© Pol J Radiol 2021; 86: e532-e541
DOI: https://doi.org/10.5114/pjr.2021.110309

Received: 04.05.2020
Accepted: 31.07.2020
Published: 13.09.2021

Polish Journal of Radiology

http://www.polradiol.com

Original paper

A fundamental study assessing the generalized fitting method in conjunction with every possible coalition of N-combinations (G-EPOC) using the appendicitis detection task of computed tomography

Tomoyuki Noguchi[1,2,3A,B,C,E], Yumi Matsushita[1D,F,G], Yusuke Kawata[4F], Yoshitaka Shida[4B], Akihiro Machitori[4G]

[1]Education and Training Office, Department of Clinical Research, Centre for Clinical Sciences, Japan
[2]Department of Radiology, National Hospital Organization Kyushu Medical Centre, Jigyohama, Chuo-ku, Fukuoka City, Fukuoka Province, Japan
[3]Department of Clinical Research, National Hospital Organization Kyushu Medical Centre, Jigyohama, Chuo-ku, Fukuoka City, Fukuoka Province, Japan
[4]Department of Radiology, National Centre for Global Health and Medicine, Japan

Abstract

Purpose: Increased use of deep learning (DL) in medical imaging diagnoses has led to more frequent use of 10-fold cross-validation (10-CV) for the evaluation of the performance of DL. To eliminate some of the (10-fold) repetitive processing in 10-CV, we proposed a "generalized fitting method in conjunction with every possible coalition of N-combinations (G-EPOC)", to estimate the range of the mean accuracy of 10-CV using less than 10 results of 10-CV.

Material and methods: G-EPOC was executed as follows. We first provided (2N-1) coalition subsets using a specified N, which was 9 or less, out of 10 result datasets of 10-CV. We then obtained the estimation range of the accuracy by applying those subsets to the distribution fitting twice using a combination of normal, binominal, or Poisson distributions. Using datasets of 10-CVs acquired from the practical detection task of the appendicitis on CT by DL, we scored the estimation success rates if the range provided by G-EPOC included the true accuracy.

Results: G-EPOC successfully estimated the range of the mean accuracy by 10-CV at over 95% rates for datasets with N assigned as 2 to 9.

Conclusions: G-EPOC will help lessen the consumption of time and computer resources in the development of computer-based diagnoses in medical imaging and could become an option for the selection of a reasonable K value in K-CV.

Key words: neural networks (computer), machine learning, learning curve, computer simulation, appendicitis, cross validation.

Introduction

Computer diagnosis systems with deep learning (DL), which have made remarkable progress in recent years, are replacing the visual recognition that previously only humans could process. Medical imaging diagnoses have shown good compatibility with DL, and diagnostic radiology might change significantly with the incorporation of DL, before other medical fields. DL is used to identify liver tumours with dynamic contrast computed tomography (CT), to detect cerebral aneurysms with magnetic resonance (MR) angiography, and to diagnose pulmonary tuberculosis, mediastinal adenopathy, and pulmonary nodules, and more [1-7]. DL compensates for the shortcomings of human intelligence that depend on the mental and physical state of the operator, and DL can be integrated in diagnostic systems.

Correspondence address:
Tomoyuki Noguchi, MD, PhD, Department of Radiology, National Hospital Organization Kyushu Medical Center, 1-8-1, Jigyohama, Chuo-ku, Fukuoka City, Fukuoka Province, Japan, Zip Code: 810-8563, phone: +81-92-852-0700, fax: +81-92-847-8802, e-mail: tnogucci@radiol.med.kyushu-u.ac.jp

Authors' contribution:
A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection

This is an Open Access journal, all articles are distributed under the terms of the Creative Commons Attribution-Noncommercial-No Derivatives 4.0 International (CC BY-NC-ND 4.0). License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

However, the use of DL is still accompanied by many difficulties, and the algorithm structure of DL continues to be analysed and improved. DL is one of the machine learning algorithms comprised of multiple layers that consist of multiple collections of interconnected nodes. DL extracts abstract features while the input data transfers from shallow layers to deeper layers, and the DL outputs an answer that essentially matches human intentions. DL has been successfully applied to natural language processing, visual object recognition, and speech recognition, which are difficult for conventional computer processing [8-10]. Among the existing relevant algorithms, DL has a high affinity for medical image recognition [6].

Although DL requires training with a large amount of data and the adjustment of internal variable parameters to output the correct answer with high accuracy, DL has the advantage in that its performance improves in proportion to the size of the data. However, DL does not always give a proper answer to real-world questions even when its use provides a perfect score for the training datasets, in a phenomenon known as 'over-fitting' [11]. In such cases, it is sometimes difficult to identify which of the data, training procedure, and/or algorithms cause the low performance of DL. Tools that rigorously evaluate the performance of DL are thus needed.

K-fold cross-validation (K-CV) is one of the verification methods most commonly used for evaluations of the performance of machine learning [12-16]. K-CV is performed as follows: 1) the dataset is split into K equal parts; 2) the ratio of the data is set as (1/K) and (1 – [1/K]) for the testing and training image numbers, respectively; 3) the DL algorithm is trained with training images; 4) the DL algorithm is tested with testing images to evaluate its judgment ability; 5) this process is repeated K times, each time with the selection of a different pair of testing and training datasets; 6) finally, the mean accuracy is determined by averaging the K test results. K-CV has the advantage of being a simple procedure without complicated mathematics, using a limited number of data under low bias.

