**Noname manuscript No.**
(will be inserted by the editor)

# Chronic gastritis classification using gastric X-ray images with a semi-supervised learning method based on tri-training

**Zongyao Li · Ren Togo · Takahiro Ogawa · Miki Haseyama**

the date of receipt and acceptance should be inserted later

**Abstract** High-quality annotations for medical images are always costly and scarce. Many applications of deep learning in the field of medical image analysis face the problem of insufficient annotated data. In this paper, we present a semi-supervised learning method for chronic gastritis classification using gastric X-ray images. The proposed semi-supervised learning method based on tri-training can leverage unannotated data to boost the performance that is achieved with a small amount of annotated data. We utilize a novel learning method named Between-Class learning (BC learning) that can considerably enhance the performance of our semi-supervised learning method. As a result, our method can effectively learn from unannotated data and achieve high diagnostic accuracy for chronic gastritis.

**Keywords** Chronic gastritis · computer-aided diagnosis · medical image analysis · convolutional neural network · semi-supervised learning

Zongyao Li (corresponding author)
Graduate School of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, 060-0814, Japan
Tel.: +81-11-706-6078
Fax: +81-11-706-6078
E-mail: li@lmd.ist.hokudai.ac.jp

Ren Togo
Faculty of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, 060-0814, Japan
Tel.: +81-11-706-6078
Fax: +81-11-706-6078
E-mail: togo@lmd.ist.hokudai.ac.jp

Takahiro Ogawa
Faculty of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, 060-0814, Japan
Tel.: +81-11-706-6078
Fax: +81-11-706-6078
E-mail: ogawa@lmd.ist.hokudai.ac.jp

Miki Haseyama
Faculty of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, 060-0814, Japan
Tel.: +81-11-706-6078
Fax: +81-11-706-6078
E-mail: miki@ist.hokudai.ac.jp

## 1 Introduction

It is known that chronic gastritis may lead to gastric cancer, and gastric X-ray images can be used for diagnosing chronic gastritis and identifying the risk of gastric cancer [29]. Clinicians can make a highly reliable diagnosis with gastric X-ray images, but there is a huge burden of reading many gastric X-ray images. Additionally, clinicians must have abundant experience and technical knowledge of reading gastric X-ray images in order to make an accurate diagnosis. To reduce the burden for clinicians and overcome the possible problem of shortage of experienced clinicians, the development of computer-aided diagnosis (CAD) systems that automatically analyze gastric X-ray images and detect chronic gastritis is needed.

In recent years, deep learning technologies have made tremendous achievements in the field of computer vision [17] and have considerably outperformed conventional machine learning methods in various tasks. Convolutional neural networks (CNNs), the most popular deep learning technology, are also widely utilized in tasks of medical image analysis [18] such as tissue segmentation [12, 19, 20, 28] and nodule detection [5, 11, 21, 23]. For the task of gastritis classification using gastric X-ray images, we have proposed methods of both conventional machine learning and deep learning in our

previous works [25, 26]. The deep learning method [26] trained two CNNs. One CNN is trained for extracting patches related to gastritis from all regions in the whole X-ray images and the other CNN is trained with the extracted gastritis-related patches to recognize gastritis accurately. The deep learning method showed an obviously better performance than that of the conventional machine learning method [25], which utilized hand-crafted features and a support vector machine (SVM) [6], and the deep learning method even outperformed diagnosis with blood inspection [16]. However, like most of the studies on utilization of deep learning in medical image tasks, our previous deep learning method is based on supervised learning and its performance is therefore dependent on a large number of gastric X-ray images being annotated by experts. Supervised learning methods always require a massive dataset to avoid overfitting and thus may not be practical for some applications of medical images. The main problem of applying deep learning technologies to medical image analysis is the difficulty in obtaining annotations. High-quality annotations of medical images require clinicians or technologists who are experienced and specialize in specific studies. One possible solution for this problem is semi-supervised learning. Semi-supervised learning methods utilize both limited annotated data and unannotated data, which are usually abundant and easy to obtain. By leveraging unannotated data, semi-supervised learning methods can outperform supervised learning methods using only a small amount of annotated data. Some researchers have presented results of their works using semi-supervised learning in medical image tasks [1–3, 22, 31], but the use of semi-supervised learning for diagnosing chronic gastritis with gastric X-ray images has not been explored.

In this paper, we present a semi-supervised learning method based on tri-training [32] for the task of chronic gastritis classification using gastric X-ray images. Tri-training is a disagreement-based semi-supervised learning method [33] that is practical and easy to implement compared with some other semi-supervised learning methods with sophisticated hyper-parameters. The method trains three models and exploits the disagreements of the models to augment the training set of each model with unlabeled data. The models can be retrained with augmented training sets iteratively to achieve better performance. We perform tri-training with three CNNs of different architectures to keep the models diverse and strengthen robustness of the results. Furthermore, we employ the Between-Class learning (BC learning) method proposed by Tokozume et al. [27] as a data augmentation method for training. The trick of BC learning can also improve the learning in terms of Fisher's criterion [10] and considerably boost the performance. Results of experiments indicate that our semi-supervised learning method can realize a high level of diagnostic accuracy for chronic gastritis even with a very limited number of annotated images. Our method can also be employed with various models and is expected to be applied to various medical image tasks.

The paper is organized as follows. In Section 2, we describe in detail the chronic gastritis classification with our semi-supervised learning method. In Section 3, we show the experimental settings and results that prove the effectiveness of our method. We further discuss the experimental results in Section 4. Finally, we conclude the paper in Section 5.

## 2 Semi-supervised learning for chronic gastritis classification

We propose a patch-based method to classify X-ray images. The original X-ray images are cropped into patches and classified into three classes including gastritis, non-gastritis and irrelevant (outside the stomach). The classification result of a whole X-ray image is obtained by performing a simplest majority voting among patches predicted to be gastritis and non-gastritis from the image. Our semi-supervised learning method has been developed with a tri-training architecture and a data augmentation method known as BC learning (or mixup [30], a training trick similar to BC learning). The procedures of our method are shown in Fig. 1. In the rest of this section, research data and patch producing are described in Subsection 2.1, and details of tri-training and BC learning are given in Subsection 2.2 and Subsection 2.3.

