Journal of Digital Health
https://ojs.luminescience.cn/JDH

LUMINESCENCE
https://luminescience.cn/

Research Article

The effect of data diversity on the performance of deep learning models for predicting early gastric cancer under endoscopy

Conghui Shi, MA[1,2,3], Jia Li, MA[1,2,3], Lianlian Wu, MD[*1,2,3]

[1] Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan 430060, Hubei Province, China
[2] Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan 430060, Hubei Province, China
[3] Key Laboratory of Hubei Province for Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan 430060, Hubei Province, China

*Correspondence: wu_leanne@whu.edu.cn

Abstract:
Aim:This study aimed to explore the effect of training set diversity on the performance of deep learning models for predicting early gastric cancer (EGC) under endoscopy.

Methods: Images of EGC and non-cancerous lesions under narrow-band imaging (ME-NBI) and magnifying blue laser imaging (ME-BLI) were retrospectively collected. Training set 1 was composed of 150 non-cancerous and 309 EGC ME-NBI images, training set 2 was composed of 1505 non-cancerous and 309 EGC ME-BLI images, and training set 3 was the combination of training set 1 and 2. Test set 1 was composed of 376 non-cancerous and 1052 EGC ME-NBI images, test set 2 consisted of 529 non-cancerous and 71 EGC ME-BLI images, and test set 3 was the combination of test set 1 and test set 2. Three deep learning models, convolutional neural network (CNN) 1, CNN 2 and CNN 3 (CNN 1, CNN 2 and CNN 3 were independently trained using training set 1, training set 2 and training set 3, respectively), were constructed, and their performances on each test set were respectively evaluated. One hundred and thirty-eight ME-NBI videos and 17 ME-BLI videos were further collected to evaluate and compare the performance of each model in real time.

Results: On the whole, the performance of CNN 3 was the best. The accuracy (Acc), sensitivity (Sn), specificity (Sp) and area under the curve (AUC) of test set 1 in CNN 3 were 87.89% (1255/1428), 90.96% (342/376), 86.79% (913/1052) and 94.60%, respectively. The Acc, Sn, Sp and AUC of test set 2 in CNN 3 were 95% (570/600), 97.92% (518/529), 73.24% (52/71) and 90.93% respectively. The Acc, Sn, Sp and AUC of test set 3 in CNN 3 were 89.99% (1825/2028), 95.03% (860/905), 85.93% (965/1123) and 94.89%, respectively. The performance of CNN 3 was also the best in videos test set. The Acc, Sn and Sp of videos test set in CNN 3 were 91.03% (142/156), 90.58% (125/138) and 94.44% (17/18), respectively.

Conclusion: The deep learning model with the most diverse training data has the best diagnostic effect.

Keywords: Deep Learning, Early Gastric Cancer, Data Diversity

Received: Dec.1, 2021; Revised: Feb.14, 2022; Accepted: Feb.17, 2022; Published: Feb.21,2022

DOI: https://doi.org/10.55976/jdh.1202214319-24

Introduction

Gastric cancer is the second leading cause of cancer-related death worldwide and one of the most common cancers in East Asia.[1] Improving the detection rate of early gastric cancer (EGC) can significantly improve the survival rate of gastric cancer patients.[2] White light endoscopy (WLE) is one of the commonly used examination methods of the upper digestive tract. However, it can only detect relatively obvious lesions, and for small lesions such as EGC, the detection rate of WLE is suboptimal. [3,4] It has been demonstrated that magnifying narrow-band imaging (ME-NBI) and blue laser imaging (ME-BLI) have higher diagnostic accuracy than WLE for EGC, and both techniques have been widely applied in clinical practice. [5-7]

However, the performance of endoscopists for using these emerging techniques varies greatly, leading to a low detection rate of EGC, especially in primary medical institutions. [8,9]

With the rapid development of artificial intelligence (AI), deep learning technologies have been widely researched in the field of digestive endoscopy. [10,11] Our research group has developed a deep learning-based system in previous studies, which can identify the differentiation state of EGC and outline the marginality of early gastric cancer under ME-NBI endoscopy. The system correctly predicts the differentiation status of EGCs with an accuracy of 83.3%, which provides great help to endoscopists. [5] However, whether deep learning methods can be used in ME-BLI has not been explored yet.

