text interpretation. Computer-aided detection (CADe) and diagnosis (CADx) systems facilitate medical decisions during clinical assessment, providing additional information for medical diagnosis. Furthermore, the incorporation of CADx systems into clinical practice can also reduce the learning curve of new acquisition modalities, by offering visual and textual clues for treatment decisions.

The success of machine learning, and more recently deep learning, has advanced the development of CADx systems to classify CRPs. For the NBI imaging mode, artificial intelligence techniques have been applied for the Hiroshima classification [20], the Kudo classification [21], and more predominantly, the NICE classification [22–24]. Alternatively, other studies have solely focused on neoplastic tissue differentiation, developing CADx based on off-the-shelf deep learning architectures [25–27]. The aforementioned CADx studies have focused on the potential of NBI applications, whereas BLI and LCI have not yet seen significant developments. In Scheeve et al. [28] a CADx pipeline was proposed for the WASP classification using handcrafted features directly obtained from existing medical knowledge. More recently, in Weight et al. [29] a CADe and CADx system was developed to detect and characterize neoplasia using a dataset of WL, BLI, and LCI polyps, after which the results were compared with non-experts and experts endoscopists. The outcomes of the study showed that the CADx system can help to improve the classification made by less experienced endoscopists for identification and classification of CRPs.

The growth in CADx systems poses new challenges for image interpretation and providing automated explanations of decision-making systems (explainable AI). The integration of new systems into clinical procedures, call for a better explainable understanding of artificial intelligence to endoscopists [30–32]. The addition of saliency maps such as GRAD-CAM in CRP workflows [33–35] allows the clinicians to visualize the decisions of the CADx and refine the diagnosis with medical expertise. Although graphic maps can improve the diagnosis, textual descriptions could greatly facilitate understanding of CADx and allow for a more precise decision-making. In the field of deep learning, image captioning is the term for automatically generating a textual description from an image. For medical applications, image captioning has been applied mainly to radiology [36–38], thanks to publicly available datasets [39,40]. In Mishra et al. [41], an image captioning pipeline was proposed for retinal diseases, while the study of Rojas-Muñoz et al. [42] reported a CADx system to guide physicians during surgical procedures, by providing medical instructions automatically generated from surgery images.

In our previous studies [43], a CADx system for BLI and LCI modalities was proposed to differentiate between HP and SSA/ADs polyps, where the results of the study were compared with a total of 19 endoscopists knowledgeable with the BASIC classification, achieving an accuracy of 95.0%, sensitivity of 93.3%, and specificity of 95.6% on a test set of 60 CRPs previously evaluated by experts and novices alike. The development of a CADx system to classify and differentiate between polyp malignancy improves the diagnosis of CRPs, but the absence of explanatory guidance in the system for in-vivo diagnosis of CRPs diminishes the potential value between AI and gastroenterologist. Such guidance for CADx systems is called explainable AI (further used as a term in this paper), which is further pursued as a concept to add trust value to an automated CADx system.

In this work, we present a CADx system, that incorporates an image captioning block based on the Bidirectional Encoder Representations from Transformers (BERT) language model [44]. The presented image captioning CADx system utilizes the foundations of our previous developed CADx [43] to transfer the knowledge of the learned polyp features. Our system is capable of providing an automated description of individual polyps in WL, BLI, and LCI modalities, according to the BASIC classification. The inclusion of automatically generated captions could provide a better diagnosis to gastroenterologists and further differentiation of HP, SSAs and ADs during in-vivo procedures. Our deep learning model is evaluated on the same test set as our previous studies, thereby preserving the consistency of our data. The results are evaluated using the BLEU (Bilingual Evaluation Understudy) [45], the ROUGE (Recall Oriented Understudy for Gisting Evaluation) [46], and the METEOR (Metric for Evaluation for Translation with Explicit Ordering) [47] score, based on each key point of the BASIC classification 1.

The contributions of our work are threefold. (1) A CADx system that automatically generates text descriptions related to the BASIC classification from a single polyp image. (2) A solution that is modality-independent and accepts the information of WL, BLI and LCI images. (3) A system capable of providing text-based suggestions to experts and novices alike, allowing for further development towards a more powerful explainable AI and automated report generation. The developed extension of the CADx system adds clarity to contribute to a smooth decision-making.

The remainder of this paper is outlined as follows. In Section 2, the acquisition and analysis of the image and textual data are explained, followed by the implementation of the CADx system. Section 3 introduces the results in a structured table, followed by Section 4, where the findings and future work of AI and endoscopy are discussed. Section 5 wraps up the contributions and benefits of the study in current endoscopy routine.

Table 1
The BASIC classification comprises a list of visual BLI features, employed by endoscopists to classify hyperplastic, sessile serrated and adenomatous polyps. A more in-depth clinical analysis can be found in the study of Subramaniam et al. [48]. In this Figure the cancer category is not present, although it is part of the official classification.