### 2.1 Research Data

In our study, we used 815 gastric X-ray images from different patients provided by The University of Tokyo Hospital. The labels of patients were obtained by cooperative evaluation of X-ray images and endoscopic images. Specifically, X-ray images were classified into four classes, normal, mild, moderate and severe, according to the atrophic level [8], and endoscopic images were evaluated with the Kimura-Takemoto seven-grade classification that contains seven classes including no atrophic change (C0), three closed types of atrophic gastritis (C1, C2, C3) and three open types of atrophic gastritis (O1, O2, O3) [15]. We labeled 240 patients with results of mild, moderate or severe for X-ray images and results of C2, C3, O1, O2 or O3 for endoscopic images as positive (gastritis). A total of 575 patients with results

of normal for X-ray images and results of C0 for endoscopic images were labeled as negative (non-gastritis). To exclude potential noises, our dataset included only patients that were diagnosed as having gastritis or non-gastritis consistently by using both X-ray images and endoscopic images. A diagnosis of C1, which is the atrophic borderline in the Kimura-Takemoto seven-grade classification, is too ambiguous to make a definite diagnosis of either gastritis nor non-gastritis. Hence, patients diagnosed as C1 were not included in our dataset.

It is very difficult for original X-ray images with a high resolution of 2048 $\times$ 2048 pixels to be directly processed by CNNs, and it is also not appropriate to downsample the original images since the fine-grained features that play a significant role in diagnosis may be harmed. Therefore, we cropped the images with a stride of 50 pixels into patches with a resolution of 299 $\times$ 299 pixels. By using the patches, the number of training samples was increased hundreds of times, which can also make the training of CNNs easier. The patches were categorized into the following three classes: region with gastritis, region without gastritis and region outside the stomach. The categorization for patches was performed in terms of hand-craft stomach boundaries and patient labels both annotated by a radiological technologist. Specifically, patches from gastritis patients were labeled as either a region with gastritis or a region outside the stomach, and patches from non-gastritis patients were labeled as either a region without gastritis or a region outside the stomach.

### 2.2 Semi-supervised Learning with Tri-training

Tri-training proposed by Zhou et al. [32] is a semi-supervised learning algorithm for discriminative models. It was originally implemented with conventional machine learning models such as an SVM and random forest [4], and it can also be combined with deep learning. In brief, in tri-training, performance promotion is achieved by augmenting the labeled dataset with unlabeled data accompanied by the corresponding predicted labels. In the whole training procedure, three models are firstly trained with three initial training sets, and then two steps are performed iteratively: 1) augmenting each of the three training sets and 2) newly training the three models with the corresponding augmented training sets obtained in 1). The key idea of tri-training is its particular augmenting architecture, in which the training set of each model is augmented by the other two models. Specifically, for each model, unlabeled samples that are categorized as the same class by the other two models are added to the training set of the model, accompanied by a consistent predicted label. Note that we do not set a threshold of probability to select unlabeled samples since we think the consistent predictions made by two models are sufficiently reliable for training and redundant hyper-parameters are not desired. As illustrated in Fig. 1, the tri-training algorithm is composed of the following three steps.

- **Step 1.** Three training sets are sampled by the bootstrap method [9] from labeled data and then three SVM classifiers are trained with the training sets. The features for training SVMs are extracted with the Inceptionv3 network [24] pre-trained on the Imagenet dataset [7].
- **Step 2.** Labels of unlabeled samples are predicted with the latest three models (SVMs or CNNs) and then the unlabeled samples that are classified into the same class by two models are added to the training set of the remaining model.
- **Step 3.** Three CNN classifiers are trained with the augmented training sets. Then the process goes back to step 2.

Note that we utilize SVMs as the initial model since the size of the labeled dataset may be extremely small and CNNs trained with a very limited number of samples tend to fall into overfitting and to be unstable. In step 3, the models trained with augmented training sets are all CNNs. We use three different network structures for the CNNs to keep the models diverse and enhance robustness of the method. The networks are constructed in the forms of ResNet [13], DenseNet [14] and the simplest architecture composed of several convolutional and fully connected layers, respectively.

Although the augmented training sets include some noise samples with incorrect labels, the models trained with augmented training sets can obviously outperform models trained with only labeled data, especially in the case of limited labeled data. Then with the outperforming models, augmented training sets of higher quality are obtained and the performances of models trained with the augmented training sets are further improved. The iteration times can be determined according to the similarity of the three models' predictions for unlabeled data.

### 2.3 Data Augmentation with BC Learning

BC learning is proposed as a novel learning method that can boost the performance of CNNs for image classification [27]. We employ the BC learning method to augment the training data in our semi-supervised learning method. The BC learning method can be simply defined as follows.

$$X = r\ X_1 + (1 - r)\ X_2 \tag{1}$$

$$Y = r\ Y_1 + (1 - r)\ Y_2 \quad (2)$$

Here, $X_1$ and $X_2$ are original image samples, $Y_1$ and $Y_2$ are one-hot label vectors of the samples, and $r$ is always a random ratio sampled from a uniform distribution of [0, 1] in the whole training. The pair of $X$ and $Y$ is a new augmented sample for training.

Different from other data augmentation methods that only consider the vicinity of a single sample, BC learning produces samples between two different samples and hence can model the feature distributions across different classes. The constraints shown in Eq. (1) and Eq. (2) can enlarge the intra-class distance and narrow the inter-class distance simultaneously in the feature distributions. Furthermore, the positional relationship among feature distributions is also regularized so that the between-class samples are not distributed around the decision boundaries of other classes. Benefiting from modeling the feature distributions, models trained with BC learning have greater generalization ability and can achieve better performances.

In our method, we produce samples between different classes as well as inside a class, namely, the original samples $X_1$ and $X_2$ may either belong to different classes or the same class. This is different from the setting that performed best in the original work on BC learning, but in our semi-supervised learning method, we think this setting can better adapt the training sets involving some noise. To prevent the augmented training sets from being polluted by more noise, as the sampling strategy, either $X_1$ or $X_2$ must be sampled only from the labeled data. Note that BC learning is utilized only for training CNNs since an SVM cannot be trained with between-class labels.

