Computers in Biology and Medicine 123 (2020) 103903

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep convolutional neural network-based anomaly detection for organ classification in gastric X-ray examination

Ren Togo [a,*], Haruna Watanabe [b], Takahiro Ogawa [b], Miki Haseyama [b]

[a] *Education and Research Center for Mathematical and Data Science, Hokkaido University, N-12, W-7, Kita-ku, Sapporo, 060-0812, Japan*

[b] *Faculty of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-Ku, Sapporo, 060-0814, Japan*

ARTICLE INFO

Keywords:
Deep learning
Medical image analysis
Gastric X-ray examination
Esophagus
Stomach
Anomaly detection
Autoencoder

ABSTRACT

Aim: The aim of this study was to determine whether our deep convolutional neural network-based anomaly detection model can distinguish differences in esophagus images and stomach images obtained from gastric X-ray examinations.

Methods: A total of 6012 subjects were analyzed as our study subjects. Since the number of esophagus X-ray images is much smaller than the number of gastric X-ray images taken in X-ray examinations, we took an anomaly detection approach to realize the task of organ classification. We constructed a deep autoencoding gaussian mixture model (DAGMM) with a convolutional autoencoder architecture. The trained model can produce an anomaly score for a given test X-ray image. For comparison, the original DAGMM, AnoGAN, and a One-Class Support Vector Machine (OCSVM) that were trained with features obtained by a pre-trained Inception-v3 network were used.

Results: Sensitivity, specificity, and the calculated harmonic mean of the proposed method were 0.956, 0.980, and 0.968, respectively. Those of the original DAGMM were 0.932, 0.883, and 0.907, respectively. Those of AnoGAN were 0.835, 0.833, and 0.834, respectively, and those of OCSVM were 0.932, 0.935, and 0.934, respectively. Experimental results showed the effectiveness of the proposed method for an organ classification task.

Conclusion: Our deep convolutional neural network-based anomaly detection model has shown the potential for clinical use in organ classification.

1. Introduction

In medical fields, diagnostic imaging techniques using such as X-ray and endoscopy have become popular. Medical images are used for the early detection of serious diseases. However, interpretation of the images requires much effort and specialized knowledge. Hence, to support doctors' diagnostic work, diagnostic supporting systems based on artificial intelligence (AI) technologies have been studied [1–4]. Deep convolutional neural networks (DCNNs) [5] have been attracting much attention since their recognition performance is better than that of other conventional machine learning techniques using manually designed features [6–9].

The main focuses of AI-based medical image analyses are disease classification, disease detection, and segmentation tasks. The disease classification task is a technique to classify medical images into a positive class if there is a disease in the target data or a negative class. The disease detection task is a technique to detect a region where a disease has occurred in a medical image. Generally, this is a complicated task since the differences in the target disease and non-related regions in an image should be recognized. The segmentation task is a technique to divide medical images into regions of organs or tissues. Nowadays, pixel-level segmentation can be realized on the basis of well-annotated large-scale medical image datasets.

The quantity and quality of data in the dataset are important factors for performing the above tasks with high level of accuracy [10,11]. If the dataset contains images other than those of the target for the task, it can cause significant degradation of performance. For example, when building a dataset for constructing a model to perform a segmentation task for a given organ, it is difficult for the model to accurately learn the region of the target organ if the training data include data for organs other than the target organ. Therefore, when introducing data-driven approaches such as deep learning, we need to pay attention to dataset construction. Many of the currently available medical image datasets

[*] Corresponding author.
E-mail addresses: togo@lmd.ist.hokudai.ac.jp (R. Togo), haruna@lmd.ist.hokudai.ac.jp (H. Watanabe), ogawa@lmd.ist.hokudai.ac.jp (T. Ogawa), miki@ist.hokudai.ac.jp (M. Haseyama).

https://doi.org/10.1016/j.compbiomed.2020.103903
Received 3 April 2020; Received in revised form 18 June 2020; Accepted 29 June 2020
Available online 8 July 2020