However, there are some issues relating to the undefined optimal K value of K-CV. Although K-values are arbitrary, 5 or 10 for K (5- or 10-CV) is recommended in general [14-17]. The most common reason for recommending 5-CV is that it takes less time to calculate than 10-CV. In addition, the Pareto principle, otherwise known as the 80/20 law, in which most of the representative values of the whole group is yielded by 20% of the group, might also be related to the recommendation of 5-CV [16,17]. However, 5-CV has a lower accuracy value compared to 10-CV because the accuracy value of DL substantially depends on the training data volume, which is 4/5 of the total data in 5-CV and 9/10 in 10-CV [14].

The advantage of 10-CV is that the division of a dataset according to decimal notation is simple and easy. The measurement error in 10-CV is relatively small compared to that in 5-CV [13-15]. However, 10-CV has a disadvantage; i.e., its use requires the long-term use of computational resources with 10 repetitions to train and test the DL. 5-CV is occasionally used to avoid such long-lasting computational processing [16]. A method that could skip some of the 10-fold repetitive processing in 10-CV without loss of accuracy is thus desirable.

Noguchi *et al*. proposed a method with which the accuracy value of 10-CV can be estimated using less than 10 results of the 10-CV by the bootstrap method in conjunction with every possible coalition of N-combinations [14]. They demonstrated that their method estimated the ranges including the mean accuracy value of 10-CV with the use of only 6 of 10 results of the 10-CV at a rate of over 95%; i.e., they showed that their boot-EPOC method could skip 4 of the 10 processing repetitions in 10-CV. However, the boot-EPOC method could not estimate the mean accuracy value using less than 6 of 10 results of the 10-CV. Here, we proposed the 'generalized fitting method in conjunction with every possible coalition of N-combinations (G-EPOC)', which can estimate the mean accuracy value of 10-CV using 2 to 9 results of the 10-CV at a rate of over 95%. We conducted the present study to validate the estimating ability of G-EPOC.

Material and methods

Study design

This study was approved by our hospital's Institutional Review Board, which waived the need for written informed consent from the patients. The study was conducted as part of a fundamental study on the development project of Computer-Assisted Diagnosis with Deep Learning Architecture (CADDELAC) for the detection of appendicitis on CT.

Terminology

'Mean accuracy': The term 'mean accuracy' in this study denotes the average score from the 10 results of 10-CV.

'Estimation': G-EPOC is used not for the detection task of the appendicitis on CT but for predicting, or in other words 'estimating', the mean accuracy of 10-CV.

'Combination' and 'Coalition': The term 'combination' is used to denote a selection of items from a collection without regard for the order of selection. The number of N combinations from X elements is often denoted by XCN and is equal to X!/N!(X-N)!.

The term 'coalition' is used as the number of N combinations for all N from X elements [18]. For example, when we extract 3 samples, named s1, s2, and s3, we can build seven coalitions as |s1|, |s2|, |s3|, |s1, s2|, |s1, s3|, |s2, s3|, and |s1, s2, s3|. With N samples, the number of coalition subsets is (2^N-1).

Methodology of G-EPOC

G-EPOC is composed of 3 steps: 1) less than 10 data sets extracted from 10 results of 10-CV are incremented by the coalition method; 2) the mean and 95% confidence interval (CI) upper and lower limits for each of incremented datasets are acquired by the primary statistical processing; and 3) these data undergo secondary statistical processing to determine the estimation range for the mean accuracy of 10-CV. This estimation range should include the mean accuracy of 10-CV. The details are provided below.

Test design

We performed 3 types of tests: a simulation test, a practical test, and a practical test using the first N sampling. In the simulation test, G-EPOC was evaluated for the datasets with severe distribution. In the practical test, we validated G-EPOC using 5 datasets of 10-CVs, which assessed the ability of DL to detect the appendicitis on CT. In the practical test using the first N sampling, we evaluated G-EPOC under the pragmatic process to obtain the estimation range. The details are as follows.

Simulation test

In the first test of G-EPOC, we estimated the ability of G-EPOC to estimate the range of the mean accuracy of 10-CV using a simulation dataset as follows:

- Simulation dataset preparation: We formed one dataset comprising 10 samples of 100 items each, simulating the 10 results of 10-CV. Each item had 0 points (for false) or 1 point (for true) in the Boolean expression manner, and the possible summed scores in each sample thus ranged from 0 to 100. We set the distribution of the scores of the 10 samples to follow the sigmoid function $1/(1 + \exp[-\alpha x])$ under $\alpha = 1.28$ and $x = 0.05$ to 0.95 with 0.1 as the increment, adjusting to the conditions for the maximal and minimal scores ranging from 0 to 100. Figure 1 illustrates the distribution of the scores in the simulation dataset.
- Data processing: If we choose a given N, which is 9 or less, out of 10 samples in the simulation dataset, we can build $(2^N - 1)$ coalition subsets as mentioned above [18]. As the primary statistical processing, we applied the coalition subsets to the probability distribution fitting which was one of the mathematical curve fitting methods. We then obtained $(2^N - 1)$ sets of 3 types of values including the means and the 95% CI upper and lower limits. From those sets, we composed 2 types of paired datasets: the mean versus the 95% CI upper limit, and the mean versus the 95% CI lower limit.
  As the secondary statistical processing, we applied these datasets to the generalized linear model fitting, which was another mathematical curve fitting method. We then obtained a total of 6 types of values: 2 of the mean values, 2 of the 95% CI upper limits, and 2 of the 95% CI lower limits.
  We regarded a range between the 2 mean values obtained in the secondary statistical processing as the standard estimated range. We also generated a wide estimated range from the maximal and minimal values of the 6 values in the secondary statistical processing for the maximal upper limit and the minimal limit, respectively. If the estimation range limits were less than 0 or greater than 100, they were replaced with 0 or 100, respectively. Various mathematical distribution models can be used for the curve fitting method. In the primary and secondary statistical processing, we used 3 types of distribution; the normal distribution, the Poisson distribution, and the binominal distribution. We thus obtained a total of 9 standard estimated ranges and 9 wide estimated ranges. Figure 2 provides a schematic explanation of the process for making the estimation ranges from N samples.
- Judgement: We scored the estimation success rates if the range by G-EPOC included the true mean of all 10 samples in the simulation dataset for all combinations with N assigned as 2, 3, 4, …, and 9, with which the numbers of combinations were ${}_{10}C_2 = 45$, ${}_{10}C_3 = 120$, ${}_{10}C_4 = 210$, …, and ${}_{10}C_9 = 10$, respectively. Finally, we selected the best series of statistical processing for each of N samples according to the following conditions: (1) more than 95% of the number rate of the successful estimated ranges, and (2) the narrowest averaged estimated range width.