		Hyperplastic	Adenoma	Sessile serrated
Surface	*Presence of mucus*	No	No	Yes
	Regular/irregular	Regular	Regular/irregular	Regular/irregular
	Pseudodepression	No	Yes	No
	Depression	No	No	No
Pit pattern	*Featureless*	Yes	No	No
	Type	Round	Not round	Round pits with/wo dark spots
	Distribution	Homogeneous	Homo/heterogeneous without focal loss	Homo/heterogeneous
Vessels	*Present*	Yes/no	Yes	yes/no
	Type if present	Lacy	Pericryptal	Pericryptal

2. Methodology

2.1. Data acquisition

The data collection was carried out in a prospective fashion, according to a pre-defined image acquisition protocol, in the Maastricht University Medical Center (MUMC), Catharina Hospital Eindhoven (CZE), both in the Netherlands, and the Queen Alexandra Hospital in Porthsmouth, United Kingdom. The training dataset includes polyps acquired in WL, BLI, LCI and I-Scan (HDWL, Modes 1, 2 and 3) modalities. A total of 468 patients were included, of which only 95 contained textual descriptions of the polyps. The data collected for the test set was obtained from a prospective, endoscopist-blinded, non-interventional study, conducted both at the MUMC and CZE. The study was in accordance with the declaration of Helsinki as well as the General Data Protection Regulation. A total of 19 endoscopists optically diagnosed 60 colonoscopy images, containing a single polyp acquired in WL, BLI modalities (later referred to as test data). Two person groups were derived from the medical professionals. The first group consisted of six expert endoscopists from the international BLI-expert group, who were knowledgeable in using BLI and BLI Adenoma Serrated International Classification (BASIC) [19,48] (Table 1) and had an experience of more than 2000 colonoscopies. The second group consisted of thirteen Dutch novices with limited colonoscopy experience (less than 400 colonoscopies) without prior experience in using BLI or BASIC. More extended information of the study can be found at van der Zander et al. [35]. All collected data was fully anonymized prior to the study. Out of all the collected training data, 95 patients had additional textual descriptions of the diagnosed polyps, with the potential to be used as ground truth for training the explainable AI system. The textual descriptions were provided by expert endoscopists following the BASIC classification, and were supplementary information connected to each polyp image. Each polyp lesion was acquired at least with WL and BLI modalities with some patients containing more than one lesion. The testing set consisted of an additional 60 patients which also included the same textual classification descriptions. Each polyp lesion was uniquely acquired with WL, BLI and the LCI modalities and selected according to good image quality and availability of the corresponding histology results (gold standard). Each polyp image was described by up to 9 BASIC terms, based on the second-left column of Table 1.

2.2. Data preprocessing

2.2.1. Image data

In order to obtain optimal analysis, the central region of the image was automatically selected as the ROI. The cropped region ensures a coverage of the polyp area, as well as its surrounding texture. The dataset was sequentially normalized by subtracting the mean and by dividing the standard deviation of the pre-trained ImageNet data. As last step, each input image was resized to 299 × 299 pixels in the RGB color space. To increase the generalization of the network, data augmentation was used to enhance the model capabilities. In this study, the training images are augmented by a combination of flipping, shifting and ±90° rotation, contrast enhancement, blurring and scaling.

2.2.2. Text data

For the analysis of the textual data, each polyp image was characterized by the descriptors grouped in four distinct blocks. The first block is associated with the Paris classification (morphology and size) and the three remaining to the BASIC classification (surface, pit pattern and vessels), as shown in Fig. 1. For each descriptor, a standardized sentence was constructed resulting in a total of four sentences for each polyp sequence. Each set of descriptors was carefully analyzed to correct for any inconsistency with the BASIC classification. If more than one block presented a disagreement then the whole sentence was discarded, otherwise only the erroneous block was removed. The blocks that remained, were considered as 'gold standard' for the text analyses. Each of the polyp sequence was restricted to a maximum length of 45 tokens or words, including the starting word ['CLS'] and the ending word ['SEP'] which are used in the BERT dictionary. In case of a sequence containing a lesser amount of tokens, the special word ['PAD'] was used to reach the maximum desired length. Each sentence within the sequence was separated by the ending word ['SEP'].

After a thorough analysis of all polyp sequences, a total of 95 patients were selected for the training set, comprising a total of 6525 polyp sequences and 507 images. The test set consisted of 55 patients with a total

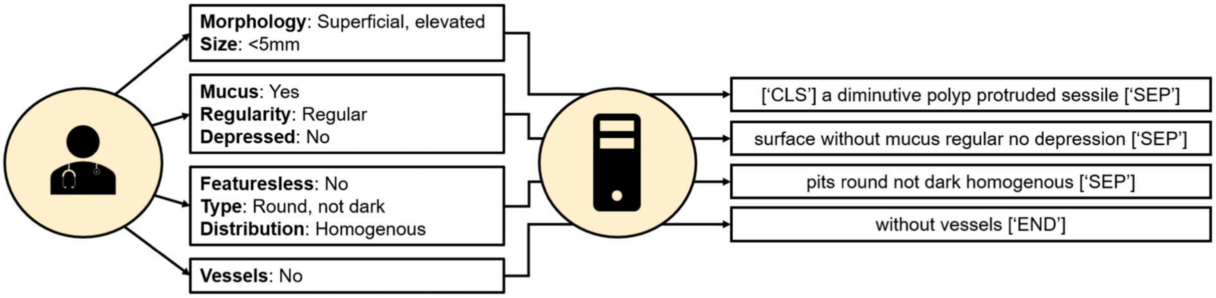

Morphology: Superficial, elevated
Size: <5mm
Mucus: Yes
Regularity: Regular
Depressed: No
Featureless: No
Type: Round, not dark
Distribution: Homogenous
Vessels: No
['CLS'] a diminutive polyp protruded sessile ['SEP']
surface without mucus regular no depression ['SEP']
pits round not dark homogenous ['SEP']
without vessels ['END']

Fig. 1. Example of a polyp sequence created from individual polyp descriptors.