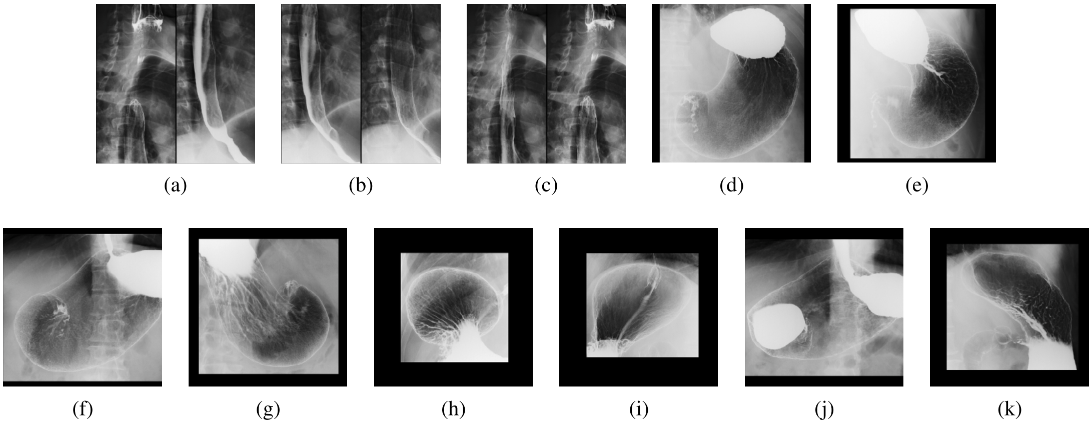


Fig. 1. Examples of images taken in gastric X-ray examinations. (a)–(c) are esophagus X-ray images and (d)–(k) are gastric X-ray images.

are manually modified and annotated, and it takes a lot of effort to construct a large dataset with good quality. However, there have been very few studies that focused on the construction of datasets.

Our previous works revealed that deep learning-based approaches are useful for the task of chronic atrophic gastritis classification using gastric X-ray images [12,13]. The symptoms of chronic atrophic gastritis are shown in gastric X-ray images as subtle changes only in the inside of the stomach. Therefore, we divided an X-ray image into small patches and estimated the presence of gastritis for each patch in our previously proposed method. Our previous analysis using a total of 6,520 gastric X-ray images for 815 subjects has already achieved high performance (The sensitivity, specificity, and harmonic mean were 0.962, 0.983, and 0.972, respectively.).

In our previous method, a dataset including only gastric X-ray images was used for the chronic atrophic gastritis classification task [12, 13]. However, in a gastric X-ray examination, organs and tissues other than the stomach can be targeted for the imaging simultaneously. The images of these organs are treated as a series of the same examination results. In other words, a gastric X-ray examination always includes images of other organs that are not necessary for the training of a chronic atrophic gastritis classification model. In the case of constructing an AI-based disease classification model, images of other organs or tissues can become noise that can cause deterioration of classification performance. Therefore, when constructing the dataset for chronic atrophic gastritis classification, X-ray images of organs and tissues other than the stomach need to be removed. In our previous studies, the clinical application issue was not considered; images of other organs were manually excluded from the dataset, a task that took a significant amount of time and labor. Construction of high-quality datasets is thus the next challenging task for realizing fully automated AI-based supporting systems. Our previous chronic atrophic gastritis classification deep learning model can distinguish the presence of gastritis, non-gastritis, and regions outside the stomach at the patch level. Although this method was able to recognize subtle differences in the presence or absence of gastritis at the patch level, it was difficult to apply it directly to organ classification tasks that require an understanding of the overall features in the image. Hence, new methods that can automatically recognize the stomach and other organs are desired.

In this study, we aimed to realize a high-quality dataset construction method, namely, an automated organ classification method. Images taken in a gastric X-ray examination include esophagus images, duodenum images, and images of regions other than stomach regions. These images are not necessary for the chronic atrophic gastritis classification task. Besides, since the number of these images is much smaller than the number of stomach images, it is difficult to obtain data on such organs for model training. Hence, we take an anomaly detection approach [14] that is effective when using an unbalanced dataset. We introduce a deep learning-based anomaly detection model as our organ classification task and expand it to recognize the characteristics of gastric X-ray images as normal images. In our method, a large number of gastric X-ray images are used as normal images for the classifier training, and the small number of esophagus images are detected as abnormal images. We show the effectiveness of our method through several experiments.

Contributions of this paper are summarized as follows:

- We propose a new automated organ classification method based on an anomaly detection approach and show its effectiveness through several experiments.
- This approach can contribute to enhancing the efficiency of dataset construction, which is necessary for training and evaluating AI models in medical image analysis.