Data diversity is a common problem in the development of AI systems with health records. [12] In clinical practice, we can use different brands or types of equipment to do the same examination on patients. Different devices have some different characteristics, such as color, imaging resolution, shape, etc. [13] Although these differences in details rarely affect endoscopists' diagnosis, they may interfere with the performances of AI models. Should we extensively collect highly heterogeneous images for training, or should we train different deep learning models for different patterns of images? There are no definitive answers to these questions.

Therefore, in this study, we trained and tested deep learning models with different data diversity and explored the effect of training set diversity on the performance of deep learning models for predicting EGC under endoscopy.

Methods

Data acquisition

The endoscopic images of non-cancerous lesions and EGC were retrospectively obtained from the Renmin Hospital of Wuhan University, including 2414 ME-BLI images (2034 non-cancerous and 380 EGC images) and 3242 ME-NBI images (1881 non-cancerous and 1361 EGC images). The images were divided into training sets and test sets, and images of the training and test sets were from different patients. Training set 1 was composed of 1505 non-cancerous and 309 EGC ME-NBI images, training set 2 was composed of 1505 non-cancerous and 309 EGC ME-BLI images, and training set 3 was the combination of training set 1 and 2. Among them, training set 1 and training set 2 were from different patients with different instruments (ME-NBI and ME-BLI). Test set 1 was composed of 376 non-cancerous and 1052 EGC ME-NBI images, test set 2 was composed of 529 non-cancerous and 71 EGC ME-BLI images, and test set 3 was the combination of test set 1 and test set 2. The detailed sample distribution is shown in Table 1.

Table 1 Composition of data sets

	Training set 1 (NBI)	Training set 2 (BLI)	Training set 3 (NBI+BLI)	Test set 1 (NBI)	Test set 2 (BLI)	Test set 3 (NBI+BLI)
ECG	309	309	618	1052	71	1123
Non-cancerous lesions	1505	1505	3010	376	529	905

NBI: narrow-band imaging, BLI: blue laser imaging, EGC: early gastric cancer.

Poor quality images (resulted from defocus, halation, blurs and so on) were excluded by two doctoral students, and then the images were evaluated by two experienced endoscopists (>10 years of experience) based on Magnifying Endoscopy Simple Diagnostic Algorithm for Early Gastric Cancer (MESDA-G) and the pathologic results. If there was disagreement between the two endoscopists, a reassessment was carried out to reach a consensus.[6] All the images used were acquainted with two instruments (Olympus Optical Co. Ltd. Tokyo, Japan; Fujifilm Co. Kanagawa, Japan).

Model construction and test

ECG recognition models were trained with ResNet 50. [14] Convolutional neural networks (CNN) 1, CNN 2 and

CNN 3 were independently trained using training set 1, training set 2 and training set 3, respectively. Finally, the performance of each model was evaluated in all three test sets.

The optimal parameters were obtained after the repeated training of the model, of which the batch size of parameters was 64, the learning rate was 0.0001, and the iteration ordinal number was 30. Dropout, data augmentation and early stopping were used to reduce the overfitting risk of the model.[15-17] When images of the training set and test set were assigned, image enhancement should be used if necessary to balance the number of images in the two categories, including image translation, rotation, mirroring and cropping. All the algorithms were written in Python 3.6.5, with Keras 2.1.5 and TensorFlow1.12.2 as the backends. The models were trained on the NVIDIA Geforce GTX1080 server, the GPU of which had 8GB of memory.

Running the models on videos

The image-enhanced endoscopic videos of non-cancerous lesions and EGC were retrospectively obtained from the Renmin Hospital of Wuhan University, including 17 ME-BLI videos (11 non-cancerous and 6 EGC videos) and 282 ME-NBI videos (127 non-cancerous and 55 EGC videos). With pathological results as the gold standard, all videos were cut into single lesion videos. Frame -wise prediction was used on the videos at 25 frames per second (fps). The noise was smoothed by the rule of outputting cancer only when more than seven of ten consecutive images were cancer, otherwise, the model outputs non-cancer.

Ethics

This study was approved by the Ethics Committee of the Renmin Hospital of Wuhan University (WDRY2019-K091). The board exempted the informed consent of patients because this was a retrospective study.