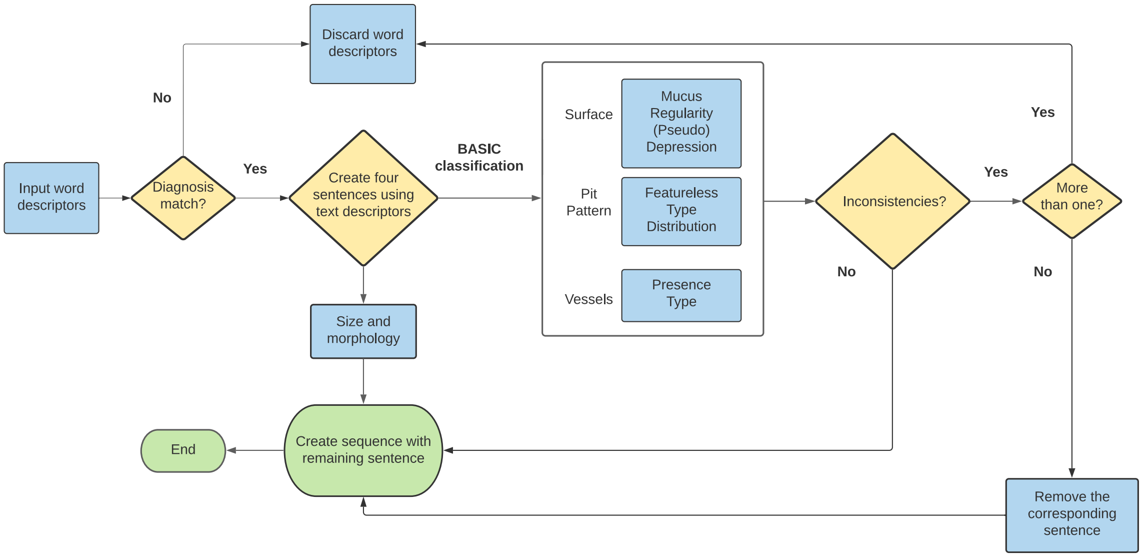


Fig. 2. Data cleaning process of polyp sentences to remove inconsistencies between diagnosis and ground truth (BASIC classification).

of 1857 sequences and 165 images (note that 5 patients were removed due to data inconsistency). A more detailed diagram can be found in Fig. 2.

2.3. CADx captioning model

The model employed in this study is divided in two distinct blocks: (1) the polyp image module which serves as encoder and image feature extractor, and (2) the BERT Module which allows for the learning of the polyp sequences. The main stages are depicted in Fig. 3.

2.3.1. Polyp image module for visual information

The image encoding section of the model receives a single image as input of any of the modalities present in the training set (WL, BLI and LCI). The employed architecture consists of the base network EfficientNet-B4 [49], where the classification layers are removed and replaced by a global average pooling layer (GAP). This leads to an $N \times 1$ vector where N is the size of the GAP layer.

2.3.2. BERT module for textual information

The learning of the polyp sequences is achieved through the language model. A pre-trained BERT transformer is used to obtain the language features. The main task of this model is to learn and predict the next word in the input sequence. The module receives a matrix of polyp sequences, where the first sequence contains the start token, the subsequent sequence is the combination of the previous sequence plus the next sequence word, and so forth, until the end of the polyp sequence is reached. This leads to an $L \times M$, where L is the max sequence and M the size of the BERT dictionary.

2.3.3. Combining visual and textual information

To merge the contents of both networks, a concatenation layer is added to combine the polyp image module and the BERT module. This operation is depicted in Fig. 3 with the *Concatenate* box. On one hand, the output of the BERT module is dictated by the maximum length sequence and the total amount of tokens in the original trained dictionary. On the other hand, the output of the polyp image module is only a feature vector dictated by the last convolutional layer of the image model. Since these outputs are intrinsically different, the smaller one (GAP layer) is redistributed to become aligned with the language model and facilitate smooth integration. To adequately implement the combination of the output of both modules, the GAP layer is repeated L times to match it with the maximum length of the text sequence resulting in a $L \times M \times N$ matrix. Following the concatenation layer, the output is supplied to a Long Short-Term Memory (LSTM) to capture the temporal relation between words. The model concludes with a dense layer of the size of the BERT dictionary where each prediction yields a probability for each word in the dictionary.