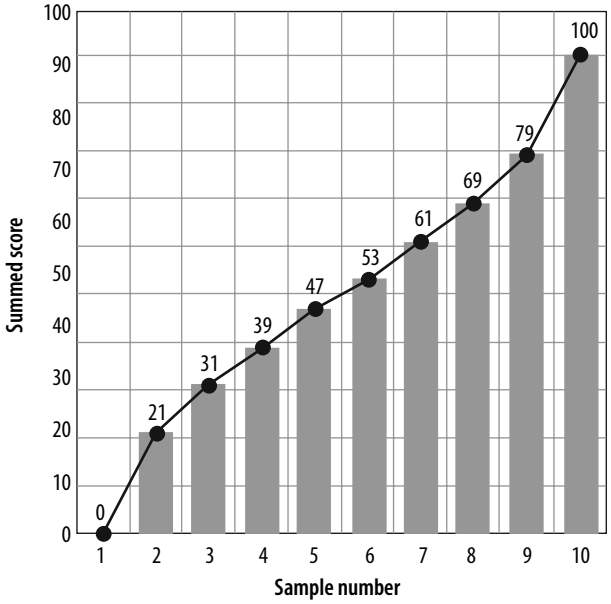

Figure 1. Distribution of the scores of 10 samples in the simulation dataset. We set the distribution to follow the sigmoid function $1/(1 + \exp(-\alpha x))$ under $\alpha = 1.28$ and $x = 0.05$ to 0.95 with 0.1 as the increment, adjusting to the conditions for 100 and 0 for the maximal and minimal scores, respectively

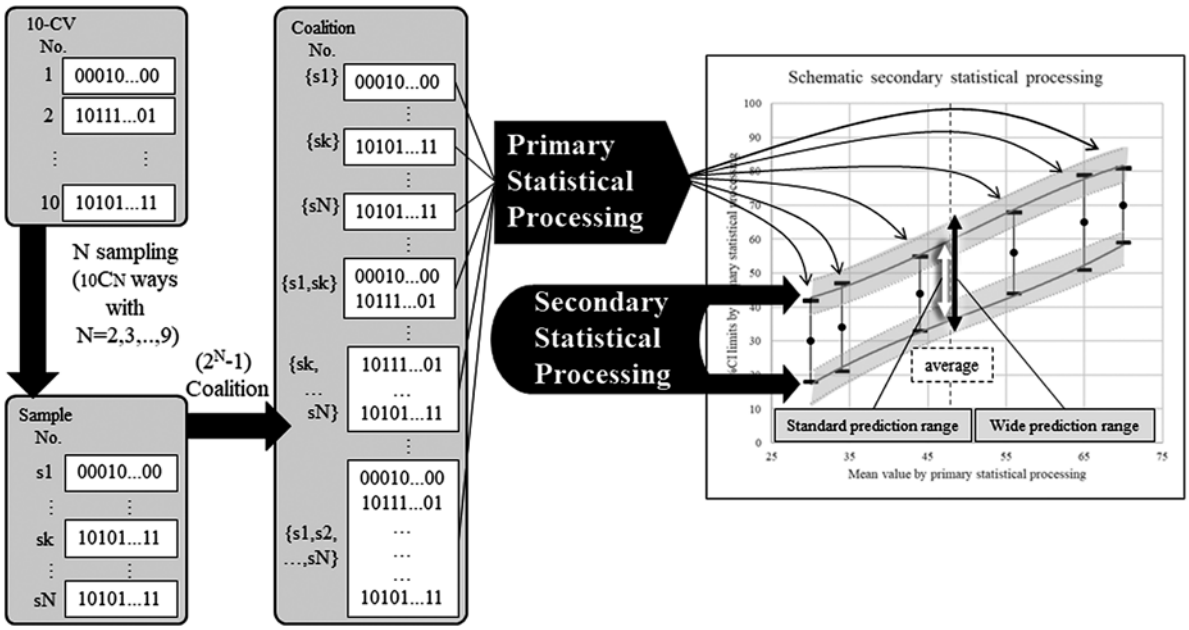

Figure 2. The process for making the estimation ranges from N samples. After we made (2N – 1) possible coalitions from N samples of 10 results of 10-CV, we applied them to 3 types of probability distribution fitting and then 3 types of generalized linear model fitting as the secondary processing. We then obtained the standard estimated range and the wide estimated range from the upper, mean, and lower limits of 95% CIs with N assigned as 2 to 9

Practical test

In the second test for G-EPOC, we assessed the ability of G-EPOC to estimate the mean accuracy of 10-CV validating the practical datasets as follows:

- Patients: As the appendicitis group, we enrolled a total of 485 patients with appendicitis, who underwent a contrast-enhanced computed tomography examination (CECT) due to acute abdomen in the period from January 2010 to March 2019 (male/female, 269/216; age range/average, 6-98/38.5 years). They were diagnosed based on 1-mm slice thickness CT as well as clinical findings. The following CT criteria were used for diagnosing appendicitis [19-22]: appendix diameter of more than 6 mm, appendiceal wall enhancement, presence of appendicoliths, periappendiceal fat stranding, and concomitant abscess.
  For the control group, we randomly selected 485 patients who underwent a CECT due to acute abdomen enrolled backward from March 2019 to December 2016, in order to include the same number of patients as in the appendicitis group (male/female, 245/240; age range/average, 13-97/64.2 years). Their results were negative for appendicitis but might be positive for other acute abdominal disorders. Their detailed diagnoses are not provided here because they were too varied to describe and they were not relevant to the present study. Figure 3 shows representative examples of CT images in the appendicitis and control groups.
- Computed tomography: Multidetector row CT was performed using clinical CT units (Aquilion CX, Aquilion 64, and Aquilion ONE, Canon Medical Systems, Ohtawara, Japan; SOMATOM Definition Flash, Siemens Healthineers, Erlangen, Germany; Discovery CT750 HD; GE Healthcare, Chicago, IL). CT images were generated using a body window algorithm, a 220-500-mm field of view (FOV), 512 × 512 matrix, and a 1-mm slice. The iodine contrast agent was intravenously injected according to the following protocols: iodine volume per weight, 600 mgI/kg; Contrast agent, iopamidol (300 mgI/ml), iopamidol (370 mgI/ml), iohexol (300 mgI/ml), iohexol (350 mgI/ml), or iomeprol (300 mgI/ml); injection speed, 5 to 2.0 ml/s bolus injection; start of helical scanning, 70 to 230 s after the injection of the contrast. Decreased volumes were adopted in cases of the patient's declining renal function based on the physician's request.
- Extraction of positive and negative image data: We chose 20,690 positive images that included any cross-sections of the enlarged and enhanced appendix indicating appendicitis extracted from the patients with appendicitis. We used multiple slices from the same subject as described previously [4,6,14].
  We extracted 297,636 negative images from the control group, covering the entire abdominal areas.
- Practical dataset preparation: We prepared five datasets of 124, 250, 374, 500, and 624 pairs of the randomly extracted positive and negative CT images from both the appendicitis group and control group, respectively. The sizes of five datasets were adjusted to meet 100, 200, 300, 400, and 500 pairs of the training images with the ratio of 1/10, 1/10, and 8/10 for the testing, verification, and training image numbers for each of the datasets.

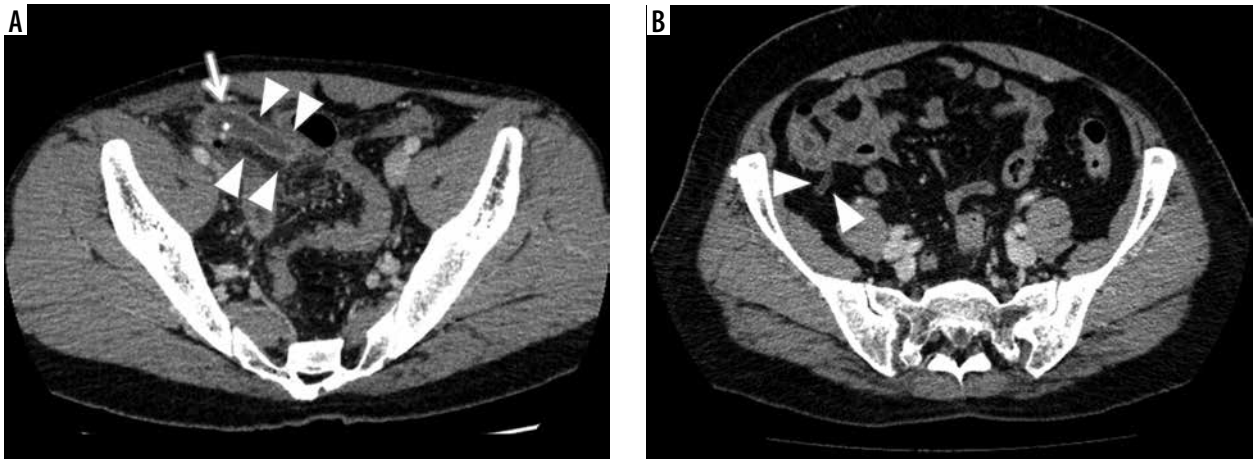

Figure 3. Representative examples of CT images in the appendicitis and control groups. **A)** Contrast-enhanced abdominal CT of a 32-year-old male in the appendicitis group shows an enlarged appendix with thickened wall enhancement (arrowheads) in conjunction with fluid and an appendicolith inside (arrow) at the right lower quadrant in the abdomen. **B)** The CECT of a 54-year-old female in the control group shows a normal gas-filled appendix with non-enhanced thin wall (arrowheads) at the apex of the cecum

We regularized the sampling size based on the size of the training images instead of the total sample images because 10-CV could be more sensitive to the former than the latter [14]. We did not use all the 20,690 positive images because we realized that a 10-CV using 20,690 positive images provided more than 90% of the mean accuracy, which had a ceiling effect and was thus not suitable for the current test.