## 3 Experiments

### 3.1 Implementation Details

#### *3.1.1 Network Architectures*

The architectures of the three CNNs are described here and shown in Fig. 2. In all three networks, each convolutional layer is followed by the ReLU function and a batch normalization layer.

**The simplest network** The simplest network is composed of three convolutional layers and two fully connected layers. The filter sizes of the convolutional layers are 5×5×32, 5×5×64 and 3×3×64, respectively. Each convolutional layer is followed by a max-pooling layer. Two fully connected layers with 100 units and 3 units respectively are connected to the top of the convolutional layers.

**ResNet-based network** The ResNet-based network is composed of a convolutional layer, three residual blocks and two fully connected layers. The first convolutional layer and two fully connected layers are the same as those of the simplest network. Each residual block consists of two units, and each unit contains two convolutional layers. The filter sizes of the residual blocks are 5×5×32, 5×5×64 and 3×3×64, respectively. Three max-pooling layers are inserted after the first convolutional layer, the first residual block and the second residual block, respectively.

**DenseNet-based network** The DenseNet-based network is composed of a convolutional layer, three dense blocks, two transition layers and a fully connected layer. Different from the ResNet-based network, the fully connected layer with 100 units is replaced with a global average pooling layer in this network. Each dense block consists of three units, and each unit contains a convolutional layer of 1×1×64 and a convolutional layer of $S \times S$×16. The filter size $S$ for each block is 5, 5 and 3, respectively. Two transition layers with a compression factor of 0.5 follow the first and the second dense blocks, respectively. Three max-pooling layers are inserted after the first convolutional layer and the two transition layers, respectively.

#### *3.1.2 Training Details*

All of the networks were trained with the Stochastic Gradient Descent (SGD) optimizer. The initial learning rate was set to 0.001. In training, when the average loss of the last ten epochs declined by less than 0.01 compared with that of the former ten epochs, the learning rate was decreased to 0.0001. When the loss decline became less than 0.001, the training was stopped. Cross entropy loss was utilized to train CNNs when BC learning was not employed, while Kullback-Leibler divergence loss was utilized for training with BC learning. All of the CNNs were trained with a mini-batch size of 64. In tri-training, the iteration was performed twice.

### 3.2 Dataset

As mentioned in Subsection 2.1, 815 gastric X-ray images with a resolution of 2048×2048 pixels from different patients provided by The University of Tokyo Hospital were used in the experiments. Specifically, 200 images including 100 gastritis images and 100 non-gastritis images were used as training data, and 615 images including 140 gastritis images and 475 non-gastritis images were used as test data. Some examples of the gastric X-ray images are shown in Fig. 3. In Fig. 3, (a)

and (b) are gastritis images and (c) and (d) are non-gastritis images. Patches cropped from the 200 images of training data include 45,127 patches of regions with gastritis, 42,785 patches of regions without gastritis and 48,385 patches of regions outside the stomach.

### 3.3 Evaluation Method and Metrics

Evaluation of a gastric X-ray image is conducted by using all of the three models to categorize all patches from the image. Firstly, predictions of the three models for a patch are averaged as the final prediction for the patch. Then categorization for an image is performed by a simple majority voting among patches predicted to be regions with and without gastritis from the image. In the majority voting, only patches with a high level of confidence are involved, making the diagnoses more accurate. Here, the confidence is denoted by the predicted probability. The threshold of the confidence was set to 0.7 in all experiments since the results with this threshold were more stable than the results using thresholds of 0.5, 0.6, 0.8 and 0.9. In evaluation, we used the original patches directly rather than performing BC learning.

We used sensitivity, specificity and harmonic mean of these two metrics as evaluation metrics for our method. The metrics are defined as follows.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{Harmonic mean} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}} \tag{5}$$

Here, TP, TN, FP and FN refer to true positive, true negative, false positive and false negative, respectively. Since there exists a trade-off relationship between sensitivity and specificity, the performance cannot be evaluated with sensitivity or specificity independently. The harmonic mean of these two metrics are utilized to evaluate the overall performance. In real-world clinical use, the balance between sensitivity and specificity can be regulated by adapting the voting system. For example, if the user places more importance on sensitivity, the majority voting can be adapted to a threshold-controlling voting that lets the gastritis patches win with a number less than the non-gastritis patches.

### 3.4 Experimental Results

The results obtained by our method when using 10, 20, 50, 100, 200 annotated images are shown in Table 1. The convergence process of our method when using 100 annotated images is shown in Table 2. As supplemental experiments, we also performed two sets of comparison experiments. Firstly, we compared our method with its ablated version, tri-training without BC learning, to certify the effectiveness of BC learning. The results are shown in Table 3. Table 4 shows the results of a comparison of our semi-supervised learning method and supervised learning with CNNs using no unannotated data. The results shown in Table 1 were obtained over 3 runs and the results shown in Table 2, Table 3 and Table 4 were obtained over 8 runs. The harmonic mean is expressed as the mean and standard deviation of the multiple runs. In each run, annotated images were randomly sampled from the 200 images of training data, but the ratio of gastritis and non-gastritis images was kept to 1:1. In the case of 200 annotated images, the models were trained in the form of supervised learning accompanied by BC learning instead of performing tri-training.

#### *3.4.1 Results Obtained by Our Semi-supervised Learning Method*

As shown in Table 1, our method achieved a high performance for diagnosing chronic gastritis even with an extremely small number of annotated images such as 10 and 20 images. Compared with supervised learning using 200 annotated images, almost the same performance could be realized by using our semi-supervised learning method with only half of the 200 annotated images. Additionally, it is clear that the performance can be boosted by increasing the number of annotated images. Table 2 shows the convergence process of our method when using 100 annotated images. The performances after the second iteration showed no further improvements.

#### *3.4.2 Results of Supplemental Experiments*

Table 3 and Table 4 show results using only 100 annotated images. We think that the conclusions based on results obtained by using different numbers of annotated images will remain unchanged. As shown in Table 3, our method, tri-training with BC learning, achieved better performance than its ablated version without the use of BC learning when using 100 annotated images. Therefore, it was confirmed that data augmentation with BC learning can greatly boost the performance of our semi-supervised learning method. As can be seen in Table 4, semi-supervised learning benefited from unannotated images and thus outperformed supervised learning when using 100 annotated images.