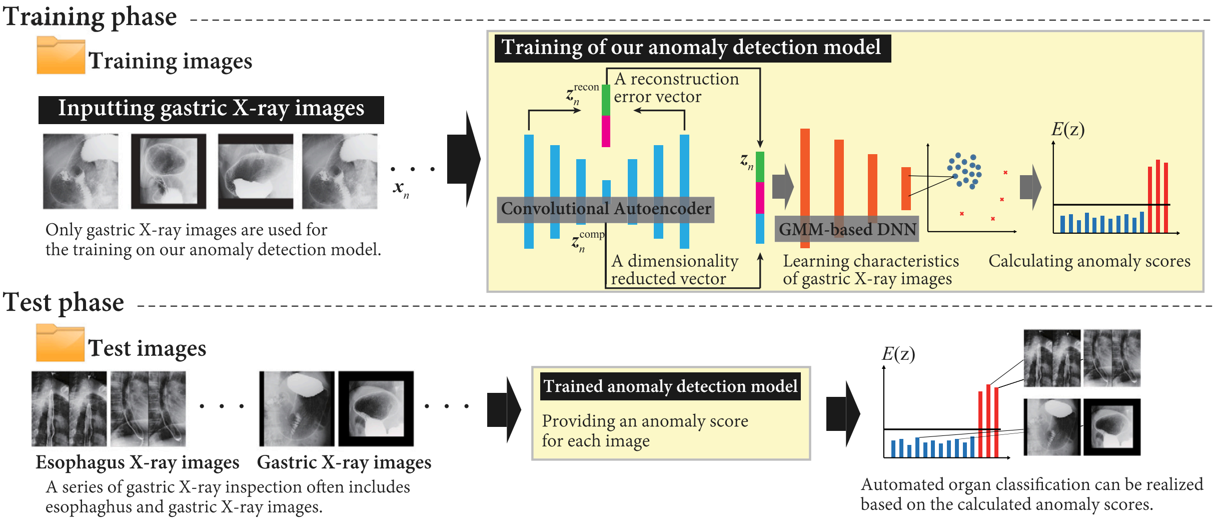
2. Methods

X-ray images of our study subjects are shown in Section 2.1. An anomaly detection model for organ classification is presented in Section 2.2. Finally, statistical analyses and comparative methods are explained in Section 2.3. This study was reviewed and approved by the institutional review board of The University of Tokyo. Data were completely anonymized prior to analysis.

2.1. *Study subjects*

Our target was X-ray images taken in gastric X-ray examinations for the diagnosis of chronic atrophic gastritis. All of the X-ray images were 16-bit gray-scale and 2048 × 2048 pixels. The images always include small numbers of esophagus images that are not used for the diagnosis of chronic atrophic gastritis.

An example of stomach and esophagus X-ray images obtained by X-ray examination of a patient for whom images were used in this study is shown in Fig. 1. Figs. 1 (a)–(c) show esophagus X-ray images taken in the gastric X-ray examination to investigate the condition of the esophagus. These images were taken in early stage of the examination. Figs. 1 (d)–(k) show gastric X-ray images taken from different angles. Gastric X-ray images taken in the following eight imaging positions were used in this study: double-contrast frontal view in the supine (Fig. 1(d)), double-contrast right anterior oblique view in the near-supine (Fig. 1(e)), double-contrast left anterior oblique view in the near-supine (Fig. 1(f)), double-contrast frontal view in the prone (Fig. 1(g)), double-contrast frontal view in the prone (Fig. 1(h)), double-contrast left lateral view in the horizontal (Fig. 1(i)), double-contrast left anterior oblique view in the near-supine (Fig. 1(j)), and double-contrast right anterior oblique view in the near-supine (Fig. 1(k)).


Training phase
Training images
Inputting gastric X-ray images
x_n
Training of our anomaly detection model
z_n^{recon}
A reconstruction error vector
Convolutional Autoencoder
z_n^{comp}
A dimensionality reducted vector
z_n
GMM-based DNN
Learning characteristics of gastric X-ray images
$E(z)$
Calculating anomaly scores
Test phase
Test images
Esophagus X-ray images
Gastric X-ray images
Trained anomaly detection model
Providing an anomaly score for each image
$E(z)$

Only gastric X-ray images are used for the training on our anomaly detection model.

A series of gastric X-ray inspection often includes esophaghus and gastric X-ray images.

Automated organ classification can be realized based on the calculated anomaly scores.

Fig. 2. Overview of our anomaly detection method for organ classification.

2.2. *Anomaly detection model for organ classification*

Fig. 2 shows an overview of our proposed method. An anomaly detection approach is used to automatically exclude esophagus X-ray images from the obtained data. Our method consists of two procedures, training phase and test phase. Each procedure is explained in detail in the following subsections.

2.2.1. *Training phase*

The proposed method consists of a state-of-the-art anomaly detection method, deep autoencoding gaussian mixture model (DAGMM) [15]. DAGMM has two deep neural networks, a compression network and an estimation network. The compression network performs dimensionality reduction and reconstruction of input samples with an autoencoder. The estimation network obtains a membership probability in the gaussian mixture model (GMM) using low-dimensional representations. In the original DAGMM [15], a deep autoencoder [16] model is used in the compression network. The deep autoencoder was intended for low-dimensional datasets such as the KDDCUP dataset [17], Thyroid dataset [17], and Arrhythmia dataset [17]. However, since X-ray images are high-dimensional data, an approach that differs from the original DAGMM model is needed. Therefore, we newly introduce a convolutional autoencoder [18] model for obtaining high-dimensional representation of X-ray images.