- Image data processing: We used one of the representative DL convolutional neural networks, AlexNet, which was designed by Alex Krizhevsky and won the ImageNet Large-Scale Visual Recognition Challenge in 2012, in the pre-trained state with over a million images from the ImageNet database (ImageNet. http://www.image-net.org) [23]. We adopted the transfer learning method for training AlexNet as described previously [4,6,14]. The AlexNet that we used in this study did not support the Digital Imaging and Communications in Medicine (DICOM) format. Using the free software program Fiji (https://imagej.net/Fiji), we converted these images from the DICOM format into Portable Network Graphics (PNG) with 16-bits and 227 × 227 pixels of the image size with adjustment of the brightness and contrast of the original DICOM image by using the look-up-table (LUT) function in the Fiji program.
- 10-CVs for validating AlexNet enabled for the task of detecting appendicitis: We applied 10-CV to the above-mentioned 5 practical datasets as follows. (1) The dataset was split into 10 equal parts. (2) The ratio of the data was set as 1/10, 1/10, and 8/10 for the testing, validation, and training image numbers, respectively. (3) AlexNet was trained with the training images. (4) AlexNet was validated with the validation images to prevent over-training. (5) AlexNet was tested with the testing images to evaluate its judgment ability. (6) This process was repeated 10 times, each time with a different 3 parts of testing, validation, and training datasets. (7) The mean accuracy was determined by averaging the 10 results of the testing.

  As the referential standard, we determined the mean accuracies and the 95%CI ranges with the t-distribution calculated from the result datasets of the 10-CV in each of the 5 practical datasets.
- Testing and judgment: We changed each of the result datasets to the Boolean expression consisting of 0 points for false and 1 point for true. We applied each of the N samples extracted from the result datasets of the 10-CV in each of the 5 practical datasets to the best G-EPOC processing determined in the simulation test as described above. We scored the number rates of the estimation range, which successfully included each of the mean accuracies of 10-CVs of the 5 practical datasets for all combinations with N assigned as 2 to 9. We then recorded the successful estimation rates, the average of the successful estimated range widths, and the averages of upper and lower limits of the successful estimated ranges. We plotted the mean accuracy line of the 5 practical datasets in conjunction with the averaged estimation range areas (ERAs) whose upper and lower boundary lines were the averaged upper and lower limits of the successful estimated ranges by G-EPOC.

Test using the first N sampling

In the third test for G-EPOC, we estimated the ability of G-EPOC in the following pragmatic situation:

In a pragmatic situation, G-EPOC users do not necessarily calculate all combinations with a given N, which is assigned as 2 to 9, out of the result datasets of 10-CV. Instead, they may stop performing 10-CV halfway and use the first N of the result datasets of 10-CV for making the estimation range. Moreover, they may regard the midpoint of the estimation range by G-EPOC as a surrogate estimated value for the mean accuracy of 10-CV. Assuming such a case, we made the following judgment:

Table 1. Results of the simulation test

Number of samplings from 10 samples in simulation dataset	Distribution model used in the primary statistical processing	Distribution model used in the secondary statistical processing	Use of wide estimated range	Successful estimation rate (%)	Average of successfully estimated range width (%)
2	Binominal distribution	Binominal distribution	Yes	100	88.8
3	Normal distribution	Binominal distribution	Yes	100	73.3
4	Normal distribution	Binominal distribution	Yes	100	58.7
5	Normal distribution	Binominal distribution	Yes	100	46.2
6	Normal distribution	Binominal distribution	Yes	100	36.2
7	Normal distribution	Binominal distribution	Yes	99	28.6
8	Normal distribution	Binominal distribution	Yes	98	22.9
9	Poisson distribution	Normal distribution	No	100	13.6

We obtained the midpoints between the upper and lower limits of the estimation range in each of the first sampling trial with N assigned as 2 to 9. We recorded the number N as successfully estimated by G-EPOC when the 95% CI ranges with the t-distribution of the 5 practical datasets included the midpoints provided by G-EPOC. We plotted the mean accuracy line and the 95% CI range areas with the t-distribution of the 5 practical datasets, plus the midpoint lines afforded by G-EPOC from the result datasets of the 10-CV in each of the 5 practical datasets for the first N sampling. We also evaluated the expected time-saving (%) of the first N of the result datasets of 10-CV calculated from (C – (A + B))/C, where A was equal to the processing time (s) of the first N sampling of G-EPOC, B was equal to the acquisition time (s) of the first N sampling of the practical datasets of 10-CV, and C was equal to the acquisition time (s) of all 10 result datasets of 10-CV.

Other conditions

Hardware and software for computational processing

We used a custom-built image processing computer (TEGARA Corp., Hamamatsu, Japan) containing a Quadro P2000 5 GB graphics processing unit (Nvidia Corp., Santa Clara, CA), an Intel Xeon E5-2680v4 2.40 GHz processor (Intel Corp.), 1.0 TB of hard disk space, and 64 GB of RAM. We used MATLAB software (ver. 2018b; MathWorks Inc., Natick, MA) for all statistical computational processing. The AlexNet algorithm that we used was distributed as add-on software of MATLAB.

Results

Simulation test

For the best series of the primary and secondary statistical processing, we identified the binominal and binominal distribution models for N as 2, the normal and binominal distribution models for N as 3 to 8, and the Poisson and normal distribution models for N as 9. We adopted the wide estimated ranges for N as 2 to 8 and the standard estimated range for N as 9. As a result, we obtained 98% to 100% successful estimation rates with all numbers of sampling out of 10 samples in the simulation dataset. When the N was increased from 2 to 9, the averaged widths of the successful estimated ranges narrowed from 88.8% to 13.6%. Table 1 summarizes the results of the simulation test.