In addition, BC learning can also improve the performance of supervised learning. Student's t-test was performed for statistical significance analysis using results of 8 runs for BC learning and 8 runs for semi-supervised learning. For the contrast between using and not using BC learning shown in Table 3, we obtained a t-value of 2.73 and $p < 0.05$. However, for the contrast between semi-supervised learning and supervised learning shown in Table 4, the t-value was 1.63 and the p-value was less than 0.2, which represents a low level of significance. We think that this is because the number of annotated images was relatively large and the same as the number of unannotated images. In general, unannotated medical images are easier to collect than annotated medical images and thus the number of unannotated images is much larger. With more unannotated images, the promotion of performance attributed to semi-supervised learning can be more significant.

## 4 Discussion

We certified the effectiveness of our method with a series of experiments. As shown in Table 1, we confirmed that the models can gain greater capacity with a larger amount of labeled data. However, the improvement became very slight after the number of annotated images exceeded 50 since the total number of images of training data is always 200. Further improvement can be expected if there are more unannotated data in the training sets with an increase in the number of annotated images. Moreover, the standard deviation of the harmonic mean shown in Table 1 clearly decreased, indicating enhancement of stability of the models, with an increase in the number of annotated images. The main factor affecting the stability of the final models is the stability of the initial SVM models. Since chronic gastritis of different atrophic levels has relatively diverse features and the images may be of different quality for training a diagnosis model of chronic gastritis, the performances of initial models trained with a small number of randomly sampled images may vary significantly. Hence, the models can gain stronger stability from more annotated images. Note that the supervised learning also has a slight deviation since there is still randomness in the network initialization and training process. Also, the results presented in Table 3 and Table 4 show that BC learning can not only boost the overall performance but also stabilize the performance for both supervised learning and semi-supervised learning.

Some examples of TP, TN, FN and FP when training the models with our semi-supervised learning method using 100 annotated images are shown in Fig. 4∼Fig. 7, respectively. To illustrate how the regions contribute to the final diagnosis, we also show the probability heat maps of the whole images. For true positive and false positive images, the values in the heat maps represent the probabilities of gastritis. For true negative and false negative images, the values in the heat maps represent the probabilities of non-gastritis. As shown in the figures, the final decisions were mostly made by the regions inside the stomach, while the regions outside the stomach were assigned low probabilities for both gastritis and non-gastritis. Fig. 6 (a) shows a sample that is clinically difficult to diagnose, while Fig. 6 (b) includes massive barium sulfate that flowed out into the bowel so that many patches became noisy. In Fig. 7 (a), the textures did not appear clearly, which made the recognition more difficult. Similar to Fig. 6 (b), Fig. 7 (b) was also falsely diagnosed due to the barium sulfate flowing out. The above-mentioned problems can probably be solved by integrated diagnosis with multiple gastric X-ray images of a patient taken at different positions since images taken at other positions may have higher quality and better features for diagnosis.

In this paper, we showed how performances of automatic classification models based on supervised learning can be improved with a practical semi-supervised learning method in the task of gastritis classification using gastric X-ray images. In addition to this task, it is expected that our method can be used for various classification tasks in the field of medical image analysis due to its high ability of generalization and flexibility. Semi-supervised learning is innately fit for applications of medical image analysis since well-annotated medical images that can be used for training supervised learning models are always costly and scarce. With semi-supervised learning, unannotated data will not be wasted and can contribute to further improvements of performance. It is expected that utilization of unannotated data with semi-supervised learning will help in the construction of computer-aided systems with a high level of robustness.

There are some limitations in our study. Firstly, the data used for training and testing in the experiments were all obtained from the same medical facility. The reliability of our method for gastric X-ray images obtained from different facilities has not been confirmed. The method may need to be adapted for dealing with gastric X-ray images from different facilities, and that is one of our future works. Secondly, as mentioned above, the method fails to stabilize when the number of annotated images becomes small. Also, the method only performs a two-class classification between gastritis and non-gastritis. Stabilization of the method and realization of multi-grade diagnosis are also aims of our future works.

## 5 Conclusion

A semi-supervised learning method for chronic gastritis classification using gastric X-ray images is presented in this paper. The method has been developed with a tri-training architecture and a data augmentation method of BC learning. The high performance for diagnosis achieved by our method indicates its effectiveness and shows its promise for practical applications. The performance may be further improved by using more unlabeled data.

**Acknowledgements** The clinical data were acquired at The University of Tokyo Hospital in Japan. This study was partly supported by JSPS KAKENHI Grant number JP17H01744. We express our thanks to Katsuhiro Mabe of the Junpukai Health Maintenance Center and Nobutake Yamamichi of The University of Tokyo Hospital.

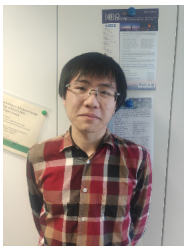
## Conflict of interest

The authors declare that they have no conflict of interest.