Formally, let $\boldsymbol{x}_n$ $(n = 1, 2, \cdots, N;$ N being the number of training images) represent gastric X-ray images for training. Our anomaly detection model's goal is to learn the characteristics of $\boldsymbol{x}_n$ and provide an anomaly score to an input test image. A compression network for dimension reduction is used to obtain sophisticated features for $\boldsymbol{x}_n$. The compressed features and the reconstruction error features are combined as low-dimensional representations, and they are treated as input vectors of the estimation network. The compression network calculates a dimensionality reduction vector $\boldsymbol{z}_n^{\mathrm{comp}}$ from the convolutional autoencoder architecture and errors between a reconstruction image $\boldsymbol{x}'_n$ and an input sample $\boldsymbol{x}_n$ with a convolutional autoencoder. The images are resized to 224 × 224 pixels when they are inputted into the convolutional autoencoder. From the reconstruction image $\boldsymbol{x}'_n$, we obtain a reconstruction error vector $\boldsymbol{z}_n^{\mathrm{recon}}$ as follows:

$$\boldsymbol{z}_n^{\mathrm{recon}} = \left[\frac{\|\boldsymbol{x}_n - \boldsymbol{x}'_n\|_2}{\|\boldsymbol{x}_n\|_2}, \frac{\boldsymbol{x}_n^{\top}\boldsymbol{x}'_n}{\|\boldsymbol{x}_n\|_2\|\boldsymbol{x}'_n\|_2} \right]^{\top}. \tag{1}$$

The compressed vector $\boldsymbol{z}_n^{\mathrm{comp}}$ and the reconstruction error vector $\boldsymbol{z}_n^{\mathrm{recon}}$ are combined as a low-dimensional compressed feature representation $\boldsymbol{z}_n$ as follows:

$$\boldsymbol{z}_n = \left[\boldsymbol{z}_n^{\mathrm{comp}\top}, \boldsymbol{z}_n^{\mathrm{recon}\top} \right]^{\top}. \tag{2}$$

By coupling two feature vectors, the low-dimensional representation $\boldsymbol{z}_n$ can be considered to have valuable information of $\boldsymbol{x}_n$.

Next, the extracted low-dimensional representation $\boldsymbol{z}_n$ is used as an input vector for the estimation network constructed with a deep neural network (DNN), and we obtain a membership probability γ_n under the framework of the GMM as follows:

$$\boldsymbol{p}_n = \mathrm{D}(\boldsymbol{z}_n), \tag{3}$$

$$\gamma_n = \mathrm{softmax}(\boldsymbol{p}_n), \tag{4}$$

where D is a DNN, $\boldsymbol{p}_n$ is the output vector of a DNN, and γ_n is the output value of a softmax function. Given a batch of N and its membership predictions $\forall 1 \le k \le K$, the parameters of GMM are as follows:

$$\hat{\phi}_k = \sum_{n=1}^{N} \frac{\hat{\gamma}_{nk}}{N}, \tag{5}$$

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} \hat{\gamma}_{nk}\boldsymbol{z}_n}{\sum_{n=1}^{N} \hat{\gamma}_{nk}}, \tag{6}$$

$$\hat{\Sigma}_k = \frac{\sum_{n=1}^{N} \hat{\gamma}_{nk}(\boldsymbol{z}_n - \hat{\mu}_k)(\boldsymbol{z}_n - \hat{\mu}_k)^{\top}}{\sum_{n=1}^{N} \hat{\gamma}_{nk}}, \tag{7}$$

where $\hat{\phi}_k$, $\hat{\mu}_k$, and $\hat{\Sigma}_k$ are respectively mixture probability, mean value, and co-variance value for the component k of GMM respectively, and $\hat{\gamma}_n$ is the membership probability for the low-dimensional representation $\boldsymbol{z}_n$. From the above, the sample's energy can be calculated as follows:

$$E(\boldsymbol{z}_n) = -\log\left(\sum_{k=1}^{K} \hat{\phi}_k \frac{\exp(-\frac{1}{2}(\boldsymbol{z}_n - \hat{\mu}_k)^{\top}\hat{\Sigma}_k^{-1}(\boldsymbol{z}_n - \hat{\mu}_k))}{2\pi\sqrt{|\hat{\Sigma}_k|}} \right), \tag{8}$$

where $|\cdot|$ donates the determinant of a matrix. In the test phase, if the calculated sample energy is high, this image is classified as another organ group.