2.4. Training

The model is trained in two separate stages. The first stage

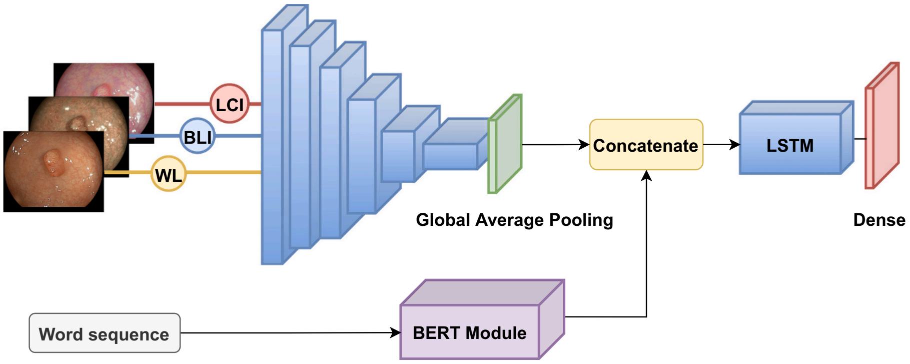
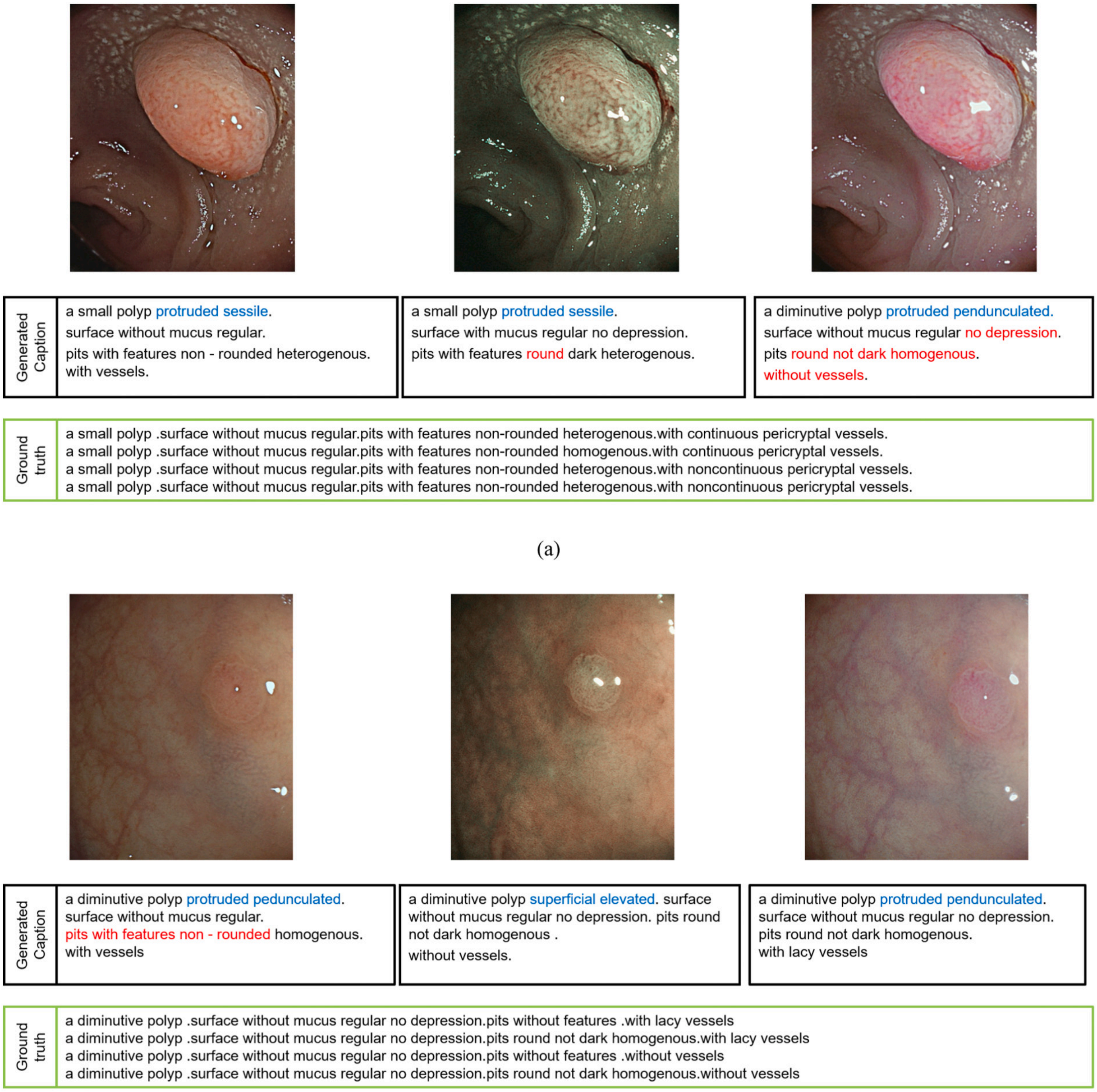


Fig. 3. End-to-end framework of the proposed CADx system. The training batch contains independent images of each modality (WL, BLI and LCI) and a polyp sequence which describes the associated image. The output of the system is a probability for each word in the text dictionary.



Generated Caption	a small polyp protruded sessile. surface without mucus regular. pits with features non - rounded heterogenous. with vessels.	a small polyp protruded sessile. surface with mucus regular no depression. pits with features round dark heterogenous.	a diminutive polyp protruded pendunculated. surface without mucus regular no depression. pits round not dark homogenous. without vessels.
Ground truth	a small polyp .surface without mucus regular.pits with features non-rounded heterogenous.with continuous pericryptal vessels. a small polyp .surface without mucus regular.pits with features non-rounded homogenous.with continuous pericryptal vessels. a small polyp .surface without mucus regular.pits with features non-rounded heterogenous.with noncontinuous pericryptal vessels. a small polyp .surface without mucus regular.pits with features non-rounded heterogenous.with noncontinuous pericryptal vessels.		

(a)

Generated Caption	a diminutive polyp protruded pedunculated. surface without mucus regular. pits with features non - rounded homogenous. with vessels	a diminutive polyp superficial elevated. surface without mucus regular no depression. pits round not dark homogenous . without vessels.	a diminutive polyp protruded pendunculated. surface without mucus regular no depression. pits round not dark homogenous. with lacy vessels
Ground truth	a diminutive polyp .surface without mucus regular no depression.pits without features .with lacy vessels a diminutive polyp .surface without mucus regular no depression.pits round not dark homogenous.with lacy vessels a diminutive polyp .surface without mucus regular no depression.pits without features .without vessels a diminutive polyp .surface without mucus regular no depression.pits round not dark homogenous.without vessels		

(b)

Fig. 4. Example of three generated captions for (a) an AD polyp and (b) a HP polyp acquired in WL, BLI and LCI. The black box shows the generated captions while the green box contains the human-reference generated from the descriptors of endoscopists. In the one hand, the words highlighted in red represent the generated text wrongly identified in the human-reference. On the other hand the blue words show the generated text not present in the ground truth captions, but which belongs to the training corpus. The remaining words in black are correctly identified in the human-reference polyp sentences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

individually trains the polyp image module, which was originally designed in our previous study [43] where a polyp classification model was trained to differentiate between benign and pre-malignant polyps. The model was initialized with Imagenet weights and trained with Adam optimizer with batch size of 8 and using an exponential learning rate, with hard restarts at every two epochs, ranging from 1e − 2 to 4e − 3. The model was trained for 100 epochs or until convergence was reached on the validation set, using a Titan Xp GPU. For this study, all the previous learned weights are transferred to the polyp image module and all the layers are frozen until the GAP layer. A transfer learning approach is used for the BERT module, where the pre-trained weights are loaded into the language model and all the layers are kept frozen until the concatenation layer. The combined models are again trained with an Adam optimizer using a learning rate of 1e − 3 and a batch size of 5. The model is trained for 3 epochs using a Xeon E5–1660 v4 with clock frequency of 3.20 GHz and 16-core CPU to replace the Titan Xp GPU, as its performance was not sufficient for our experiments. As a result, the total training took 5 days of continuous computing.