## References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 253–260. Springer (2017)
2. Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 311–319. Springer (2017)
3. Bechar, M.E.A., Settouti, N., Barra, V., Chikh, M.A.: Semi-supervised superpixel classification for medical images segmentation: application to detection of glaucoma disease. Multidimensional Systems and Signal Processing **29**(3), 979–998 (2018)
4. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
5. Chen, G., Zhang, J., Zhuo, D., Pan, Y., Pang, C.: Identification of pulmonary nodules via ct images with hierarchical fully convolutional networks. Medical & biological engineering & computing pp. 1–14 (2019)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee (2009)
8. Dheer, S., Levine, M., Redfern, R., Metz, D., Rubesin, S., Laufer, I.: Radiographically diagnosed antral gastritis: findings in patients with and without helicobacter pylori infection. The British journal of radiology **75**(898), 805–811 (2002)
9. Efron, B.: Bootstrap methods: another look at the jackknife. In: Breakthroughs in statistics, pp. 569–593. Springer (1992)
10. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)
11. Han, G., Liu, X., Zheng, G., Wang, M., Huang, S.: Automatic recognition of 3d ggo ct imaging signs through the fusion of hybrid resampling and layer-wise fine-tuning cnns. Medical & biological engineering & computing **56**(12), 2201–2212 (2018)
12. Hatipoglu, N., Bilgin, G.: Cell segmentation in histopathological images with deep learning algorithms by utilizing spatial relationships. Medical & biological engineering & computing **55**(10), 1829–1848 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
15. Kimura, K., Takemoto, T.: An endoscopic recognition of the atrophic border and its significance in chronic gastritis. Endoscopy **1**(03), 87–97 (1969)
16. Kudo, T., Kakizaki, S., Sohara, N., Onozato, Y., Okamura, S., Inui, Y., Mori, M.: Analysis of abc (d) stratification for screening patients with gastric cancer. World journal of gastroenterology: WJG **17**(43), 4793 (2011)
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015)
18. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
19. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on, pp. 565–571. IEEE (2016)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
21. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
22. She, Q., Hu, B., Luo, Z., Nguyen, T., Zhang, Y.: A hierarchical semi-supervised extreme learning machine method for eeg recognition. Medical & biological engineering & computing **57**(1), 147–157 (2019)
23. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging **35**(5), 1285–1298 (2016)
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)
25. Togo, R., Ishihara, K., Ogawa, T., Haseyama, M.: Estimation of salient regions related to chronic gastritis using

gastric x-ray images. Computers in biology and medicine **77**, 9–15 (2016)
26. Togo, R., Yamamichi, N., Mabe, K., Takahashi, Y., Takeuchi, C., Kato, M., Sakamoto, N., Ishihara, K., Ogawa, T., Haseyama, M.: Detection of gastritis by a deep convolutional neural network from double-contrast upper gastrointestinal barium x-ray radiography. Journal of gastroenterology pp. 1–9 (2018)
27. Tokozume, Y., Ushiku, Y., Harada, T.: Between-class learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5486–5494 (2018)
28. Vivanti, R., Joskowicz, L., Lev-Cohain, N., Ephrat, A., Sosna, J.: Patient-specific and global convolutional neural networks for robust automatic liver tumor delineation in follow-up ct studies. Medical & biological engineering & computing **56**(9), 1699–1713 (2018)
29. Yamamichi, N., Hirano, C., Ichinose, M., Takahashi, Y., Minatsuki, C., Matsuda, R., Nakayama, C., Shimamoto, T., Kodashima, S., Ono, S., et al.: Atrophic gastritis and enlarged gastric folds diagnosed by double-contrast upper gastrointestinal barium x-ray radiography are useful to predict future gastric cancer development based on the 3-year prospective observation. Gastric cancer **19**(3), 1016–1022 (2016)
30. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
31. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 408–416. Springer (2017)
32. Zhou, Z.H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on knowledge and Data Engineering **17**(11), 1529–1541 (2005)
33. Zhou, Z.H., Li, M.: Semi-supervised learning by disagreement. Knowledge and Information Systems **24**(3), 415–439 (2010)



**Zongyao Li** received his B.S. degree in Flight Vehicle Design and Engineering from Zhejiang University, China in 2017. He is currently pursuing an M.S. degree at the Graduate School of Information Science and Technology, Hokkaido University. His research interests are deep learning in medical image analysis and other applications. He is a student member of the IEEE.

**Ren Togo** received his B.S. degree in Health Sciences from Hokkaido University, Japan in 2015. He received his M.S. and Ph.D. degrees in Graduate School of Information Science and Technology, Hokkaido University, Japan in 2017 and 2019, respectively. He is also a radiological technologist. His research interests are machine learning and its applications. He is a member of the IEEE.

**Takahiro Ogawa** received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He is currently an assistant professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of ITE Transactions on Media Technology and Applications. He is a member of the IEEE, EURASIP, IEICE, and Institute of Image Information and Television Engineers (ITE).

**Miki Haseyama** received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 1995 to 1996. She is currently a professor in the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of ITE Transactions on Media Technology and Applications, a Director, International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE). She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITE) and Acoustical Society of Japan (ASJ).

**Table 1** Performances of our method using different numbers of annotated images.

| Number of annotated images | Sensitivity | Specificity | Harmonic mean |
|---|---|---|---|
| 10 | 0.857 | 0.864 | 0.860 ± 0.025 |
| 20 | 0.855 | 0.889 | 0.870 ± 0.016 |
| 50 | 0.871 | 0.945 | 0.906 ± 0.016 |
| 100 | 0.922 | 0.907 | 0.914 ± 0.001 |
| 200 | 0.893 | 0.953 | 0.922 ± 0.001 |

**Table 2** Convergence process of our method when using 100 annotated images.

| Model | Sensitivity | Specificity | Harmonic mean |
|---|---|---|---|
| SVM | 0.763 | 0.967 | $0.852 \pm 0.022$ |
| CNN (first iteration) | 0.892 | 0.910 | $0.901 \pm 0.013$ |
| CNN (second iteration) | 0.915 | 0.914 | $0.914 \pm 0.009$ |

**Table 3** Comparison of tri-training with and without BC learning using 100 annotated images.

| Method | Sensitivity | Specificity | Harmonic mean |
|---|---|---|---|
| Tri-training with BC learning (our method) | 0.915 | 0.914 | **0.914 ± 0.009** |
| Tri-training without BC learning | 0.812 | 0.963 | 0.880 ± 0.028 |

**Table 4** Comparison of our method and supervised learning using 100 annotated images.

| Method | Sensitivity | Specificity | Harmonic mean |
|---|---|---|---|
| Our method | 0.915 | 0.914 | **0.914 ± 0.009** |
| Supervised learning with BC learning | 0.891 | 0.919 | 0.903 ± 0.013 |
| Supervised learning without BC learning | 0.798 | 0.951 | 0.873 ± 0.018 |

