

# A Random Forest based predictor for medical data classification using feature ranking

Alam et al., Bangladesh Department of CSE, 2019

DOI: 10.1016/j.jim.2019.100180

## INTRODUCTION

As artificial intelligence (AI) becomes more prevalent in our lives, its effect on medical data analysis is also growing. It's currently being utilized to assist and help doctors and healthcare professionals in diagnosing diseases [1]. The end goal isn't to replace them, but to support their decisions. That's where decision support systems (DSS) come in. Machine learning (ML) fits into this by treating diagnosis like a classification problem, where the model tries to guess what's wrong with the patient based on existing data.

However , this isn't a simple process. Real-world medical data is quite messy, inconsistent, and hard to normalize [2]. So far, majority of the studies have been about specific cases [1-3], which limits their usefulness. This paper, however, takes does the opposite. Instead of one dataset, it tries to explore a general approach that can be applied to all ( or most ) medical problems.

**Table 1**  
Brief description of the datasets used in this research.

DataSet	ID	No. Of Features	Training Samples	Testing Samples
Wisconsin Breast Cancer	WBC	9	499	200
Pima Indians Diabetes	PID	8	576	192
Bupa	Bp	6	200	145
Hepatitis	Hp	19	80	75
Heart-Statlog	HtS	13	180	90
SpectF	SF	44	176	91
SaHeart	SHt	9	304	158
PlanningRelax	PRx	12	120	62
Parkinsons	PkS	22	130	65
Hepatocellular Carcinoma (HCC)	HCC	49	110	55

To achieve this, the authors worked with 10 individual datasets. They applied a feature ranking method to pick out the most relevant data

points and used Random Forest as the main classification algorithm. Not only did the model perform well, but the study also highlighted the most important features in each dataset. This could actually help medical professionals understand the data better and make more informed decisions.

## METHODS AND MATERIALS

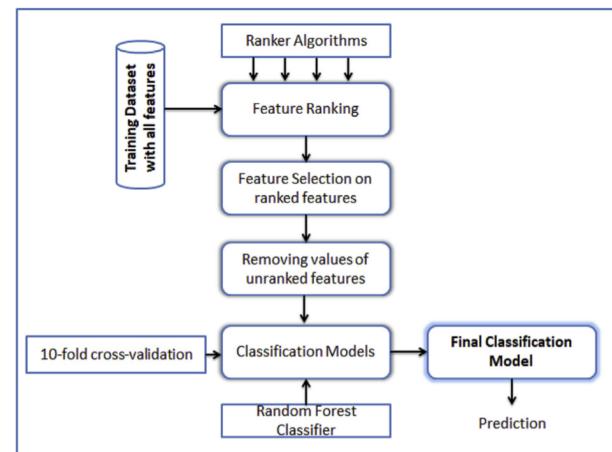


Fig. 1. Model construction overview.

The data sets of this paper were provided by the UCI Machine Learning Repository [4]. To properly test the model, each disease dataset was randomly splitted into training and testing sets. This way, the model will not be overfitted to the data [2]. Before training, the researchers checked which features were more useful for our model by utilizing feature ranking [5-7].

After ranking the features, only the top ones were selected for the final model. Then, the Random

Forest model was used to build and train the classifier. Multiple models were built for each dataset to see which ones perform better. Feature ranking was an important step, identifying and separating useful data for making accurate predictions. Features with negative correlation coefficients were found to hurt model performance, so they were excluded.

## RESULTS

In each dataset, researchers first trained a basic and beginner model using all features, then trained three more models using only the top features. After the comparison, the best-performing model was chosen. Then, only the base model and the best one were used for final testing, instead of testing all four. In almost every case, the models using ranked features gave better results. This explains the reasoning behind removing unimportant features and how it leads to improving both training and testing accuracy.

Compared to other methods, the models outperformed most of the existing approaches on many datasets. In the breast cancer dataset, a deep learning model from another study did slightly better, however small [8]. In the heart disease dataset, one other method had higher accuracy, but again, the difference was small [9]. Overall, the model in this study performed quite well.

## CONCLUSION

Essentially, classifying medical data is quite a complex achievement. To remedy that, a general method that focuses on singling out important features before model training was proposed by the authors. 10 separate databases were tested and, in each case, the results were consistently better when using feature selection instead of doing it with all features. The aim was to improve

the overall accuracy while simplifying the data input process. As tested, SVM, Bayes Network, and Random Forest gave the best results regarding all the datasets. While the paper doesn't introduce new studies or theories, the consistent results make it a strong and impressive contribution to the field.