3. Evaluation and results

The proposed model is evaluated on a dataset of 55 patients that were imaged with WL, BLI and LCI. Therefore, for each patient, three images of a single polyp were acquired. Each lesion is associated with one or more sentences, describing the size, morphology and the BASIC classification. Measuring the quality of generated captions is a

Table 2
Evaluation results in the unity interval for the automatically generated captions of the test set. The first two rows present the results for the entirety of the sentence and for the sequence containing the BASIC descriptors (which excludes morphology and size). The last four rows belong to the evaluation of each individual block from a polyp sequence.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Complete sentence	0.67	0.54	0.46	0.36	0.80	0.58	0.83	0.50
BASIC descriptors	0.50	0.38	0.27	0.09	0.66	0.43	0.70	0.55
Morphology & size	0.52	0.34	0.29	0.00	0.87	0.62	0.89	0.69
Surface	0.89	0.87	0.85	0.77	0.91	0.86	0.92	0.96
Pit pattern	0.52	0.40	0.34	0.05	0.83	0.72	0.85	0.76
Vessels	0.43	0.13	0.02	0.00	0.53	0.11	0.59	0.45

challenging task, because there is no right, or simple way to measure the correctness of one sentence to another. Nevertheless, three metrics are employed to assess the results of the proposed algorithm, in order to compare the generated descriptions with the human-based reference. All the employed metrics are based on the calculation of the n-gram sequence, where a unigram is defined as a sequence of one word, a bigram as a pair of words, and so forth. In the evaluation we employ the following metrics.

(1) The BLEU score [45] is a measure of precision, which allows for the calculation of segments of words (several n-grams) to evaluate the amount of words in the generated caption appearing in the human references. We compute the BLEU score at four different n-grams (BLEU-1, BLEU-2, BLEU-3 and BLEU-4).
(2) The ROUGE score [46] is an alternative to the BLEU score and is a measure of recall, where the calculated metric is based on the amount of segment words in the human references that appear in the generated captions. In addition to ROUGE-1 and ROUGE-2, a commonly used ROUGE score is the ROUGE-L (Longest Common Subsequence), where the longest co-occurring sequence is calculated between the reference and the prediction.
(3) METEOR is a less common score [47] but designed to address the weaknesses in the BLEU score. METEOR evaluates a sentence by calculating a score based on the explicit unigrams between the generated sentence and all the human-reference sentences.

The evaluation is performed using the entirety of the generated caption as well as the information of each individual sentence. As such, the scores are individually computed for the morphology and size, the surface, the pit pattern and the vessel. Additionally, the three BASIC descriptors are evaluated as a single sequence as well. Although some generated sentences may be too short to be evaluated based on n-grams, we have decided nonetheless to calculate every descriptor to observe its outcome. As final remark, each generated description is associated with more than one human-reference sentence, hence the reported numbers are obtained from the mean of all available comparisons between the generated captions and the references. Summarizing, all the metrics are evaluated at the level of a complete sentence, at the level of the BASIC text descriptors and at a level of an individual descriptor.

4. Discussion

In-vivo classification of polyps is a challenge for medical doctors during a live colonoscopy. To aid clinicians, current CADe and CADx systems focus on the detection or classification of the observed polyps. Advances towards a more explainable CADx are required for improved diagnosis of CRC [50,51]. In this work, we propose a system capable of automatically generating informative captions based on the BASIC classification. Fig. 4 illustrates visual and textual examples of an HP polyp and an AD polyp, as well as its generated captions for each of three acquired modalities.

4.1. General performance

The performance of the system is shown in Table 2. From the observed results, the BLEU score for all experiments is marginally lower than the ROUGE score. The observed trend may indicate that the majority of the test corpus or dictionary is being represented in the generated captions. The similarity and standardization of the morphology and BASIC descriptors between the training and the test set does contribute to the homogeneity on the generated captions. On the contrary, an intermediate BLEU score (about 0.50) may indicate that not enough words from the test dictionary are being generated by the proposed system. This can be associated to the difficulties of BLEU score at evaluating complex words, such as those found in the morphology and pit pattern description. Similar results are observed in the BASIC subdivision, where an identical trend is shown all across the scores. Larger n-grams, both in BLEU and ROUGE, are not suited to evaluate short sequences, such as found in the sub-divisions of morphology and the individual BASIC descriptors. A more significant evaluation can be extracted from ROUGE-L and METEOR. On one hand, ROUGE-L automatically obtains the longest sequence of n-grams between reference and generated caption, so that it can give insights to the word occurrence or shorter sentences of varying size. On the other hand, METEOR calculates the harmonic mean or F_1-score, based solely on unigram precision. Moreover and different from the other metrics, METEOR can be calculated from multiple references and compared to one generated sentence accordingly.