According to our DAGMM-based anomaly detection model, the parameters of the two networks can be trained by minimizing the following objective function:

$$J = \frac{1}{N}\sum_{n=1}^{N} L(\boldsymbol{x}_n, \boldsymbol{x}'_n) + \frac{\lambda_1}{N}\sum_{n=1}^{N} E(\boldsymbol{z}_n) + \lambda_2 P(\hat{\Sigma}), \tag{9}$$

where $L(\boldsymbol{x}_n, \boldsymbol{x}'_n)$ is a reconstruction error that can be used by the convolutional autoencoder in the compression network. If the calculated reconstruction error L becomes low, it means that low-dimensional representation can preserve the key information of input samples. Therefore, the reconstruction error L is expected to be always low. In this study, L_2-norm is used as the reconstruction error. Next, according to the second term $E(\boldsymbol{z}_n)$, the learned model produces low energy for the input sample of gastric X-ray images. On the other hand, the model produces high energy for the input sample of esophagus X-ray images. The third term P is a penalty term to avoid the singularity problem in the GMM. We penalize small values on the diagonal entries by P.

2.2.2. Test phase

The trained model can calculate the energy that represents anomaly scores to test X-ray images. Since this model is trained with data for gastric X-ray images, if a given image is a gastric X-ray image, the anomaly score becomes low. If the given image is an esophagus X-ray image, the anomaly score becomes high. In the test phase, all of the X-ray images that are resized to 224 × 224 pixels are inputted into the trained model, and each image's energy is calculated. By defining the degree of abnormality using a specific threshold value ξ, we can classify the test X-ray images into stomach images and images of other organs. Specifically, we determine the classification of images as follows:

$$E(\boldsymbol{z}) = \begin{cases} \text{stomach} & (\text{if } E < \xi) \\ \text{other organs} & (\text{otherwise}), \end{cases} \tag{10}$$

where $\boldsymbol{z}$ is a low-dimensional representation of a test image. In this way, it becomes possible to classify X-ray images into images of the stomach and images of other organs.

2.3. Statistical analyses and comparative methods

From the data for 6012 subjects, a training dataset containing data for 5,912 subjects, a validation dataset containing data for 50 subjects, and a test dataset containing data for 50 subjects were constructed as shown in Fig. 3. Note that the validation and test datasets were constructed from subjects that included a complete set of images for the standard eight imaging positions as shown in Fig. 1. The imaging positions of the validation and test datasets were reviewed by a clinician who has specialized knowledge of gastritis diagnosis. The training dataset had 47,212 gastric X-ray images, the validation dataset had 400 gastric and 197 esophagus X-ray images, and the test dataset had 400 gastric and 206 esophagus X-ray images. We trained our anomaly detection model with the training dataset and decided the threshold value ξ by the validation dataset. We classified the test dataset based on the threshold value ξ and evaluated our model. Sensitivity (Sen), specificity (Spe), and the harmonic mean of sensitivity and specificity (HM) were calculated for our evaluation. These criteria are defined as follows:

$$\text{Sen} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}, \tag{11}$$

$$\text{Spe} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}, \tag{12}$$

$$\text{HM} = \frac{2 \times \text{Sen} \times \text{Spe}}{\text{Sen} + \text{Spe}}, \tag{13}$$

For comparison, the following three comparative methods were used. One-Class Support Vector Machine (OCSVM) [19] is one of the popular anomaly detection models. We extracted features using a pre-trained Inception-v3 network model [20], and we trained the OCSVM-based classifier with extracted features. Moreover, to confirm the effectiveness of using the convolutional autoencoder, we also used a method with extracted features as input vectors of the DAGMM as a comparison method. Finally, a deep learning model, AnoGAN [21], was used as a comparative method. AnoGAN is a popular deep learning-based anomaly detection models for medical image classification tasks.

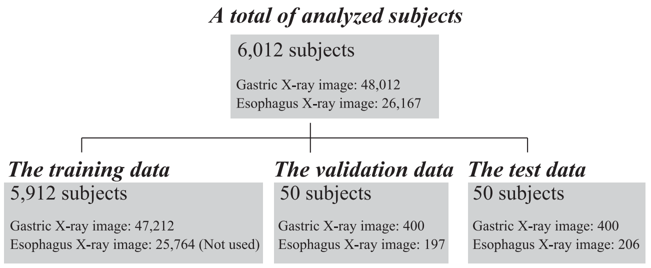

A total of analyzed subjects
6,012 subjects
Gastric X-ray image: 48,012
Esophagus X-ray image: 26,167
The training data
5,912 subjects
Gastric X-ray image: 47,212
Esophagus X-ray image: 25,764 (Not used)
The validation data
50 subjects
Gastric X-ray image: 400
Esophagus X-ray image: 197
The test data
50 subjects
Gastric X-ray image: 400
Esophagus X-ray image: 206

Fig. 3. Dataset construction flowchart in this study.

Table 1
Hyper-parameters of our model used in the experiment.