## REFERENCES

- [1] Chabat F, Hansell DM, Yang G-Z. Computerized decision support in medical imaging. *IEEE Eng Med Biol Mag* 2000;19(5):89–96.
- [2] Mohapatra P, Chakravarty S, Dash P. An improved cuckoo search based extreme learning machine for medical data classification. *Swarm and Evolutionary Computation*, vol. 24. 2015. p. [25–49]. <https://doi.org/10.1016/j.swevo.2015.05.003><https://www.sciencedirect.com/science/article/pii/S2210650215000413>.
- [3] Duda RO, Hart PE. *Pattern classification and scene analysis*. NY, USA: A Wiley- Interscience Publication, John Wiley & Sons; 1973.
- [8] Karthik S, Perumal RS, C.Mouli PVSSR. Breast cancer classification using deep neural networks. *Knowledge Computing and Its Applications* 2018. p. 227–41.  
[https://doi.org/10.1007/978\981\10\6680\1\\_12](https://doi.org/10.1007/978\981\10\6680\1_12)[https://link.springer.com/chapter/10.1007/978/discretionary{-}X\981/discretionary{-}X\10/discretionary{-}X\6680/discretionary{-}X\1\\_12](https://link.springer.com/chapter/10.1007/978/discretionary{-}X\981/discretionary{-}X\10/discretionary{-}X\6680/discretionary{-}X\1_12)
- [9] Eshtay M, Faris H, Obeid N. Improving extreme learning machine by competitive swarm optimization and its application for medical diagnosis problems. *Expert Syst Appl* 2018;104:134152.  
<https://doi.org/10.1016/j.eswa.2018.03.024><https://www.sciencedirect.com/science/article/pii/S0957417418301696>

- [4] University of California at irvine (uci) machine learning repository. <https://archive.ics.uci.edu/ml/datasets.html>.
- [5] Dash M, Liu H. Feature selection for classification. *Intell Data Anal* 1997;1(1–4): [131]–56]. [https://doi.org/10.1016/S1088-467X\(97\)00008-5](https://doi.org/10.1016/S1088-467X(97)00008-5) <http://www.lsi.us.es/~riquelme/publicaciones/kes03.pdf>.
- [6] Ruiz R, Riquelme JC, Aguilar-Ruiz JS. Fast feature ranking algorithm. In: Palade V, Howlett RJ, Jain LC, editors. KES 2003 (LNAI 2773). 2003. p. 325331 <http://www.lsi.us.es/~riquelme/publicaciones/kes03.pdf>.
- [7] J. Novaković, P. Strbac, D. Bulatović, Toward optimal feature selection using ranking methods and classification algorithms, *Yugosl J Oper Res* ([21]). doi:10.2298/YJOR1101119N. URL [http://elib.mi.sanu.ac.rs/files/journals/yjor/41/yujo\\_rn41p119/discretionary{-}135.pdf](http://elib.mi.sanu.ac.rs/files/journals/yjor/41/yujo_rn41p119/discretionary{-}135.pdf).

# Effects of Deep Learning on radiology residents' performance in identifying esophageal cancer on CT

Yasaka et al., British Journal of Radiology, 2023

DOI: 10.1259/bjr.20220685

## INTRODUCTION

Esophageal cancer is the mission of this research paper due to its infamous low survival rate (SR)<sup>1</sup>, which is mainly because of delayed diagnosis. According to the American Cancer Society (ACS), the 5-year SR for localized esophageal cancer is only 47%<sup>2</sup>. However early detection could significantly improve outcomes, especially if the cancer is caught in its initial stages<sup>2</sup>.

Ranked as the 8th most common cancer worldwide in 2020 by the World Health Organization (WHO)<sup>3</sup>, esophageal cancer has a strong association with cigarette smoking<sup>1</sup>. While smoking is also a risk factor for other types of cancer, such as lung<sup>4</sup> and pancreatic cancers<sup>5</sup>, this presents an advantage: incidental detection. Since CT scans are frequently used in clinics<sup>6</sup>, esophageal cancer may be unintentionally identified in its early stage during scans for other conditions.