4.2. Individual performance and textual inconsistencies

From the observations of individual descriptors, the surface outperforms the rest of the descriptors, both in ROUGE-L (0.92) and METEOR (0.96). This observation can be explained by the homogeneity on surface descriptors and the simplicity of its wording. Less rewarding results are observed for the vessels descriptor, where the complexity of the sentence as well as the sparsity of the words in the training vocabulary may have affected its performance. Out of all descriptors, the vessels were the sequences that were most affected by the harsh preprocessing, which diminished the amount of useful training sentences compared to the rest of the corpus. This loss is caused by regular occurrences of inconsistencies between the diagnosis and the BASIC classification (ground truth), which lead to discarding word descriptions (see Fig. 2). For example, an HP polyp cannot contain pericryptal vessels and an AD/SSA cannot contain lacy vessels according to the BASIC classification, see Table 1, where such occurrence leads to an inconsistency in diagnosis.

4.3. Quality of the generated text

The quality of the generated sentences is quite comparable with the ground truth. To illustrate the quality, some visual and textual results are presented in Fig. 4. The incorrectly generated words which are not present in the human-sentence references, are indicated in red color. A remark that was encountered during the inspection of test human-references and the training captions is that the latter did contain more

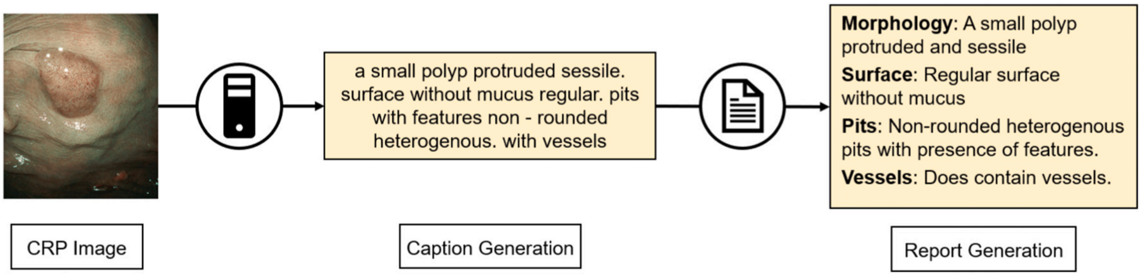

a small polyp protruded sessile. surface without mucus regular. pits with features non - rounded heterogenous. with vessels
Morphology: A small polyp protruded and sessile
Surface: Regular surface without mucus
Pits: Non-rounded heterogenous pits with presence of features.
Vessels: Does contain vessels.
CRP Image
Caption Generation
Report Generation

Fig. 5. Visual representation of the proposed workflow of our CADx system. From an input polyp image obtained from the endoscope until an automated report generated from the automatic caption system.

information and deeper description of the size and morphology. Hence, the generated captions from the test set do contain extra information not present in the references (denoted in color blue in Fig. 4). The question arises whether the generation of extra words is helpful or not for the clinical application. This aspect considered differently by the metrics. For example, a BLEU will degrade by this extra information, whereas ROUGE score will positively affected by this aspect. As an illustration, we have computed the corresponding scores of the top experiment (AD) in Fig. 4, which results for white light in (BLEU-1 0.70 vs ROUGE-L 0.88), BLI (0.62 vs 0.67) and LCI (0.47 vs 0.61). The HP example follows the same trend where white light scores (BLEU-1 0.59 vs ROUGE-L 0.76), BLI (0.75 vs 0.88) and LCI (0.74 vs 0.92). This result confirms our statement, but is only a single example and more study would further assess this aspect. The assessment of the text is based on the BERT capacity to generate text, further studies should focus on other modules such as GPT family or variants of BERT such as ALBERT or RoBERTa.

4.4. Human interfacing

Our system is trained with sentences generated from individual polyp descriptors. In order to ease the training process, a standardization of all the sentences is applied. Although this approach helps on reducing the complexity of the sentences, it does otherwise and imposes a burden on the semantic understanding during in-vivo colonoscopy. In Fig. 5, a graphic representation is shown of a plausible post-processing step to transfer the computer-generated captions into sentences with a semantic meaning, with the aim to improve the diagnosis of any observed polyp. The presented example in Fig. 5 poses an interesting structure of reporting the outcome to the endoscopist, but a much broader test would be needed to measure the acceptance and usage in the workflow of the medical experts. The proposed system was developed with static images captured under strict high quality imaging conditions, however in clinical practice video endoscopy is employed and such image quality cannot be guaranteed. Hence future work should focus a clear criteria for the captured data and the incorporation on real-time video.

4.5. Integration of polyp characterization

The descriptions generated by our system do not contemplate the characterization of the polyp label. In image captioning task such as the specific subject of the image is usually incorporated in the description. In our case, the word "polyp" is present in all descriptions, but not the more accurate classification such as "HP", "AD" and "SSA". A decision of excluding the polyp classification was made with the aim to focus the training more on the polyp descriptors, rather than the polyp type. As future work towards a better polyp image description, this aspect could be incorporated and classified with an already existing CADx system and supplemented with our polyp captioning system. Further studies should focus on including the polyp type into the generated captions with emphasis on checking if the classification results are in accordance with the polyp description generated from the BASIC classification.

In this work, multiple human-sentences were used to compare the results of the generated sentences. During the pre-processing steps, all the references that did not comply with the diagnosis were excluded, in order to reduce the amount of errors produced by endoscopists. The lack of a concise and unique description for each polyp implies a burden when comparing the generated captions to the human-reference, which leads to various interpretations. Therefore, the results of this study are still in a research stage and should be taken as supportive material for clinical decision rather than an absolute truth, and is advisable to employ the generated captions as a side tool together with the clinician's own expertise.