Practical test

We obtained 98% to 100% successful estimation rates for all numbers of sampling in the 5 practical datasets by applying those datasets to the best series of the successful estimation processing obtained in the simulation test. Table 2 shows the mean accuracies of the 5 practical datasets, the averages of those estimated range widths, and the averages of the upper and lower limits of those estimated ranges by G-EPOC. Figure 4 illustrates the mean accuracy line of the 5 practical datasets and ERAs with the N assigned as 2 to 9 by G-EPOC. The mean accuracy line is completely covered by ERAs by G-EPOC. The upper and lower limits of the ERAs approached the mean accuracies of 10-CV results as the number of samplings increased.

Table 2. Results of the practical test

Number of samplings from 10 samples in simulation dataset	Successful estimation rate (mean/range)	Average of successfully estimated range width (mean/range)
2	100.0% (98-100)	87.6% (85.5-91.3)
3	100.0% (100-100)	73.6% (67.6-79.8)
4	99.7% (94-100)	60.8% (49.7-76.4)
5	98.8% (92-100)	51.0% (36.8-76.1)
6	98.0% (88-100)	43.9% (30.4-83.8)
7	98.0% (90-100)	39.2% (26.1-89.5)
8	98.5% (89-100)	36.2% (21-95.6)
9	99.0% (72-100)	13.8% (8.7-16.4)

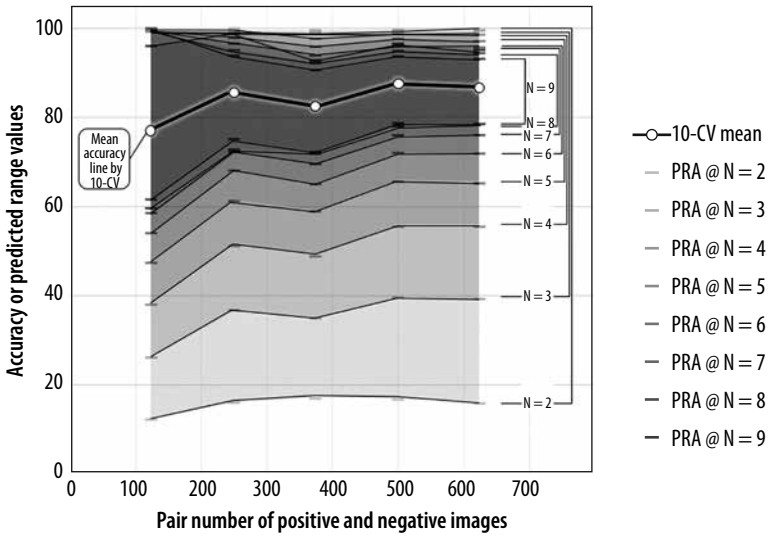

Figure 4. The mean accuracy line for 5 practical datasets and the estimation range areas (ERAs) by G-EPOC. The mean accuracy line is completely covered by ERAs by G-EPOC

Test using the first N sampling

The midpoints of the ranges between the upper and lower limits of the first N sampling with the N assigned as 8 and 9 by G-EPOC were included by the 95% CI range with the t-distribution for all 5 practical datasets. However, some of the midpoints with the N assigned as 5 to 7 and all of those with the N assigned as 2 to 4 missed the 95% CI range with the t-distribution. Figure 5 provides the mean accuracy line and the 95% CI range areas with the t-distribution calculated from the 5 practical datasets, plus the midpoint lines from the result datasets of the 10-CV in each of the 5 practical datasets for the first N sampling by G-EPOC with the N assigned as 2 to 9.

Tables 4-6 show the processing time (sec) of the first N sampling of G-EPOC, the acquisition time (sec) of the first N sampling of the practical datasets including the total 10 result datasets of 10-CV, and the expected time-saving (%) of the first N of the result datasets of 10-CV. The acquisition times of the first N sampling of the practical datasets ranged from 363 to 8118 sec, whereas the processing time of G-EPOC ranged from 0.61 to 3.64 sec. The expected time-saving (%) was thus approximately 10% per one result of 10-CV skipped.