This is where Deep Learning (DL), specifically Convolutional Neural Networks (CNN), plays a big role in radiology<sup>7</sup>. Studies have shown that DL models can successfully detect, stage, and classify lesions from CT images<sup>8-10</sup>. By building and training these models, authors can improve the process of cancer detection. DL systems are best viewed as decision support services (DSS)<sup>11</sup> that improve accuracy and aid radiologists to make life-saving decisions.

## METHODS AND MATERIALS

As shown in Fig1, Yasaka et al. built a deep learning model that spots signs of cancer from a single CT scan. Like usual in machine learning, the data was split into training, validation, and test sets. After training with labeled scans (both with and without cancer), radiologists could use the model to help detect possible cancer in new patient scans.

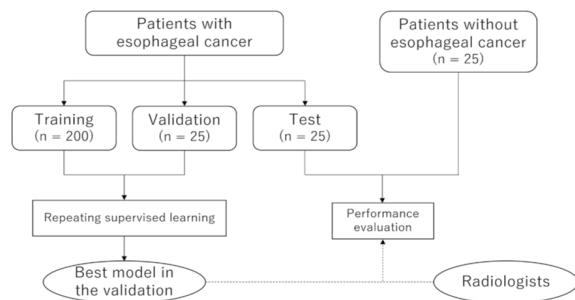


Fig1. Flowchart of the study

CT images were enhanced with contrast to make them easier for the model to understand. More tweaks were done using Python libraries like Pillow and pydicom. The lower parts of the images that didn't show the esophagus were cut out. To add some variety, the images were randomly flipped, rotated, slightly moved, or given noise.

Training and validation ran 15 times in each session, with the number of patients slowly increased by 10 each round. This created 15 models per session. They picked the model with the best results from Session 20 and used that one on the test set.

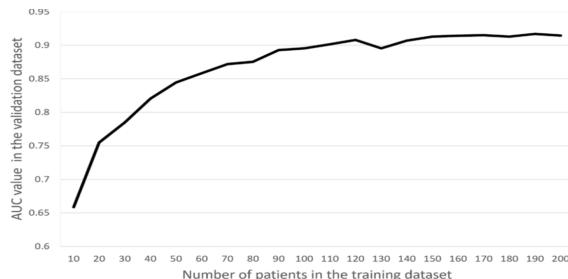


Fig3. Relationship between the training data and validation data

## RESULTS AND DISCUSSION

The relation between the number of patients used in the training data and the performance of the CNN in the validation data set is shown in Fig3. As the number of patients in the training data set increased, CNN performance tended to improve. This trend became less obvious when the number of patients in the training data set exceeded 150. In the validation data set, the accuracy was 86%. While less-experienced readers benefited the most in diagnostic accuracy and confidence, a minor drop in specificity was observed for one reader, showing that AI helps but doesn't always improve every aspect equally.

In this paper, authors found that the DL model built may prove useful to the radiologists, specifically the less experienced readers in the field of esophageal cancer. Yasaka et al. managed to increase the R<sup>2</sup>-score of the previously made model to detect esophageal cancer on CT images<sup>12</sup> from 84% to 92%. With all that said, it is imperative to remind ourselves that the fundamental role of deep learning models is to provide support to radiologists and radiology residents rather than replacing them in clinical settings.

To wrap it up, there were several limitations to the model that are worth mentioning:

1. Although the training split of the data was relatively small, its performance were capped

once we reached about 150 patients, showing signs of diminishing returns and suggesting that adding more data might not improve the model's accuracy any more.

2. The cases with esophageal cancer were clearly visible on the images, meaning the model would probably not do that good noticing the cancer in its earlier stages.

3. Due to how the images were preprocessed, the model isn't really capable of noticing patterns when patients were in lateral position when scanning.

## CONCLUSION

In conclusion, the DL model was helpful mostly for radiologists with lesser experience. There also might be a risk that residents with little experience can lose confidence in diagnosing patients without the expert help of DL model, resulting in an abundance of reliance on artificial intelligence and deep learning technology.