5. Conclusions

In this study, we have presented a CADx system for automatic caption generation for colonoscopy images obtained with three different image modalities (WL, BLI and LCI). From a single input image, our system automatically generates a textual polyp description, based on the BASIC classification. The proposed CADx system is trained with a dataset of 6525 polyp sequences correlated to 507 polyp images and evaluated on an independent test set of 55 patients with 165 polyp images and 1857 human-reference sequences. The model demonstrates a good performance at generating sentences which are comparable to the human references. Besides its optimal performance, one of the downsides we have found during training is that one single GPU is not sufficient for training such a complex model, which restricted the training to only the usage of a CPU. The lack of variability on the number of patients with available textual data also restricted the diversity of the generated polyp sequences. Further studies should aim at collecting a broader dataset to enhance the qualities of both the generating model and the human dictionary. Our study opens the possibility towards future automatic report generation during in-vivo classification of colorectal polyps. The combination of existing detection and classification systems with the proposed system could potentially improve the diagnosis of polyps and facilitate the learning curve of the BASIC classification for experts and novice endoscopists. Overall, the presented study can facilitate the diagnosis of colorectal polyps in two ways. First, the presented system may improve the cooperation and trust between a CADx system and gastroenterologist by providing an automatic analysis and reporting of colorectal polyps. Second, the system may decrease the burden and involved cost of histological examinations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 721766. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2021.102178.

References

[1] Jass JR. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 2017;50(1):113–30.

[2] Ignjatovic A, East JE, Suzuki N, Vance M, Guenther T, Saunders BP. Optical diagnosis of small colorectal polyps at routine colonoscopy (detect inspect characterise resect and discard; discard trial): a prospective cohort study. Lancet Oncol 2009;10(12):1171–8.

[3] Hassan C, Pickhardt PJ, Rex DKA. Resect and discard strategy would improve cost-effectiveness of colorectal cancer screening. Clin Gastroenterol Hepatol 2010;8 (10):865–869.e3.

[4] S. Tsuji, Y. Takeda, K. Tsuji, N. Yoshida, K. Takemura, S. Yamada, H. Doyama, Clinical outcomes of the "resect and discard" strategy using magnifying narrow-band imaging for small (<10 mm) colorectal polyps, Endosc Int Open 6 (12) (2018) E1382–E1389.

[5] Neumann H, Sen H Neumann, Vieth M, Bisschops R, Thieringer F, Rahman K, et al. Leaving colorectal polyps in place can be achieved with high accuracy using blue light imaging (BLI). Unit Eur Gastroenterol J 2018;6(7):1099–105.

[6] Kandel PW, Wallace MB. Should we resect and discard low risk diminutive colon polyps. Clin Endosc 2019;52(3):239.

[7] The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. Gastrointest Endosc 2003;58(6):S3–43.

[8] Endoscopic classification review group update on the Paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy 2005;37(6):570–8.

[9] Har-Noy O, Katz L, Avni T, Battat R, Bessissow T, Yung DE, et al. Chromoendoscopy narrow-band imaging or white light endoscopy for neoplasia detection in inflammatory bowel diseases. Dig Dis Sci 2017;62(11):2982–90.

[10] Kudo S, Tamura S, Nakajima T, Yamano H, Kusaka H, Watanabe H. Diagnosis of colorectal tumorous lesions by magnifying endoscopy. Gastrointest Endosc 1996; 44(1):8–14.

[11] East JE, Guenther T, Kennedy RH, Saunders BP. Narrow band imaging avoids potential chromoendoscopy risks. Gut 2007;56(8):1168–9.

[12] Vişovan II, Tanțău M, Pascu O, Ciobanu L, Tanțău A. The role of narrow band imaging in colorectal polyp detection. Bosn J Basic Med Sci 2017;17(2):152.

[13] Hayashi N, Tanaka S, Hewett D, Kaltenbach T, Sano Y, Ponchon T, et al. Endoscopic prediction of deep submucosal invasive carcinoma: validation of the narrow-band imaging international colorectal endoscopic (NICE) classification. Gastrointest Endosc 2013;78(4):625–32.

[14] Iwatate M, Hirata D, Sano Y. NBI international colorectal endoscopic (NICE) classification. In: Endoscopy in early gastrointestinal cancers. vol. 1. Springer; 2018. p. 69–74.

[15] Sano Y, Tanaka S, Kudo S, Saito S, Matsuda T, Wada Y, et al. Narrow-band imaging (NBI) magnifying endoscopic classification of colorectal tumors proposed by the Japan NBI expert team. Dig Endosc 2016;28(5):526–33.

[16] IJspeert J, Bastiaansen B, van Leerdam M, Meijer G, van Eeden S, Sanduleanu S, et al. Dutch workgroup serrated polyps & polyposis (wasp). development and validation of the wasp classification system for optical diagnosis of adenomas, hyperplastic polyps and sessile serrated adenomas/polyps. Gut 2015;65(6):963–70.

[17] Bouwens MW, de Ridder R, Masclee AA, Driessen A, Riedl RG, Winkens B, et al. Optical diagnosis of colorectal polyps using high-definition i-scan: an educational experience. World J Gastroenterol 2013;19(27):4334–43.

[18] Yoshida N, Dohi O, Inoue K, Yasuda R, Murakami T, Hirose R, et al. Blue laser imaging and linked color imaging for the detection and characterization of colorectal tumors. Gut Liver 2019;13(2):140–8.

[19] Bisschops R, Hassan C, Bhandari P, Coron E, Neumann H, Pech O, et al. Basic (bli adenoma serrated international classification) classification for colorectal polyp characterization with blue light imaging. Endoscopy 2018;50:211–20.