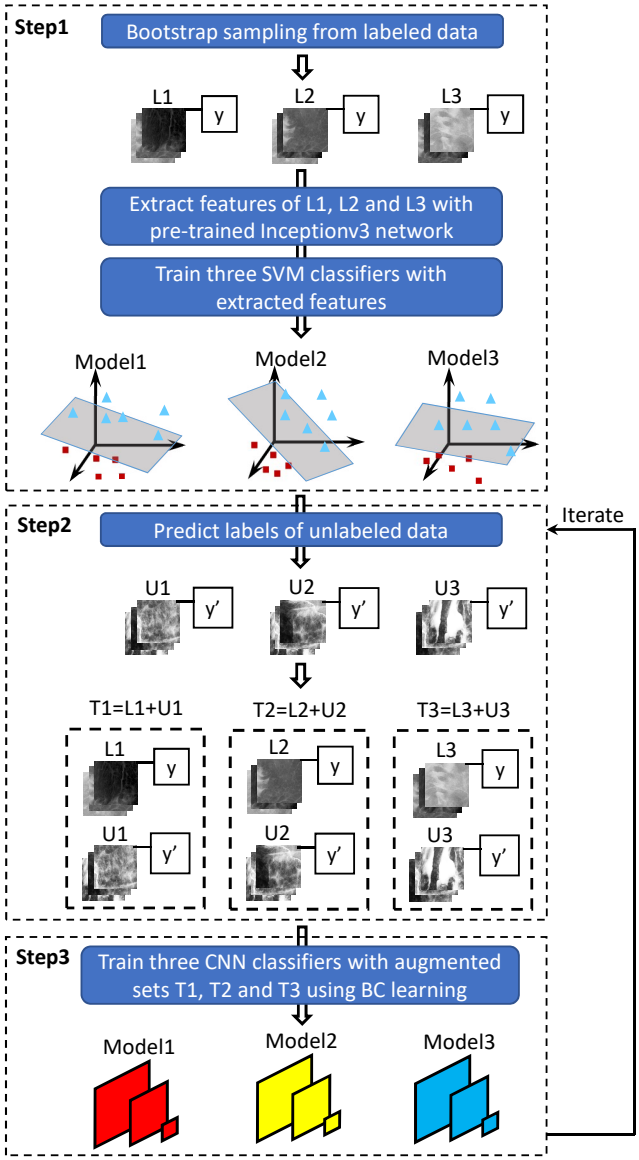



**Fig. 1** Procedures of the proposed method. 'L1', 'L2' and 'L3' are randomly sampled training sets from labeled data. 'U1', 'U2' and 'U3' are selected from unlabeled data. Specifically, 'U1' is the set of samples for which 'Model2' and 'Model3' agree, 'U2' is the set of samples for which 'Model1' and 'Model3' agree, and 'U3' is the set of samples for which 'Model1' and 'Model2' agree. 'y' denotes ground truth labels, and '$y'$' denotes predicted labels.