## REFEREENCES

1. Lagergren J, Smyth E, Cunningham D, Lagergren P. Oesophageal cancer. Lancet 2017; 390: 2383–96. [https://doi.org/10.1016/S0140-6736\(17\)31462-9](https://doi.org/10.1016/S0140-6736(17)31462-9)
2. Esophagus Cancer. American Cancer Society Web site. Available from: <https://www.cancer.org/cancer/esophagus-cancer.html> (accessed 25 Jul 2021)
3. World Health Organization Web site. Available from: <https://gco.iarc.fr/today/data/factsheets/cancers/6-Oesophagus-fact-sheet.pdf> (accessed 25 Jul 2021)

4. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. *Eur Respir J* 2016; 48: 889–902. <https://doi.org/10.1183/13993003.00359-2016>
- Hidalgo M. Pancreatic cancer. *N Engl J Med* 2010; 362: 1605–17. <https://doi.org/10.1056/NEJMra0901557>
5. Mettler FA Jr, Thomadsen BR, Bhargavan M, Gilley DB, Gray JE, Lipoti JA, et al. Medical radiation exposure in the U.S. in 2006: preliminary results. *Health Phys* 2008; 95: 502–7. <https://doi.org/10.1097/HP.0000326333.42287.a2>
6. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. *Jpn J Radiol* 2018; 36: 257–72. <https://doi.org/10.1148/radiol.2017170706>
7. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver fibrosis: deep convolutional neural network for staging by using gadoxetic acid-enhanced hepatobiliary phase MR images. *Radiology* 2018; 287: 146–55. <https://doi.org/10.1148/radiol.2017171928>
8. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Deep learning for staging liver fibrosis on CT: a pilot study. *Eur Radiol* 2018; 28: 4578–85. <https://doi.org/10.1007/s00330-018-5499-7>
9. Yamashita R, Mittendorf A, Zhu Z, Fowler KJ, Santillan CS, Sirlin CB, et al. Deep convolutional neural network applied to the liver imaging reporting and data system (LIRADS) version 2014 category classification: a pilot study. *Abdom Radiol (NY)* 2020; 45: 24–35. <https://doi.org/10.1007/s00261-019-02306-7>
10. European Society of Radiology (ESR). What the radiologist should know about artificial intelligence—an ESR white paper. *Insights Imaging* 2019; 10: : 44. <https://doi.org/10.1186/s13244-019-0738-2>
11. Takeuchi M, Seto T, Hashimoto M, Ichihara N, Morimoto Y, Kawakubo H, et al. Performance of a deep learning-based identification system for esophageal cancer from CT images. *Esophagus* 2021; 18: 612–20. <https://doi.org/10.1007/s10388-021-00826-0>

# Early detection of esophageal cancer: Evaluating AI algorithms with multi- institutional narrowband and white-light imaging data

Baik et al., PLOS ONE, 2025  
DOI: 10.1371/journal.pone.0321092

## INTRODUCTION

According to data, esophageal cancer is one of the most common cancers in the world<sup>[1]</sup>, and what makes it worse is the fact that patients tend to be diagnosed way too late in the disease<sup>[2]</sup>, resulting in quite a poor survival rate. With all that said, if the cancer were to be found early in its stages, it's quite curable<sup>[3]</sup>. Unfortunately, doctors have been mainly using white-light imaging (WLI) during the endoscopy procedure to diagnose a patient, and that doesn't yield the best results as it's quite cumbersome and ineffective to do<sup>[4]</sup>. Another useful technique is to identify the structure from the surface of the esophagus using narrowband imaging (NBI)<sup>[5]</sup>. NBI works much better for helping diagnose and detect early signs of esophageal cancer<sup>[6]</sup>.

Due to the cancer's complexity and the challenge it is to accurately detect<sup>[7]</sup>, the procedure barely takes place in a person's life and is often overlooked, resulting in a shortage of high-definition (HD) scans<sup>[8]</sup>. With the advancements of technology and the effect of artificial intelligence (AI) on medical practices<sup>[9]</sup>, there is hope that using this tool might ease and optimize the early detection of such cancers and be of aid to lesser experienced doctors<sup>[10]</sup>.

Using deep learning (DL), specifically its Convolutional Neural Networks' (CNNs) variants, has been utilized for detecting signs from images<sup>[11]</sup>. Even though there are many CNN-based studies, data for esophageal cancer is

still limited compared to others, which has led to poor performance on new inputs<sup>[12]</sup>. Here, Baik et al. are building a new AI model that helps with recognizing early signs of esophageal cancer by checking information on already collected datasets that contain both WLI and NBI formats.

## MATERIALS AND METHODS

Dataset came from a collection of 2,674 images from 619 patients who had done WLI procedure between 2016 and 2020, plus 480 images from 121 patients with NBI. Every image from one patient was put into either the training, validation, or test set, making sure there is zero overlap and data leakage. Since WLI and NBI samples had multiple dimensions, everything was resized to 640x640.