[20] Kominami Y, Yoshida S, Tanaka S, Sanomura Y, Hirakawa T, Raytchev B, et al. Computer-aided, diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. Gastrointest Endosc 2016;83:643–9.

[21] Mori Y, Kudo SE, Misawa M, Saito Y, Ikematsu H, Hotta K, et al. Use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann Intern Med 2018;169:357–66.

[22] Chen PJ, Lin MC, Lai MJ, Lin JC, Lu HHT, Accurate VS. Classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology 2018;154: 568–75.

[23] Zhang X, Chen F, Yu T, An J, Huang Z, Liu J, et al. Real-time gastric polyp detection using convolutional neural networks. PloS One 2019;14(3):e0214133.

[24] Byrne MF, Chapados N, Soudan F, Oertel C, Linares-Pérez M, Kelly R, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94–100.

[25] Komeda Y, Handa H, Watanabe T, Nomura T, Kitahashi M, Sakurai T, et al. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology 2017;93(Suppl. 1):30–4.

[26] Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 2018;155(4):1069–78.

[27] Rodriguez-Diaz E, Baffy G, Lo W-K, Mashimo H, Vidyarthi G, Mohapatra SS, et al. Real-time artificial intelligence-based histologic classification of colorectal polyps with augmented visualization. Gastrointest Endosc 2021;93(3):662–70.

[28] Scheeve T, Schreuder RM, van der Sommen F, IJspeert JE, Dekker E, Schoon EJ. Computer-aided classification of colorectal polyps using blue-light and linked-color imaging. In: SPIE medical imaging 2019: computer-aided diagnosis 2019 (10950); 2019. p. 12.

[29] J. Weigt, A. Repici, G. Antonelli, A. Afifi, L. Kliegis, L. Correale, C. Hassan, H. Neumann, Performance of a new integrated cade/cadx system for detection and characterization of colorectal neoplasia. preprint, Endoscopy.

[30] van der Sommen F, de Groof J, Struyvenberg M, van der Putten J, Boers T, Fockens K, et al. Machine learning in gi endoscopy: practical guidance in how to interpret a novel field. Gut 2020;69(11):2035–45.

[31] Le Berre C, Sandborn WJ, Aridhi S, Devignes M-D, Fournier L, Smaïl-Tabbone M, et al. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology 2020;158(1):76–94.

[32] R. Pannala, K. Krishnan, J. Melson, M. A. Parsi, A. R. Schulman, S. Sullivan, G. Trikudanathan, A. J. Trindade, R. R. Watson, J. T. Maple, et al., Artificial intelligence in gastrointestinal endoscopy, VideoGIE.

[33] Jin EH, Lee D, Bae JH, Kang HY, Kwak M-S, Seo JY, et al. Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. Gastroenterology 2020;158(8):2169–79.

[34] Zhou D, Tian F, Tian X, Sun L, Huang X, Zhao F, et al. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. Nat Commun 2020; 11(1):1–9.

[35] Q. E. van der Zander, R. M. Schreuder, R. Fonollà, T. Scheeve, F. van der Sommen, B. Winkens, P. Aepli, B. Hayee, A. Pischel, M. Stefanovic, et al., Optical diagnosis of colorectal polyp images using a newly developed computer-aided diagnosis system (cadx) compared to intuitive optical diagnosis, Endoscopy (AAM).

[36] Shin H, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM. Learning to read chest X-rays: recurrent neural cascade model for automated image annotation. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR); 2016. p. 2497–506.

[37] Kisilev P, Sason E, Barkan E, Hashoul SY. Medical image captioning: learning to describe medical image findings using multitask-loss CNN. 2016.

[38] I. Allaouzi, M. B. ahmed, B. Benamrou, M. Ouardouz, Automatic caption generation for medical images, Proceedings of the 3rd international conference on smart city applications.

[39] Demner-Fushman D, Antani S, Simpson MS, Thoma G. Design and development of a multimodal biomedical information retrieval system. J Comput Sci Eng 2012;6: 168–77.

[40] Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology objects in context (roco): a multimodal image dataset. In: Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis. Springer; 2018. p. 180–9.

[41] Mishra S, Banerjee M. Automatic caption generation of retinal diseases with self-trained rnn merge model. In: Advanced computing and systems for security; 2020. p. 1–10.

[42] Rojas-Muñoz E, Couperus K, Wachs JP. The ai-medic: an artificial intelligent mentor for trauma surgery, computer methods in biomechanics and biomedical engineering. Imaging Visual. 2020;0(0):1–9.

[43] Fonolla R, Sommen FVD, Schreuder RM, Schoon EJ, De With PHN. Multi-modal classification of polyp malignancy using cnn features with balanced class augmentation. In: Proceedings - international symposium on biomedical imaging; 2019. p. 74–8.

[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 181004805.

[45] Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, association for computational linguistics; 2002. p. 311–8.

[46] Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out. Association for Computational Linguistics; 2004. p. 74–81.

[47] Lavie A, Agarwal A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the second workshop on statistical machine translation, StatMT '07. Association for Computational Linguistics; 2007. p. 228–31.

[48] Subramaniam S, Hayee B, Aepli P, Schoon E, Stefanovic M, Kandiah K, et al. Optical diagnosis of colorectal polyps with blue light imaging using a new international classification. United Eur Gastroenterol J 2019;7(2):316–25.

[49] Tan M, Le Q. Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th international conference on machine learning; 2019. pp. 97, 6105–6114.

[50] Byrne MF, Chapados N, Soudan F, Oertel C, Pérez M Linares, Kelly R, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94–100.

[51] Shung DL, Byrne MF. How artificial intelligence will impact colonoscopy and colorectal screening. Gastrointest Endosc Clin 2020;30(3):585–95.