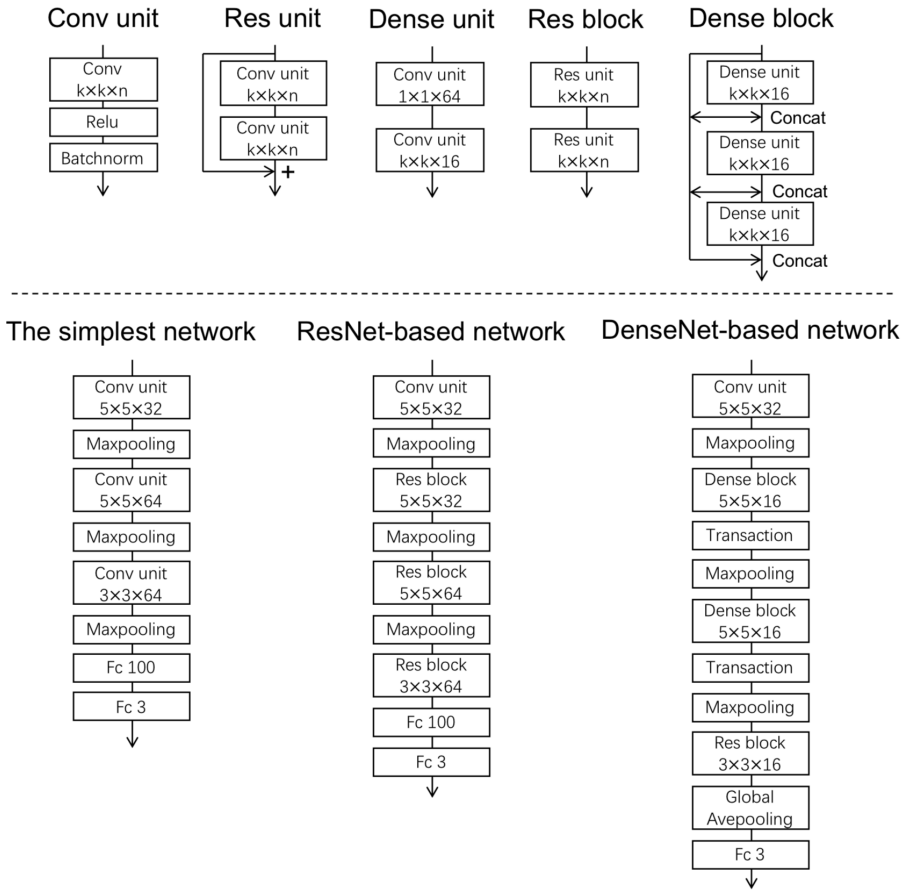



**Fig. 2** Architectures of three CNNs used in our method.

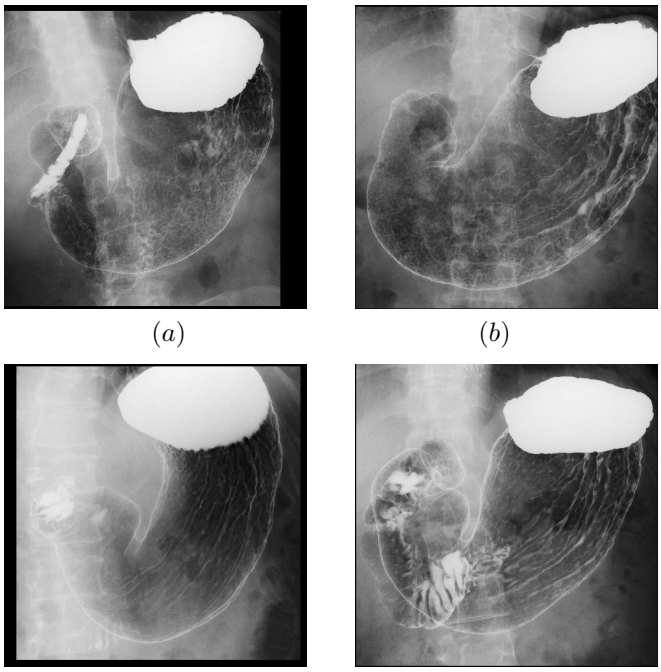



**Fig. 3** Examples of gastric X-ray images: (a) and (b) are gastritis images and (c) and (d) are non-gastritis images.

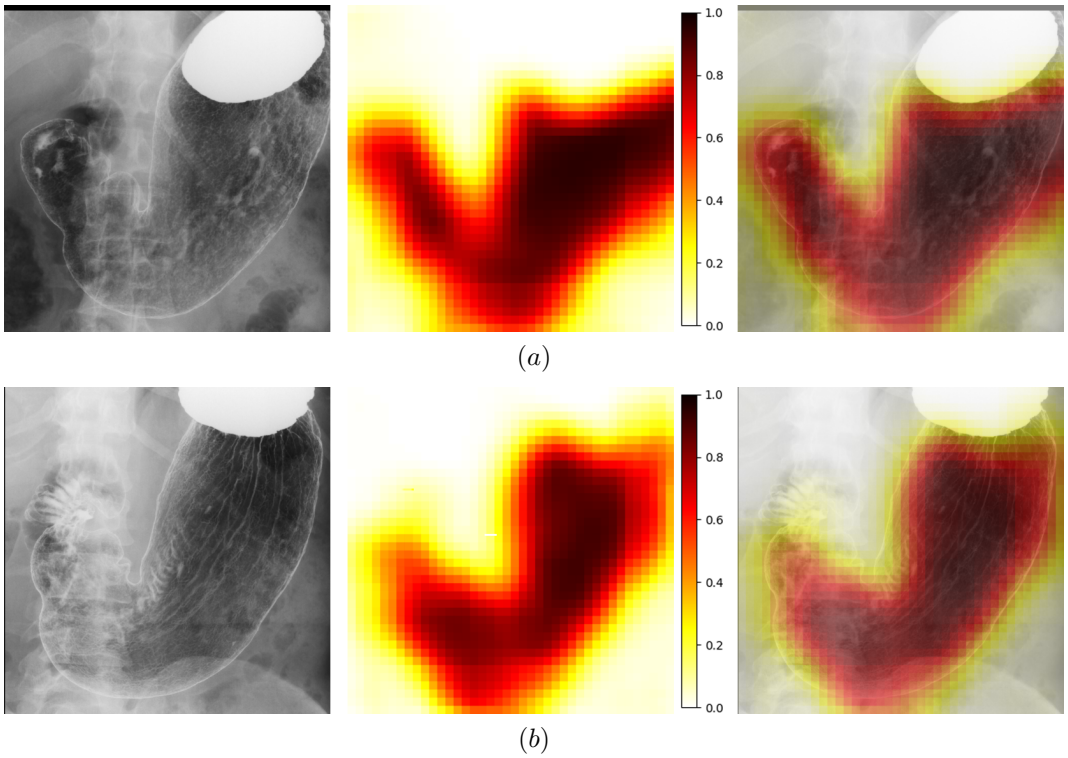



**Fig. 4** Examples of true positives when using 100 annotated images. Left column: gastric X-ray images. Middle column: heat maps of the probability for gastritis. Right column: overlapped images.

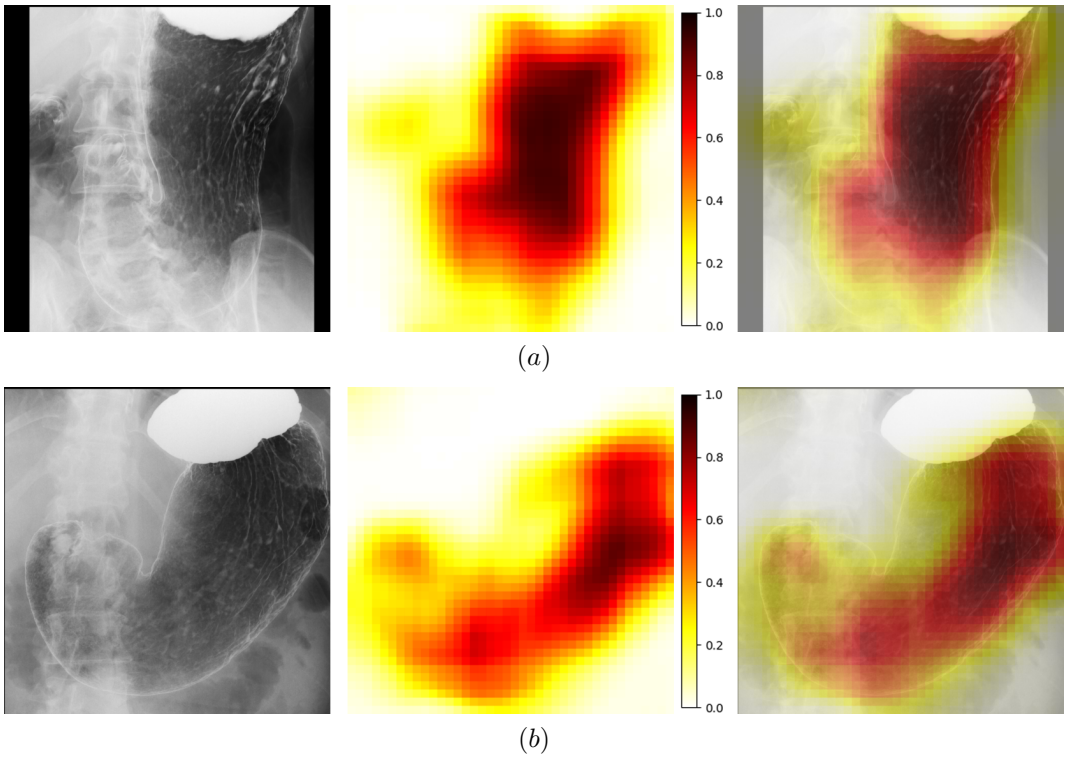



**Fig. 5** Examples of true negatives when using 100 annotated images. Left column: gastric X-ray images. Middle column: heat maps of the probability for non-gastritis. Right column: overlapped images.