Parameter	Value
Learning rate	0.0001
λ_1	0.1
λ_2	0.01
Batch size	128
Epoch	200
Encoded dimensions	300
GMM mixtures	7

3. Results

Experiments were conducted on a Linux operating system (Ubuntu 18.04; Canonical, London, England) with the Keras framework and a single NVIDIA GeForce RTX 2080 Ti GPU. Hyper-parameters of our model used in this experiment are shown in Table 1. These hyper-parameters were determined by the validation dataset. Also, the threshold of normal/abnormal was determined to be $\xi = -13.9$ by the validation dataset.

3.1. Performance evaluation

The classification performances of our anomaly detection model and the comparative methods are shown in Table 2. Sen, Spe, and HM of our method were 0.956, 0.980, and 0.968, respectively. For comparative methods, Sen, Spe, and HM of the baseline DAGMM-based anomaly detection method were 0.932, 0.883, and 0.907, respectively, and those of the OCSVM-based anomaly detection method were 0.932, 0.935, and 0.934, respectively. By comparing the classification performances of our method and the OCSVM-based method, we confirmed that deep learning-based anomaly detection showed better performance than that of the classical anomaly detection model. Also, by comparing the classification performances of our method and the DAGMM-based method, our new network architecture, including convolutional architectures, was shown to be useful for improving the classification performance. The popular deep learning model AnoGAN had the lowest classification performance among the methods used in this experiment.

Next, we show the results of the confusion matrix of our method. The confusion matrix represents information about actual and predicted classification results obtained by the classification model. In 606 test images, the number of true positive, true negative, false positive, and false negative images for our method were 391, 198, 9 and 8, respectively, as shown in Fig. 5. From the results, we can see that our model can correctly recognize the differences between gastric X-ray images and esophagus X-ray images for almost all of the test images.

More specific evaluation results focusing on an individual examination are shown in Fig. 6. Fig. 6 shows calculated energies from a series of gastric X-ray examinations of a patient. Sample energies of (a)–(k) correspond to the images of (a)–(k) in Fig. 1. Therefore, images (a)–(c) represent esophagus X-ray images and the others represent gastric X-ray images. As described above, the threshold of normal/abnormal was determined to be $\xi = -13.9$ by the validation dataset. From the

Table 2
Classification performances of our method and comparative methods.

	Sen	Spe	HM
Our method	**0.956**	**0.980**	**0.968**
DAGMM [15]	0.932	0.883	0.907
AnoGAN [21]	0.835	0.833	0.834
OCSVM [19]	0.932	0.935	0.934

Table 3
Classification performances of our method and comparative methods for different dataset division.

	Sen	Spe	HM
Our method	0.937	**0.941**	**0.939**
DAGMM [15]	0.881	0.875	0.878
AnoGAN [21]	0.452	0.480	0.466
OCSVM [19]	**0.945**	0.904	0.924

		Actual class	
		Positive	Negative
Predicted class	Positive	397	9
	Negative	8	198

Fig. 4. Examples of images classified correctly and incorrectly by our method.

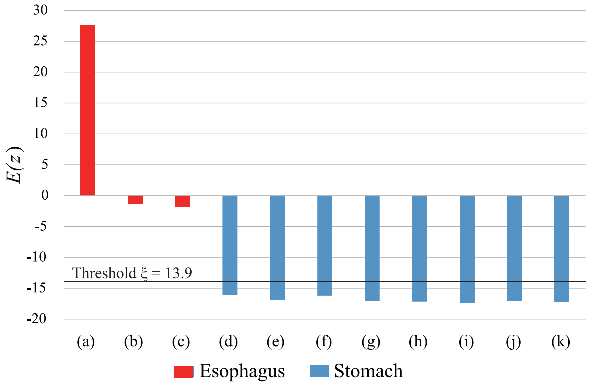
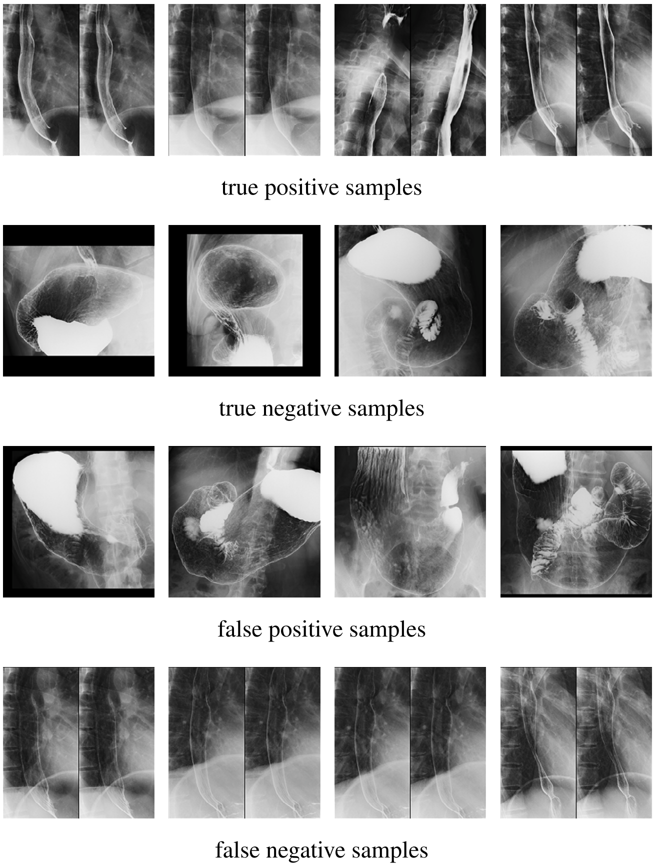