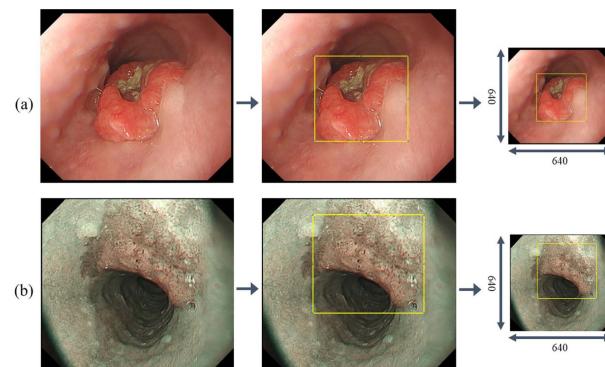


Fig 1. Labeling data for regions of interest, (a) WLI, (b) NBI

A gastroenterologist with 10+ years of experience was present during the model training to label the tumors. To detect tumors in the endoscopy

images, authors used the YOLOv5 model, which is known as a single stage object detection model (ODM). After training, model's performance and precision, and false positives per image (FPPI) were checked.

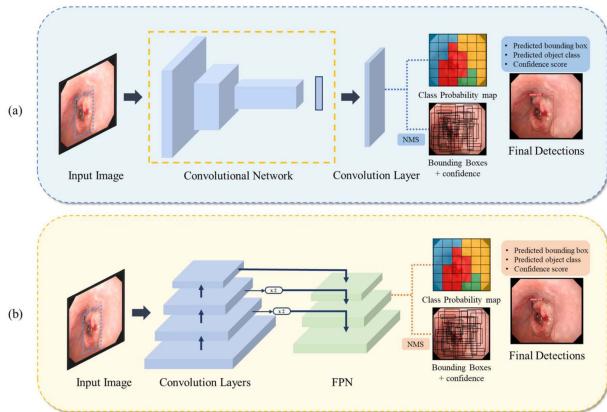


Fig 2. Tumor detection in esophageal Endoscopy, (a) YOLOv5, (b) RetinaNet.

## RESULTS AND DISCUSSION

Finally, models were tested on two types of data: normal images and images with esophageal lesions. Basically, lesion were marked as true and vice versa. For the WLI dataset, the YOLOv5 model hit a precision of 93.7%, and for NBI it scored 86.5% in precision. So, based on the results, this model could definitely help spot tumors by analyzing the images. While prior studies mainly focused on polyp detection, this one goes a step further by focusing on early-stage esophageal cancer [11-19]. That said, a few limitations are still present, such as the poor results against unseen and rare shapes the cancer might take upon.

Going forward, collecting more data, re-training, and tweaking the hyperparameters are some of the recommendations from the authors [20]. Also, optimizing the speed and size of the model could help it run in real-time, which would be useful in medical clinics, resulting in helping new and

unexperienced doctors make quick and accurate decisions on the fly.

## REFERENCES

- Pohl H, Sirovich B, Welch HG. Esophageal adenocarcinoma incidence: Are we reaching the peak? *Cancer Epidemiol Biomarkers Prev*. 2010;19(6):1468–70. <https://doi.org/10.1158/1055-9965.EPI-10-0012> PMID: 20501776
- Hur C, Miller M, Kong CY, Dowling EC, Nattinger KJ, Dunn M, et al. Trends in esophageal adenocarcinoma incidence and mortality. *Cancer*. 2013;119(6):1149–58. <https://doi.org/10.1002/cncr.27834> PMID:23303625
- Behrens A, et al. Barrett's adenocarcinoma of the esophagus: Better outcomes through new methods of diagnosis and treatment. *Dtsch Ärztebl Int*. 2011 108(2011):313. <https://doi.org/10.3238%2Farztebl.2011.0313>
- Nagami Y, Tominaga K, Machida H, Nakatani M, Kameda N, Sugimori S, et al. Usefulness of non-magnifying narrow-band imaging in screening of early esophageal squamous cell carcinoma: A prospective comparative study using propensity score matching. *Am J Gastroenterol*. 2014;109(6):845–54. <https://doi.org/10.1038/ajg.2014.94> PMID: 24751580
- Lee YC, Wang CP, Chen CC, Chiu HM, Ko JY, Lou PJ, et al. Transnasal endoscopy with narrow-band imaging and Lugol staining to screen patients with head and neck cancer whose condition limits oral intubation with standard endoscope (with video). *Gastrointest Endosc*. 2009;69(3 Pt 1):408–17.