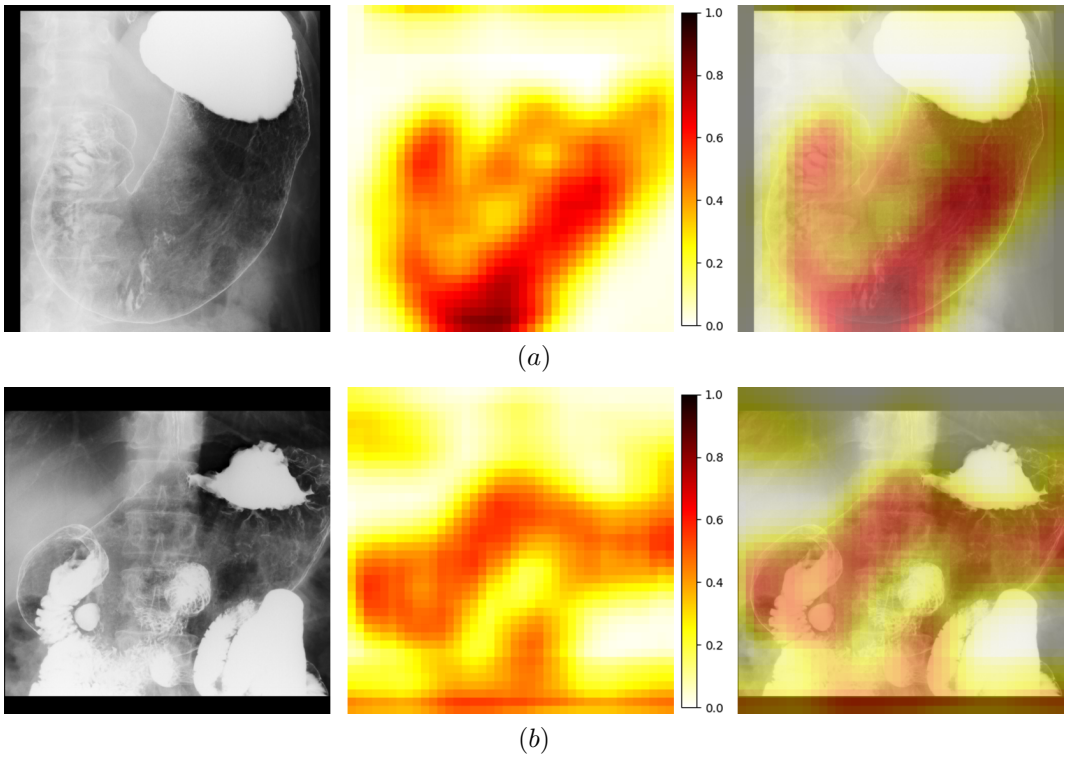



**Fig. 6** Examples of false negatives when using 100 annotated images. Left column: gastric X-ray images. Middle column: heat maps of the probability for non-gastritis. Right column: overlapped images.

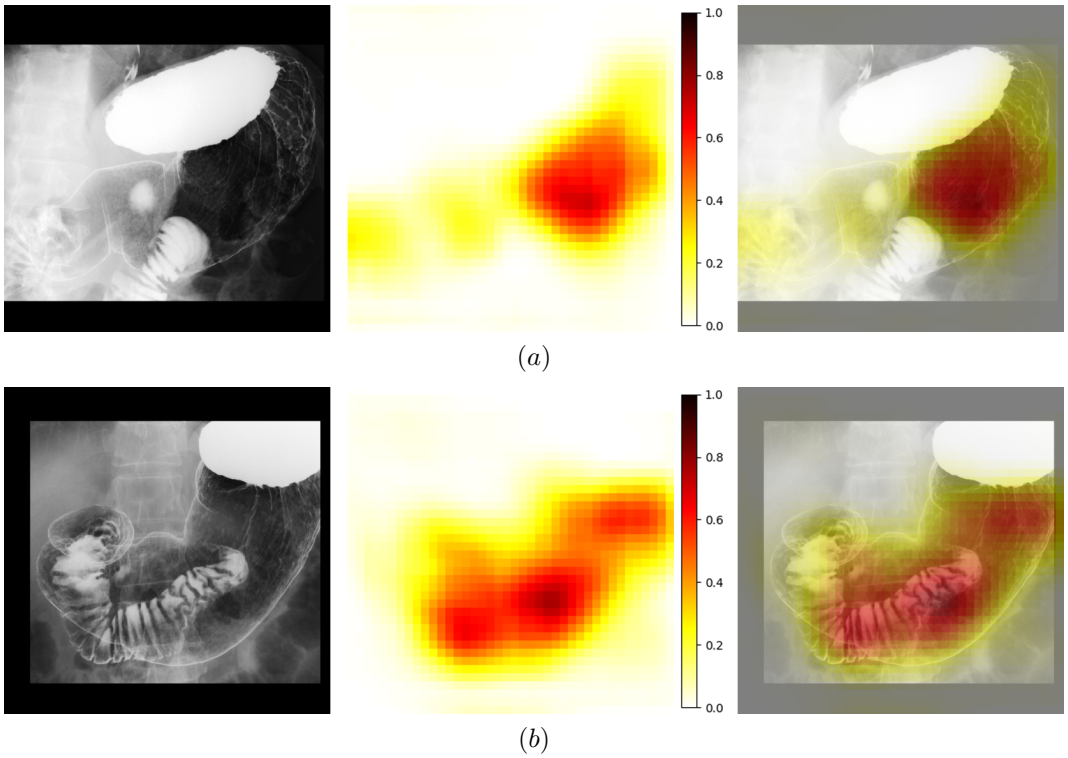



**Fig. 7** Examples of false positives when using 100 annotated images. Left column: gastric X-ray images. Middle column: heat maps of the probability for gastritis. Right column: overlapped images.