Statistical Analyses

The McNemar test was applied to compare the differences in accuracy (Acc), sensitivity (Sn), specificity (Sp), and area under the curve (AUC) among the models. Two-sided statistical tests were conducted, and the P-values < 0.05 was considered statistically significant. The statistical analysis was performed using the SPSS 25. 0 software.

Results

Performances of the models in images

Compared with CNN 1 and CNN 2, CNN 3 performed the best in all the three test sets, especially in Acc. The test results of the models are shown in Table 2 and Figure 1. The Sn, Sp, Acc and AUC of CNN 1 in test set 1 were 87.64% (922/1052), 78.72% (296/376), 85.29% (1218/1428) and 90.07%, respectively. The Sn, Sp, Acc and AUC of CNN 1 in test set 2 were 69.01% (49/71), 97.54% (516/529), 94.17% (565/600) and 93.15%, respectively. The Sn, Sp, Acc and AUC of CNN 1 in test set 3 were 86.46% (971/1123), 89.72% (812/905), 87.92% (1783/2028) and 94.03%, respectively. The Sn, Sp, Acc and AUC of CNN 2 in test set 1 were 86.12% (906/1052), 87.77% (330/376), 86.55% (1236/1428) and 94.06%, respectively. The Sn, Sp, Acc and AUC of CNN 2 in test set 2 were 63.38% (45/71), 95.46% (505/529), 91.67% (550/600) and 83.50%, respectively. The Sn, Sp, Acc and AUC of CNN 2 in test set 3 were 84.68% (951/1123), 92.27% (835/905), 88.07% (1786/2028) and 94.57%, respectively. The Sn, Sp, Acc and AUC of CNN 3 in test set 1 were 86.79% (913/1052), 90.96% (342/376), 87.89% (1255/1428) and 94.60%, respectively. The Sn, Sp, Acc and AUC of CNN 3 in test set 2 were 73.24% (52/71), 97.92% (518/529), 95% (570/600) and 90.93%, respectively. The Sn, Sp, Acc and AUC of CNN 3 in test set 3 were 85.93% (965/1123), 95.03% (860/905), 89.99% (1825/2028) and 94.89%, respectively. The findings are summarized in Table 2 and Figure 1.

Table 2 Performances of the models in images

	Model	Sn	Sp	Acc	AUC
Test set 1 (NBI)	CNN 1	87.64%(922/1052)	78.72%(296/376)	85.29%(1218/1428)	90.07%
	CNN 2	86.12%(906/1052)	87.77%(330/376)	86.55%(1236/1428)	94.06%
	CNN 3	86.79%(913/1052)	90.96%(342/376)#	87.89%(1255/1428)#	94.60%
Test set 2 (BLI)	CNN 1	69.01%(49/71)	97.54%(516/529)	94.17%(565/600)	93.15%
	CNN 2	63.38%(45/71)	95.46%(505/529)	91.67%(550/600)	83.50%
	CNN 3	73.24%(52/71)	97.92%(518/529)^	95.00%(570/600)^	90.93%
Test set 3 (NBI+BLI)	CNN 1	86.46%(971/1123)	89.72%(812/905)	87.92%(1783/2028)	94.03%
	CNN 2	84.68%(951/1123)	92.27%(835/905)	88.07%(1786/2028)	94.57%
	CNN 3	85.93%(965/1123)	95.03%(860/905)#^	89.99%(1825/2028)#	94.89%

* CNN 1 ~ 3 were the deep learning models trained using ME-NBI only, ME-BLI only and ME-NBI+ME-BLI, respectively
# Compare with CNN 1，$P < 0.05$
^ Compare with CNN 2，$P < 0.05$

Sn: sensitivity, Sp: specificity, Acc: accuracy, AUC: area under the curve, NBI: narrow-band imaging, BLI: blue laser imaging, CNN: convolutional neural networks.

