Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy

Thomas de Lange, Pål Halvorsen, Michael Riegler

Thomas de Lange, Department of Transplantation, Oslo University Hospital, Oslo 0424, Norway

Thomas de Lange, Institute of Clinical Medicine, University of Oslo, Oslo 0316, Norway

Pål Halvorsen, Michael Riegler, Center for Digital Engineering Simula Metropolitan, Fornebu 1364, Norway

Pål Halvorsen, Michael Riegler, Department for Informatics, University of Oslo, Oslo 0316, Norway

ORCID number: Thomas de Lange (0000-0003-3989-7487); Pål Halvorsen (0000-0003-2073-7029); Michael Riegler (0000-0002-3153-2064).

Author contributions: de Lange T, Halvorsen P and Riegler M contributed to the concept and design of the editorial, drafting of the manuscript and final approval of the manuscript.

Supported by the grants from Norwegian Research Council, No. 282315.

Conflict-of-interest statement: de Lange T declares no conflict of interests; Halvorsen P reports grants from Norwegian Research Council during the conduct of the study; and grants from Norwegian Research Council outside the submitted work; Riegler M declares no conflict of interests.

Open-Access: This article is an open-access article which was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

Manuscript source: Invited manuscript

Telephone: +47-22118080

Received: September 7, 2018
Peer-review started: September 7, 2018
First decision: October 4, 2018
Revised: October 25, 2018
Accepted: November 2, 2018
Article in press: November 2, 2018
Published online: December 7, 2018

Abstract

Assisted diagnosis using artificial intelligence has been a holy grail in medical research for many years, and recent developments in computer hardware have enabled the narrower area of machine learning to equip clinicians with potentially useful tools for computer assisted diagnosis (CAD) systems. However, training and assessing a computer's ability to diagnose like a human are complex tasks, and successful outcomes depend on various factors. We have focused our work on gastrointestinal (GI) endoscopy because it is a cornerstone for diagnosis and treatment of diseases of the GI tract. About 2.8 million luminal GI (esophageal, stomach, colorectal) cancers are detected globally every year, and although substantial technical improvements in endoscopes have been made over the last 10-15 years, a major limitation of endoscopic examinations remains operator variation. This translates into a substantial inter-observer variation in the detection and assessment of mucosal lesions, causing among other things an average polyp miss-rate of 20% in the colon and thus the subsequent development of a number of post-colonoscopy colorectal cancers. CAD systems might eliminate this variation and lead to more accurate diagnoses. In this editorial, we point out some of the current challenges in the development of efficient computer-based digital assistants. We give examples of proposed tools using various techniques, identify current challenges, and give suggestions for the development and assessment of future CAD systems.

Key words: Endoscopy; Artificial intelligence; Deep
learning; Computer assisted diagnosis; Gastrointestinal

Core tip: Assisted diagnosis using artificial intelligence and recent developments in computer hardware have enabled the narrower area of machine learning to equip the endoscopists with potentially powerful tools for computer assisted diagnosis systems. The success depends on various factors; optimizing algorithms, image database quality and size and comparison with existing systems.

INTRODUCTION

Gastrointestinal (GI) endoscopy is a cornerstone for diagnosis and treatment of diseases in the GI tract. About 2.8 million luminal GI cancers (esophageal, stomach, colorectal) are detected globally every year, and many of these might be prevented through improved endoscopic performance and systematic high-quality screening in high incidence areas[1]. These cancers represent a substantial health challenge for society with a mortality rate of about 65%[2], and colorectal cancer is the third most common cause of cancer mortality among both women and men[3]. Despite substantial technical improvements in endoscopes over the last 10-15 years, a major limitation of endoscopic examinations is operator variation. This variation depends on operator skill, perceptual factors, personality characteristics, knowledge, and attitude[4]. This translates into a substantial inter-observer variation in the detection and assessment of mucosal lesions[5,6], leading to an average polyp miss-rate of 20% in the colon[7]. All of these factors can to some extent be alleviated by substantial educational efforts, but they cannot be eliminated entirely[8]. Thus, developing an automated computer-based support system for the detection and characterization of mucosal lesions would be an important contribution to eliminating the current variation in endoscopists’ performance.