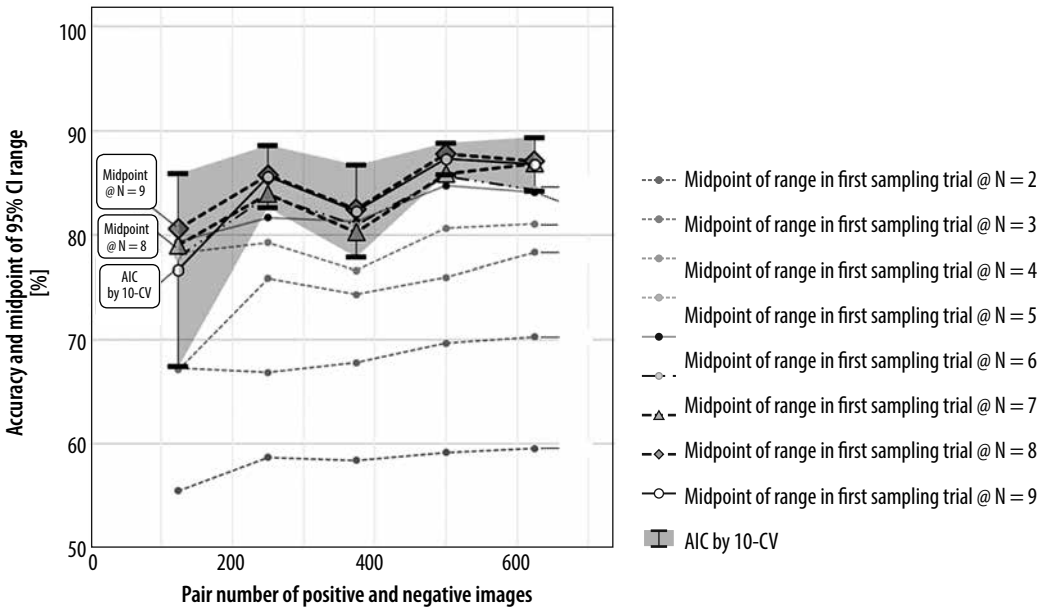

Figure 5. The mean accuracy line and 95% CI range areas with t-distribution of the 5 practical datasets, and the midpoint lines in each for the first N sampling trial. The midpoints of between the upper and lower limits of the first sampling by G-EPOC with 8 and 9 as the N were included by the range of 95% CI with the t-distribution of all 5 practical datasets. However, some of the midpoints with the N assigned as 5 to 7 and all of those with the N assigned as 2 to 4 missed the 95% CI range with the t-distribution

Table 3. Results of the random dataset test

Number of sampling from 10 samples in 10-CV for practical dataset	Pair number of positive and negative images	Averaged upper and lower limits and midpoints of the estimation ranges (midpoint [upper limit – lower limit / width])				
		124	250	374	500	624
2		56% (12.3-99.7/87.4)	58% (16.1-99.9/83.8)	58.4% (17.1-99.7/82.6)	58.4% (16.9-99.9/83)	57.9% (15.9-99.9/84)
3		62.9% (26.3-99.6/73.3)	68.3% (36.9-99.7/62.8)	67% (35.1-99/64)	69.6% (39.7-99.5/59.8)	69.3% (39.3-99.4/60.1)
4		68.7% (38.1-99.3/61.2)	75.1% (51.3-99/47.7)	73.3% (48.9-97.6/48.7)	77.1% (55.5-98.8/43.3)	77% (55.5-98.5/43)
5		73.4% (47.5-99.2/51.7)	79.6% (61.3-97.8/36.5)	77.4% (58.9-95.9/37)	81.6% (65.7-97.6/31.9)	81.2% (65.3-97.1/31.8)
6		76.9% (54-99.8/45.8)	82.3% (68.1-96.6/28.5)	79.5% (65.1-93.9/28.9)	84.2% (72.1-96.3/24.2)	83.8% (71.9-95.7/23.9)
7		79.2% (58.6-99.9/41.3)	83.7% (72.3-95.2/22.9)	81% (69.6-92.3/22.7)	85.5% (76.1-94.9/18.9)	85.2% (76-94.4/18.4)
8		80.8% (61.7-100/38.3)	84.5% (75.1-93.9/18.7)	81.4% (72.2-90.6/18.4)	86.3% (78.8-93.8/15)	85.9% (78.6-93.2/14.5)
9		77.9% (59.7-96.1/36.4)	85.8% (72.7-98.8/26.1)	82.4% (71.9-92.8/21)	87.3% (78.1-96.6/18.6)	86.8% (78.5-95.1/16.6)
10 (=total acquisition)		76.7% (72.7-88.7/16)	85.6% (80.6-91.1/10.5)	82.3% (83-90/7)	87.3% (71.6-93.2/21.6)	86.8% (78-92/14)

Table 4. Acquisition time (sec) of the first N sampling of G-EPOC

Number of the first sampling from 10 samples in 10-CV for practice dataset	Pair number of positive and negative images	Acquisition time (sec) of the first N sampling of G-EPOC				
		124	250	374	500	624
2		0.28	0.29	0.26	0.26	0.29
3		0.38	0.35	0.34	0.34	0.37
4		0.39	0.41	0.39	0.38	0.39
5		0.49	0.48	0.48	0.49	0.5
6		0.66	0.68	0.67	0.69	0.69
7		1.03	1.05	1.06	1.09	1.12
8		1.79	1.85	1.86	1.92	1.92
9		3.28	3.38	3.45	3.54	3.64

Discussion

In the simulation test, we obtained 98% to 100% successful estimation rates with all numbers of sampling out of 10 samples in the simulation dataset. Because the possible summed scores in one sample thus ranged from 0 to 100, which presumed the extreme case, our present analyses clarified that G-EPOC could estimate the range of the mean accuracy by any situation of 10-CV using less than 10 results at the rates of more than 95%. We adopted the standard or wide ranges calculated by the primary and secondary statistical processing using the binominal, normal, or Poisson distribution, although the mathematical appropriateness for the current processing was not clarified.

In the practical test, we obtained 98% to 100% successful estimation rates for all numbers of sampling in the 5 practical datasets by G-EPOC using the best series of the successful estimation processing obtained in the simulation test. Our results confirmed that G-EPOC could estimate the range of the mean accuracy by any situation of

Table 5. Acquisition time (s) of the first N sampling of the practical datasets

Number of the first sampling from 10 samples in 10-CV for practice dataset	Pair number of positive and negative images	Acquisition time (sec) of the first N sampling of the practical dataset				
		124	250	374	500	624
2		363	670	1044	1535	1642
3		639	920	1672	2174	2532
4		755	1284	2413	3293	3191
5		946	1572	3240	4043	4220
6		1155	1879	3559	4903	5249
7		1402	2243	4019	5652	5862
8		1452	2702	4366	6734	6661
9		1652	2896	4628	7779	7459
10 (= total acquisition)		1824	3506	5060	8270	8118