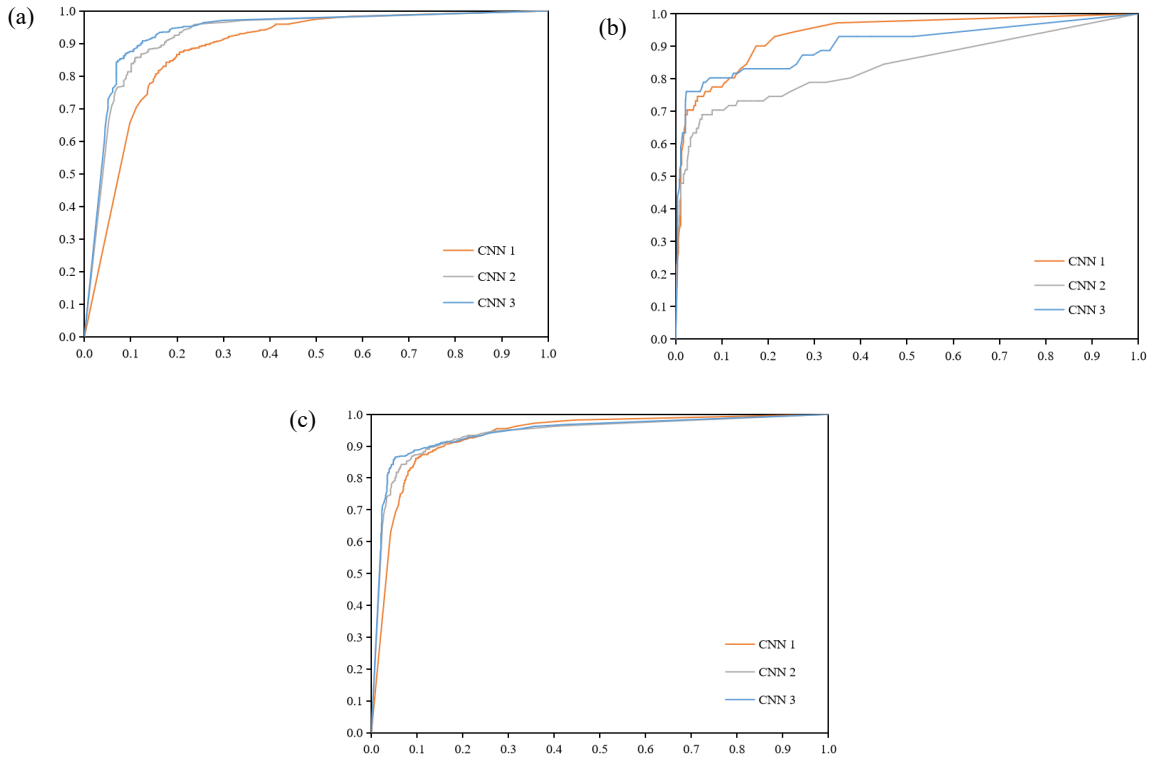

Figure 1. The ROC curves of three CNNs in three test sets. (a). Test set 1; (b). Test set 2; (c). Test set 3. ROC: receiver operating characteristic, CNN: convolutional neural network.

In test set 1 and test set 3, the AUC of CNN 3 was higher than that of CNN 1 and CNN 2, the Acc of CNN 3 was better than that of CNN 1, and the differences were statistically significant ($\chi2$= 9.460, $P < 0.05$; $\chi2$= 10.250, $P < 0.05$). In test set 2, the Acc of CNN 3 was better than that of CNN 2, and the difference was statistically significant ($\chi^2 = 15.042$, $P < 0.05$). In test set 1 and test set 3, the Sp of CNN 3 was better than that of CNN 1, and the difference was statistically significant ($\chi^2 = 40.500$, $P < 0.05$; $\chi^2 = 36.625$, $P < 0.05$). In test set 2 and test set 3, the Sp of CNN 3 was better than that of CNN 2, and the difference was statistically significant ($\chi^2 = 8.471$, $P < 0.05$; $\chi^2 = 8.862$, $P < 0.05$). In test set 1 and test set 3, the Sn of CNN 3 was better than that of CNN 2, and the difference was not statistically significant. In test set 2, the Sn of CNN 3 was better than that of model 1 and CNN 2, and the differences were not statistically significant（$P > 0.05$）

Tests of the models in videos

To explore the performances of the CNNs in a real-time clinical setting, we tested them in real image-enhanced endoscopic videos and calculated the Sn, Sp and Acc of models. The test results of the model are shown in Table 3. The Sn, Sp and Acc of CNN 1 in videos test set were 88.88%(16/18), 63.04%(87/138) and 66.03% (103/156), respectively. The Sn, Sp and Acc of CNN 2 videos test set were 88.88% (16/18), 83.33% (115/138) and 83.97% (131/156), respectively. The Sn, Sp and Acc of CNN 3 in videos test set were 94.44% (17/18), 90.58% (125/138) and 91.03% (142/156), respectively.