Artificial intelligence (AI) is the area of computer science that aims to create intelligent machines that mimic human behavior, and assisted diagnosis using AI has been a holy grail in the field of medicine for many years. Such machines have long been the realm of fiction, but recent developments in computer hardware have enabled the narrower field of machine learning to develop potentially highly accurate computer assisted diagnosis (CAD) systems. At its most basic, machine learning is the practice of using algorithms to parse data, learn from the data, and then make a prediction, and in the medical domain such systems are used to detect or classify a disease. Research and development of such systems is currently under way in many medical domains like retina scans, various cancer screening systems, and skin cancer detection[9-11]. However, there exist methodological issues that need to be addressed both for creating and improving automated diagnosis algorithms.

MACHINE LEARNING IN ENDOscopy

Automated detection of anomalies in the GI tract have been proposed for diseases such as Barrett’s esophagus, gastric cancer, angiectasia, celiac disease, and polyp detection and characterization, and a number of methods and algorithms have been tested in recent years[12-18]. The methods and algorithms range from simpler traditional machine learning methods to more recently developed deep learning approaches[19,20].

An example of a simple system is a search-based system using various global features in the images[21]. It extracts (complex) image features like color histograms and textures and feeds these features into a classifier for determining whether an object is present or not. For example, such a system might determine the presence of an object by calculating the distance of the feature vector from the vectors in the model. An important advantage of systems based on simple methods is that they can be easier to understand and their results can be easier to explain to medical personnel[22-24].

The current state-of-the-art and the most commonly used methods are based on deep neural networks. These networks work as an interconnected group of nodes, akin to the vast network of neurons in the human brain[25]. Such networks typically consist of an input and an output layer, as well as multiple hidden convolutional, pooling, fully connected, and normalization layers. Typically, each input image will pass through the layers in order to classify an object with probabilistic values between 0 and 1. There exist several variations of deep neural networks. For image and video analysis, convolutional neural networks (CNNs) are the most common. CNNs can be used to perform either segmentation (the exact marking of a finding in the image[26]) or classification (a more global point of view on the image, such as a general statement like “this image contains a polyp”[22,27,28]). Another promising method for image analysis is generative adversarial networks (GANs). GANs consist of two neural networks competing with each other in a zero-sum game framework during the training phase. The generator network generates new data instances using an inverse convolutional network by upsampling random noise to an image. The other network, the discriminator, takes the generated image and the training set and checks for authenticity. This means that the discriminator decides whether the data belong to or are
classified in the actual training dataset or not. GANs can also be defined as conditional GANs that have an image as input instead of random noise and that transform this image into another image. This can be used to create, for example, segmentation masks. An example of a GAN-based method is described by Pogorelov et al.\textsuperscript{[22,20]}. The approach presented in their papers uses conditional GANs with a normal image from the colon as input, and the algorithm segments the finding in the image. This noise segmentation is then cleaned in a post-processing step that leads to a clear segmentation. Many of these approaches have yielded promising results regarding detection accuracy, with some achieving numbers above 90%, but many run too slowly to be used in a clinically useful system providing real-time feedback. Some comparisons of different approaches are given by Pogorelov et al.\textsuperscript{[22,20]} and Riegler et al.\textsuperscript{[20]}.

**IMAGE DATABASE QUALITY**

A sufficient amount of data is vital in machine learning, and the creation of algorithms usually relies on large databases. This is especially true for deep learning, which is currently the standard for image analysis\textsuperscript{[31]}. However, the quality of the database is also essential, and it is crucial that all the images and videos are annotated correctly. The computer learns from analyzing the given data, and thus erroneous learning will lead to incorrect diagnoses. Therefore, when collecting data and making a dataset the recommendations below should be followed.

There are variations between observers, and to reduce this bias the ground truth assessment should involve at least three observers\textsuperscript{[32]}. However, the required agreement between the observers and the degree of confidence is not known and requires further studies. The goal regarding the diagnostic thresholds for such a technique is to reach more than 90% positive predictive value for correct classification of the lesions\textsuperscript{[33]}.