Table 6. Expected time-saving (%) of the first N of the result datasets of 10-CV

Number of the first sampling from 10 samples in 10-CV for practice dataset	Pair number of positive and negative images	Expected time-saving (%) of the first N of the result datasets of 10-CV				
		124	250	374	500	624
2		80%	81%	79%	81%	80%
3		69%	74%	67%	74%	69%
4		61%	63%	52%	60%	61%
5		48%	55%	36%	51%	48%
6		35%	46%	30%	41%	35%
7		28%	36%	21%	32%	28%
8		18%	23%	14%	19%	18%
9		8%	17%	9%	6%	8%

10-CV using less than 10 results in a practical situation. Figure 4 demonstrates that the upper and lower limits of the estimation range of G-EPOC approached the mean accuracies of 10-CV results as the number of samplings increased. However, the estimation ranges of G-EPOC are apparently wider compared to the 95% CI ranges of 10-CV results with t-distribution. This is a compensatory trade-off for accomplishing over 95% accuracy.

We assessed the pragmatic use of G-EPOC; i.e., the midpoint of the estimation range with the first N sampling as a surrogate estimated value. As a result, the midpoints of the estimation range with the N assigned as 8 and 9 were placed inside of the 95%CI range with the t-distribution of the 5 practical datasets. Therefore, G-EPOC might be effectively used with the N assigned as 8 or 9, in which approximately 10% to 20% would be spared because about 10% of time-saving per one result of 10-CV was confirmed.

G-EPOC has been proposed as an option for the selection of the optimal K value in K-CV; i.e., it is possible to make an almost perfect estimation even if the training and validating DL is repeated fewer than 10 times in 10-CV. Although G-EPOC does not necessarily have a direct clinical contribution, G-EPOC as an alternative to 10-CV will save time and computer resources in the process of developing the CADDELAC system. G-EPOC might be also effective for the early introduction of the CADDELAC system for clinical applications.

There are some limitations to this study. Narrower range estimations should be achieved with further investigations, although we suggested the pragmatic use of the midpoint of the estimation range with the first N sampling. The G-EPOC findings obtained herein are empirical results, and there is still no mathematical or statistical basis to support them. Further research is required.

Conclusions

G-EPOC demonstrated over 95% successful estimation for the accuracy determined by 10-CV, using less than 10 results of 10-CV, which was confirmed by the datasets obtained from the real detection task of appendicitis on CT by AlexNet. Our series will help lessen the consumption of time and computer resources in the development of computer-based diagnoses in medical imaging.

Acknowledgements

This work was supported in part by Japan Agency for Medical Research and Development (AMED) under Grant Number JP18lk1010028.

Conflicts of interest

The authors report no conflict of interest.

References

1. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 2018; 286: 887-896.
2. Nakao T, Hanaoka S, Nomura Y, et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J Magn Reson Imaging 2018; 47: 948-953.
3. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017; 284: 574-582.
4. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016; 35: 1285-1298.
5. Hua KL, Hsu CH, Hidayati SC, et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 2015; 8: 2015-2022.
6. Noguchi T, Higa D, Asada T, et al. Artificial intelligence using neural network architecture for radiology (AINNAR): classification of MR imaging sequences. Jpn J Radiol 2018; 36: 691-697.
7. Ueda D, Yamamoto A, Nishimori M, et al. Deep learning for MR angiography: automated detection of cerebral aneurysms. Radiology 2019; 290: 187-194.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521: 436-444.
9. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 2016; 35: 1153-1159.
10. Cheng PM, Malhi HS. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. J Digit Imaging 2017; 30: 234-243.
11. Lanka P, Rangaprakash D, Dretsch MN, et al. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. Brain Imaging Behav 2020; 14: 2378-2416.
12. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society Series B (Methodological) 1974; 36: 111-147.
13. Hastie T, Tibshirani R, Friedman J. 7.10.1 K-fold Cross Validation. In The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009; 241-245.
14. Noguchi T, Uchiyama F, Kawata Y, et al. A fundamental study assessing the diagnostic performance of deep learning for a brain metastasis detection task. Magn Reson Med Sci 2020; 19: 184-194.
15. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence – Volume 2. Montreal: Morgan Kaufmann Publishers Inc.; 1995, pp. 1137-1143.
16. Breiman L, Spector P. Submodel Selection and Evaluation in Regression. The X-Random Case. International Statistical Review/Revue Internationale de Statistique 1992; 60: 291-319.
17. Newman MEJ. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 2005; 46: 323-351.
18. Dang VD, Jennings NR. Generating Coalition Structures with Finite Bound from the Optimal GuaranteesProceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems – Volume 2. New York: IEEE Computer Society; 2004, pp. 564-571.
19. Balthazar EJ, Birnbaum BA, Yee J, et al. Acute appendicitis: CT and US correlation in 100 patients. Radiology 1994; 190: 31-35.
20. Balthazar EJ, Megibow AJ, Siegel SE, Birnbaum BA. Appendicitis: prospective evaluation with high-resolution CT. Radiology 1991; 180: 21-24.
21. Malone A Jr, Wolf C, Malmed A, Melliere B. Diagnosis of acute appendicitis: value of unenhanced CT. AJR Am J Roentgenol 1993; 160: 763-766.
22. Noguchi T, Gibo M, Murata S. CT findings of the normal appendix. Comparison with Ba enema study. Rinsho Hoshasen 1999; 44: 339-344.
23. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networksProceedings of the 25th International Conference on Neural Information Processing Systems – Volume 1. Lake Tahoe: Curran Associates Inc.; 2012, pp. 1097-1105.