- <https://doi.org/10.1016/j.gie.2008.05.033> PMID: 19019362
6. Kuraoka K, Hoshino E, Tsuchida T, Fujisaki J, Takahashi H, Fujita R. Early esophageal cancer can be detected by screening endoscopy assisted with narrow-band imaging (NBI). *Hepatogastroenterology*. 2009;56(89):63–6. PMID: 19453030
7. Li H, Liu D, Zeng Y, Liu S, Gan T, Rao N, et al. Single-image-based deep learning for segmentation of early esophageal cancer lesions. *IEEE Trans Image Process*. 2024;33:2676–88. <https://doi.org/10.1109/TIP.2024.3379902> PMID: 38530733
8. Pennathur A, Gibson MK, Jobe BA, Luketich JD. Oesophageal carcinoma. *Lancet*. 2013;381(9864):400–12. [https://doi.org/10.1016/S0140-6736\(12\)60643-6](https://doi.org/10.1016/S0140-6736(12)60643-6) PMID: 23374478
9. Kudo SE, Misawa M, Mori Y, Hotta K, Ohtsuka K, Ikematsu H, et al. Artificial intelligence-assisted system improves endoscopic identification of colorectal neoplasms. *Clin Gastroenterol Hepatol*. 2020;18(8):1874–1881.e2. <https://doi.org/10.1016/j.cgh.2019.09.009> PMID: 31525512
10. Walter FM, Rubin G, Bankhead C, Morris HC, Hall N, Mills K, et al. Symptoms and other factors associated with time to diagnosis and stage of lung cancer: A prospective cohort study. *Br J Cancer*. 2015;112(Suppl 1):S6-13. <https://doi.org/10.1038/bjc.2015.30> PMID: 25734397
11. Horie Y, Yoshio T, Aoyama K, Yoshimizu S, Horiuchi Y, Ishiyama A, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest Endosc*. 2019;89(1):25–32. <https://doi.org/10.1016/j.jge.2018.07.037> PMID: 30120958
12. Malick A, Soroush A, Abrams JA. Esophageal dysbiosis and esophageal squamous cell carcinoma. *Esoph Dis Microbiome*. 2023;91–114. <https://doi.org/10.1016/b978-0-323-95070-1.00014-x>
13. Ikenoyama Y, Hirasawa T, Ishioka M, Namikawa K, Yoshimizu S, Horiuchi Y, et al. Detecting early gastric cancer: Comparison between the diagnostic ability of convolutional neural networks and endoscopists. *Dig Endosc*. 2021;33(1):141–50. <https://doi.org/10.1111/den.13688> PMID: 32282110
14. Ribeiro E, Uhl A, Wimmer G, Häfner W. Exploring deep learning and transfer learning for colonic polyp classification. *Comput Math Methods Med*. 2016;2016:6584725. <https://doi.org/10.1155/2016/6584725> PMID: 27847543
15. Ding Z, et al. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. *Gastroenterology*. 157(2019):1044–1054. <https://doi.org/10.1053/j.gastro.2019.06.025>
16. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng*. 2018;2(10):741–8. <https://doi.org/10.1038/s41551-018-0301-3> PMID: 31015647
17. 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. <https://doi.org/10.1371/journal.pone.0214133> PMID: 30908513
18. Goda K, Tajiri H, Ikegami M, Yoshida Y, Yoshimura N, Kato M, et al. Magnifying endoscopy with narrow band imaging for

- predicting the invasion depth of superficial esophageal squamous cell carcinoma. *Dis Esophagus*. 2009;22(5):453–60. <https://doi.org/10.1111/j.1442-2050.2009.00942.x> PMID: 19222533
19. Nakagawa K, Ishihara R, Aoyama K, Ohmori M, Nakahira H, Matsuura N, et al. Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists. *Gastrointest Endosc*. 2019;90(3):407–14. <https://doi.org/10.1016/j.gie.2019.04.245> PMID: 31077698
20. Wang CC, Chiu YC, Chen WL, Yang TW, Tsai MC, Tseng MH. A deep learning model for classification of endoscopic gastroesophageal reflux disease. *Int J Environ Res Public Health*. 2021;18(5):2428. <https://doi.org/10.3390/ijerph18052428> PMID: 33801325