Fig. 5. Confusion matrix obtained by our method.



true positive samples

true negative samples

false positive samples

false negative samples

Fig. 6. Energies of the images in Fig. 1 calculated by our model.

results, we can see that anomaly scores of esophagus X-ray images are higher than those of gastric X-ray images.

Finally, we show samples that were correctly and incorrectly classified by our method in Fig. 4. As shown in Fig. 4, true positive and true negative samples were images that clearly showed organ characteristics. Although our method achieved high classification performance, there were some incorrectly classified samples. Most of the false positive samples (gastric X-ray images that had high anomaly scores) were images of the double-contrast frontal view of the stomach in the prone position with the head down. Moreover, some of false negative samples (esophagus X-ray images that had low anomaly scores) were images in which the esophagus was magnified.

Overall, our method achieved better performance than that of other benchmark anomaly detection methods. The effectiveness of our method was confirmed by the test dataset that included 606 images of 50 subjects.

3.2. Additional evaluation with different dataset division

Dataset bias affects the evaluation of performance of models. Therefore, we divided the dataset into different proportions than those used in the main experiment described above to verify the classification performance of the model. From the data for 6012 subjects, a training dataset containing data for 5012 subjects, a validation dataset containing data for 200 subjects, and a test dataset containing data for 800 subjects were randomly sampled. It should be noted that the test data in this experiment did not include test data used in the main experiment. As a result, the training dataset had 39,999 gastric X-ray images, the validation dataset had 1599 gastric and 897 esophagus X-ray images, and the test dataset had 6398 gastric and 3484 esophagus X-ray images.

Classification performances of our method and comparative methods for the different dataset divisions are shown in Table 3. Following the primary evaluation of the main dataset division, our method outperformed the comparative methods in this evaluation. We confirmed that the performance of the recently proposed AnoGAN significantly dcreased for the different dataset division. On the other hand, other methods including our method showed robustness for the different dataset division.

4. Discussion

We discuss our contributions to technical and clinical fields in the following subsections.

4.1. Contributions to technical fields

We have proposed a method for organ classification in images taken in gastric X-ray examinations using an anomaly detection approach as the first step in the construction of a high-quality dataset for realizing accurate classification of gastritis. Unsupervised anomaly detection is useful when using unbalanced datasets such as images of X-ray examinations because abnormal samples can be detected by the classifier trained with only normal samples. Unsupervised anomaly detection has been actively researched [14]. Previously reported unsupervised anomaly detection methods can be grouped into three categories: reconstruction-based approaches, clustering analysis, and one-class classification.

Reconstruction-based approaches perform dimensionality reduction and reconstruction of samples and detect an anomaly sample from the reconstruction error. Conventional methods in this category often use principal component analysis (PCA) [22], kernel PCA [23] and robust PCA [24]. Moreover, recent studies have shown that analysis of the reconstruction error induced by a deep autoencoder is useful [25,26].

Clustering analysis is another popular anomaly detection approach based on multivariate gaussian models, gaussian mixture models, and k-means [27–30]. Because of the curse of dimensionality, it is difficult to apply such methods directly to high-resolution image data. Conventional methods consist of two steps [14]: dimensionality reduction and clustering analysis. Training is conducted separately in the two steps. In other words, dimensionality reduction is trained without guidance from the subsequent clustering analysis. Therefore, key information for clustering analysis might be lost during dimensionality reduction.

One-class classification approaches are also often used for anomaly detection. These approaches use a discriminative boundary surrounding the normal instances that are trained by algorithms such as OCSVM [19, 31,32]. When targeting high-resolution images, high performance of these techniques usually cannot be expected because of the curse of dimensionality.