A potential problem in machine learning is overfitting. Many of the datasets show obvious examples of medical findings, and the similarity of the different images often results in overfitting. Thus, overfitting occurs when the learning algorithm learns the data too well and therefore also captures the noise of the data, e.g., when the model or the algorithm fits the data too well, or if the model or algorithm shows low bias but high variance. Therefore, too many similar samples should be avoided in order to avoid such “overtraining”. A diverse dataset is therefore recommended to better enable correct disease detection in new data.

Many datasets are limited in size, and many assess their systems using too few samples. Many argue that the dataset should be as large as possible\textsuperscript{[34]}, but others show that machine learning can also work on smaller datasets using transfer learning\textsuperscript{[35]}, which has recently found frequent use in the context of medical image problems\textsuperscript{[36,37]}. Note that there is no “one size fits all” answer. The amount of required training data is dependent on many different aspects of the experiment, but a general rule of thumb is to have around 1000 images per class for deep learning applications. In the Kvasir dataset\textsuperscript{[38]}, at least 1000 images per class are provided for different findings.

One general problem is that several of the existing datasets are cumbersome to use in terms of permission, for example, several of the listed sets in Table 1\textsuperscript{[38-44]} are restricted. To enable subsequent comparisons, it is best is to use an open dataset.

The most important take-away message is that clean and complete data are one of the most important parts of a good detection system. This means that spending the time to create a high-quality database is very important and is directly connected to the quality of the following steps.

**SYSTEM ASSESSMENT**

Comparing published research is challenging, and an increasing number of research communities are targeting this problem by creating public available datasets and encouraging reproducible experiments. In order to enable full comparisons, not only the same datasets should be used, but the datasets should also be split between training and test sets in an equal way. Furthermore, the more information the better, and one should use as many of the common metrics as possible as described by Pogorelov et al.\textsuperscript{[38]}. For detection accuracy, the raw numbers for true positives, true negatives, false positives, and false negatives are important, and metrics based on these like sensitivity (recall), precision, specificity, accuracy, Matthews correlation coefficient, and F1 score should be calculated. Finally, a metric for processing speed in terms of time per image or frame should be included, and although this depends on the hardware that is used, it gives an indication as to whether the system can run in real time.

We must also emphasize that there is a difference in how anomaly detection is defined. In the area of computer science, detection per frame or image is the standard, but in the medical domain, reporting a detection per instance (at least once in a sequence of frames of the same finding) is common. If possible, one should include both definitions.

**CONCLUSION**

Researchers have sought for many years to develop efficient AI tools to assist in medical diagnosis. Enabled by recent hardware developments, several research groups are now working on machine learning-based medical systems and have obtained promising results. Thus, we have observed a rapid increase in publications related to AI in GI endoscopy over the last two years. However, as described above, there are still large variations in the...
tested datasets, and insufficient metrics are being used. In order to enable full comparisons between methods, the same datasets should be utilized, and as many of the common metrics as possible should be used[38]. Another limitation is that the lesion characterization systems rely on advanced endoscopic functionality like narrow-band imaging, endocytoscopy, or volumetric laser endomicroscopy, to which most endoscopy units do not have access, especially in low-income countries[39]. Still, it is not proven that these techniques improve endoscopy performance, and validation in live endoscopy is still required. Therefore, there is still a long road ahead before such systems can be put into practice, and much research, development, and clinical testing still needs to be performed. To produce the best possible and the most comparable results, the recommendations given here should be followed.

REFERENCES

1. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet* 2014; 383: 1490-1502 [PMID: 24225001 DOI: 10.1016/S0140-6736(13)61649-9]

2. World Health Organization - International Agency for Research on Cancer. Estimated cancer incidence, mortality and prevalence world-wide in 2012. Available from: URL: http://globocan.iarc.fr/Default.aspx. 2012

3. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin* 2015; 65: 87-108 [PMID: 25651787 DOI: 10.3322/caac.21262]

4. Hewett DG, Kahi CJ, Rex DK. Efficacy and effectiveness of colonoscopy: how do we bridge the gap? *Gastrointest Endosc Clin N Am* 2010; 20: 673-684 [PMID: 20890701 DOI: 10.1016/j.giec.2010.07.011]