In videos test set, the Sn, Sp and Acc of CNN 3 were better than those of CNN 1 and CNN 2. The Acc and Sp of CNN 3 were better than those of CNN 1, and the differences were statistically significant ($\chi^2 = 33.581$, $P < 0.05$; $\chi^2 = 34.225$, $P < 0.05$). Other statistical results were not statistically significant. ($P > 0.05$).

Table 3 Performances of the models in real-time videos

Model	Sn	Sp	Acc
CNN 1	88.88%(16/18)	63.04%(87/138)	66.03% (103/156)
CNN 2	88.88%(16/18)	83.33% (115/138)	83.97% (131/156)
CNN 3	94.44% (17/18)	90.58%(125/138)#	91.03% (142/156)#

*:CNN 1 ~ 3 were the early gastric cancer recognition model under ME-NBI, ME-BLI, and ME-NBI and ME-BLI, respectively
#: Compare with CNN 1, P <0.05
Sn: sensitivity, Sp: specificity, Acc: accuracy, CNN: convolutional neural networks.

Discussion

This study explored the effect of data diversity on the performances of deep learning models for predicting early gastric cancer, which is different from most other studies focusing on the auxiliary diagnosis of EGC with AI. We retrospectively collected image-enhanced endoscopic images and videos, and deep learning models for predicting early gastric cancer were trained according to the types of images in the training set and compared among the models. The results showed that CNN 3 which was trained by both ME- NBI images and ME- BLI images performed better.

ME-NBI and ME-BLI are commonly used in clinical endoscopy, which is of great significance in the diagnosis of early gastric cancer. Compared with traditional white light, ME-NBI and ME-BLI are of better diagnostic performance. [5-7] The difference in the diagnosis of EGC by endoscopists is a major clinical problem, especially for junior doctors. Endoscopists in different regions also differ significantly in the use of ME-NBI and ME-BLI in the diagnosis of EGC. These factors seriously affect the detection rate of early gastric cancer, thus affecting the prognosis of these patients. A good deep learning system for predicting early gastric cancer can solve these problems.

In this study, we found that the performance of deep learning models for predicting early gastric cancer trained by the two training sets was superior to that trained by the single training set. These results suggest that if the types of training sets are increased, deep learning models trained with a variety of training sets may have better performances in diagnosing EGC. The model trained by multiple training sets may have more significant potential in assisting endoscopists in clinical practice.

In clinical practice, if early intervention for cancer can be carried out, the survival rate and later quality of life of patients will be greatly improved, so the diagnosis of early cancer is of great significance in the treatment and prognosis of patients. The early diagnosis of the whole digestive tract tumor (such as colorectal cancer), not only of EGC, is very important. [18] Increasing the diversity of training sets of deep learning-based early gastric cancer recognition models can improve their performances. [19] Then, can other deep learning-based early gastrointestinal cancer recognition models also improve the diagnosis rate of early cancer by this method?

This study also has many limitations. First, this is a retrospective test, and it is difficult to collect complete and sufficient data because some information cannot be included in the medical records. Second, the sample size of the ME-BLI image and video is small and should be strengthened in further study. Third, this is a single-center trial, and further multi-center trials should be conducted to prove the robustness of the results in this study.

In summary, we constructed three deep learning models with different data diversity in training sets and fully validated each model in different test sets. The results showed that the model with the most data diversity performed the best, indicating that in the case of a limited sample size, we should collect heterogeneous images for training deep learning models, rather than use different patterns of images to train different deep learning models.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [published correction appears in CA Cancer J Clin. 2020 Jul;70(4):313]. *CA: A Cancer Journal for Clini*cians 2018; 68(6):394-424. doi:10.3322/caac.21492.

[2] de Vries AC, Kuipers EJ. Epidemiology of premalignant gastric lesions: implications for the development of screening and surveillance strategies. *Helicobacter*. 2007; 12 Suppl 2:22-31. doi:10.1111/j.1523-5378. 2007. 00562. x.

[3] Panteris V, Nikolopoulou S, Lountou A, Triantafillidis JK. Diagnostic capabilities of high-definition white light endoscopy for the diagnosis of gastric intestinal metaplasia and correlation with histologic and clinical data. *European Journal of Gastroenterology & Hepatology*. 2014; 26(6):594-601. doi:10.1097/MEG.0000000000000097.

[4] Quénéhervé L, Neunlist M, Bruley des Varannes S, Tearney G, Coron E. Nouvelles stratégies d'analyse endoscopique des maladies digestives [Novel endoscopic techniques to image the upper

gastrointestinal tract]. *Medecine Sciences: M/S*. 2015; 31(8-9):777-783. doi:10.1051/medsci/20153108017.

[5] Ling T, Wu L, Fu Y, et al. A deep learning-based system for identifying differentiation status and delineating the margins of early gastric cancer in magnifying narrow-band imaging endoscopy. *Endoscopy*. 2021; 53(5):469-477. doi:10.1055/a-1229-0920.

[6] Muto M, Yao K, Kaise M, et al. Magnifying endoscopy simple diagnostic algorithm for early gastric cancer (MESDA-G) [published correction appears in Dig Endosc. 2016 Jul;28(5):630]. *Digestive Endoscopy*. 2016; 28(4):379-393. doi:10.1111/den.12638.

[7] Sivanathan V, Tontini GE, Möhler M, Galle PR, Neumann H. Advanced endoscopic imaging for diagnosis of inflammatory bowel diseases: Present and future perspectives. *Digestive Endoscopy*. 2018; 30(4):441-448. doi:10.1111/den.13023.

[8] Li L, Chen Y, Shen Z, et al. Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging. *Gastric Cancer*. 2020; 23(1):126-132. doi:10.1007/s10120-019-00992-2.

[9] Lau JYW, Yu Y, Tang RSY, et al. Timing of Endoscopy for Acute Upper Gastrointestinal Bleeding. *New England Journal of Medicine*. 2020; 382(14):1299-1308. doi:10.1056/NEJMoa1912484.

[10] Min JK, Kwak MS, Cha JM. Overview of Deep Learning in Gastrointestinal Endoscopy. *Gut and Liver*. 2019; 13(4):388-393. doi:10.5009/gnl18384.

[11] Wu L, He X, Liu M, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. *Endoscopy*. 2021; 53(12):1199-1207. doi:10.1055/a-1350-5583.

[12] Fatoum H, Hanna S, Halamka JD, Sicker DC, Spangenberg P, Hashmi SK. Blockchain Integration With Digital Technology and the Future of Health Care Ecosystems: Systematic Review. *Journal of Medical Internet Research*. 2021; 23(11): e19846. Published 2021 Nov 2. doi:10.2196/19846.

[13] Tao GL, Liu YK, Tang JJ, et al. *Zhonghua Shao Shang Za Zhi*. 2021; 37(8):747-751. doi:10.3760/cma.j.cn501120-20200318-00179.

[14] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition. 2016 *IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, 2016; 770 - 778. doi:10.1109/ CVPR. 2016. 90.

[15] Baldi P, Sadowski P. The Dropout Learning Algorithm. *Artificial Intelligence*. 2014; 210:78-122. doi: 10.1016/j.artint.2014.02.004.

[16] Tanner MA,Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*,1987; 82(398):528 - 540. doi: 10.2307/2289457

[17] Prechelt L. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*. 1998; 11(4):761-767. doi: 10. 1016 / s0893 - 6080(98)00010 - 0.

[18] Tauriello DV, Calon A, Lonardo E, Batlle E. Determinants of metastatic competency in colorectal cancer. *Molecular Oncology*. 2017; 11(1):97-119. doi:10.1002/1878-0261.12018.

[19] Clark RD. Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics. *Journal of Computer-Aided Molecular Design*. 2003; 17(2-4):265-275. doi:10.1023/a:1025366721142.