5. Lee SH, Jang BI, Kim KO, Jeon SW, Kwon JG, Kim EY, Jung JT, Park KS, Cho KB, Kim ES, Park CG, Yang CH; DeagsGyengbok Gastrointestinal Study Group. Endoscopic experience improves interobserver agreement in the grading of esophagitis by Los Angeles classification: conventional endoscopy and optimal band image system. *Gut Liver* 2014; 8: 154-159 [PMID: 24672656 DOI: 10.5009/gnl.2014.8.2.154]

6. van Doorn SC, Hazewinkel Y, East JE, van Leerdam ME, Rastogi A, Pellié S, Sanduleanu-Dasculescu S, Bastiaanssen BA, Fockens P, Dekker E. Polyp morphology: an interobserver evaluation for the Paris classification among international experts. *Am J Gastroenterol* 2015; 110: 180-187 [PMID: 25331346 DOI: 10.1038/ajg.2014.326]

7. Lanspa SJ, Lynch HT. Quality indicators for colonoscopy and the risk of interval cancer. *N Engl J Med* 2010; 363: 1371; author reply 1373 [PMID: 20879889 DOI: 10.1056/NEJMc1006842]

8. Rondonotti E, Soncini M, Girelli CM, Russo A, Ballardini G, Bianchi G, Cantù P, Centenara L, Cesarì P, Cortelezzi CC, Gozzini C, Lupinacci G, Maino M, Mandelli G, Mantovani N, Moneghini D, Morandi E, Putignano R, Schalling R, Tarella M, Vitagliano P, Villa F, Zatelli S, Conte D, Masci E, de Franchis R; AIGO, SIED and SIGE Lombardia. Can we improve the detection rate and interobserver agreement in capsule endoscopy? *Dig Liver Dis* 2012; 44: 1006-1011 [PMID: 22858420 DOI: 10.1016/j.dld.2012.06.014]

9. Guishan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 2016; 316: 2402-2410 [PMID: 27898976 DOI: 10.1001/jama.2016.17216]

10. Ciompi F, Chang K, van Riel SJ, Setio AAA, Gerke PK, Jacobs C, Scholten ET, Schaeffer-Prokop C, Wille MMW, Marchianò A, Pastorino U, Prokop M, van Ginneken B. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep* 2017; 7: 46479 [PMID: 28422152 DOI: 10.1038/srep46479]

11. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115-118 [PMID: 28117445 DOI: 10.1038/nature21056]

12. Swager AF, van der Sommen F, Klomp SR, Zinger S, Meijer 600; Open academic

| Dataset | Findings | Frames | Usage |
|---------|----------|-------|-------|
| CVC-356<sup>60</sup> | Polyps | 1706 | ©, by request |
| CVC-612<sup>50</sup> | Polyps | 1962 | ©, by request |
| CVC-12K<sup>50</sup> | Polyps | 11954 | ©, by request |
| Kvasir<sup>59</sup> | Polyps, esophagitis, ulcerative colitis, Z-line, pylorus, cecum, dyed polyp, dyed resection margins, stool | 8000 | Open academic |
| Nerthus<sup>51</sup> | Stool - categorization of bowel cleanliness | 1350 | Open academic |
| GIANA’17<sup>52</sup> | Angiectasia | 600 | ©, by request |
| ASU-Mayo polyp database<sup>53</sup> | Polyps | 18781 | ©, by request |
| CVC-ClinicDB | Polyps | 612 | ©, by request |
| ETIS-Larib Polyp DB | Polyps | 1500 | ©, by request |
| KID<sup>54</sup> | Angiectasia, bleeding, inflammations, polyps | 2500 + 47 videos | Open academic |

*WJG | www.wjgnet.com | 5060 | December 7, 2018 | Volume 24 | Issue 45 |
research purposes. *Endosc Int Open* 2017; 5: E477-E483 [PMID: 28580415 DOI: 10.1055/s-0043-105488]

Wang Z, Meng Q, Wang S, Li Z, Bai Y, Wang D. Deep learning-based endoscopic image recognition for detection of early gastric cancer: a Chinese perspective. *Gastrointest Endosc* 2018; 88: 198-199 [PMID: 29935613 DOI: 10.1016/j.gie.2018.01.029]

P- Reviewer: Hashimoto R, Shichijo S  S- Editor: Ma RY
L- Editor: A  E- Editor: Huang Y