The recently proposed deep learning-based anomaly detection model DAGMM can detect anomaly samples based on reconstruction errors using compressed features. A DAGMM-based model solves the problem of key information being lost in conventional clustering analysis by training dimensionality reduction and clustering analysis at the same time. DAGMM also estimates data density in a low-dimensional representation for more robust anomaly detection of high-dimensional data, unlike one-class classification approaches. However, although anomaly detection approaches have attracted attention in the medical field [33], this work is, to the best of our knowledge, the first work in which high performance for organ classification based on a DAGMM-based architecture was realized. Our experimental results showed that the DAGMM-based model is useful for organ classification in a gastric X-ray examination.

4.2. Contributions to clinical fields

Although there are several methods used for assessment of stomach conditions such as blood tests, biopsy, and endoscopy, a gastric X-ray examination is still the most widely used and most effective method for evaluation of stomach conditions since it enables direct observation of the stomach [34]. Many gastric X-ray examinations can be performed in one day, and they are therefore suitable for mass screening in East Asian countries [35]. However, the number of clinicians who are specialized in the diagnosis of chronic atrophic gastritis from gastric X-ray images has decreased due to the diversification of inspection approaches. Hence, the introduction of AI-based supporting systems is crucial in this field [36].

As found in our previous works, we revealed that deep learning-based chronic atrophic gastritis classification approaches can achieve high classification performance at the clinical level. However, these techniques still cannot be used in clinical applications. One of the main reasons for this obstacle is dataset construction. Our organ classification model, based on an anomaly detection approach, tackled this challenging problem and achieved high classification performance for dataset construction. This approach can reduce the labor required for dataset preparation.

In medical image analysis for clinical applications, anomaly detection approaches are more suitable than supervised learning approaches in many cases. Unlike general images, medical images have confidential information, and annotation for them requires specialized knowledge. Therefore, methods for the construction of datasets to train AI-based methods with as little effort as possible are needed. Anomaly detection does not require annotations for abnormal images since it can learn the characteristics of normal images. The effectiveness of an anomaly detection approach for organ classification was shown in this paper.

Our approach will also be useful unbalanced data classification problems and rare disease detection problems. Each medical facility has various types of data, such as examination equipment, number of examinations, and number of cases, and it may be difficult to obtain uniform data for supervised learning. In the field of medical screening, most test results are often negative, and positive data are often overwhelmingly less than negative data. Although it is difficult to apply general supervised learning in such a situation, the anomaly detection approach makes it possible to construct a diagnostic support system using only negative data. In addition, for rare diseases, an abnormality detection-based diagnostic support system can be applied to mitigate the risk of overlooking by clinicians without collecting their data.

Our study has several limitations. The proposed method was evaluated on the gastric X-ray images reviewed by clinicians. However, in actual gastric X-ray examinations, a small number of stomach images with leakage of barium, which are not useful for gastritis detection, are also included. For clinical applications, such images should also be removed when constructing the dataset for gastritis detection. Therefore, we classified images that are useful for detecting gastritis from all gastric X-ray images taken during gastric X-ray examinations. This is one of the issues that need to be addressed in the future.

5. Conclusion

We have presented a deep convolutional neural network-based anomaly detection method for organ classification in images taken in gastric X-ray examinations. We have proposed a new anomaly detection architecture inspired by DAGMM for the realization of automated classification of gastric and esophagus X-ray images. Experiments using data for 6012 subjects showed the effectiveness of our method. Our approach will contribute to database formatting for supervised machine learning in medical image analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

In this work, data were provided by The University of Tokyo Hospital in Japan. We express our thanks to Katsuhiro Mabe of the Junpukai Health Maintenance Center, and Nobutake Yamamichi of The University of Tokyo. This study was partly supported by JSPS KAKENHI, Japan under Grant JP20K19857, JP17H01744, and JP18K11367.

References

[1] J.L. Seixas, S. Barbon, R.G. Mantovani, Pattern recognition of lower member skin ulcers in medical images with machine learning algorithms, in: Proceedings of the IEEE International Symposium on Computer-Based Medical Systems, CBMS, 2015, pp. 50–53.

[2] A. van Opbroek, M.A. Ikram, M.W. Vernooij, M. de Bruijne, Transfer learning improves supervised image segmentation across imaging protocols, IEEE Trans. Med. Imaging 34 (5) (2015) 1018–1030.

[3] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S.C.H. Hoi, M. Satyanarayanan, A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 30–44.

[4] B. Tanoori, Z. Azimifar, A. Shakibafar, S. Katebi, Brain volumetry: An active contour model-based segmentation followed by SVM-based classification, Comput. Biol. Med. 41 (8) (2011) 619–632.

[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the International Conference on Neural Information Processing Systems, NeurIPS, 2012, pp. 1097–1105.

[6] K. Hatano, S. Murakami, H. Lu, J. Kooi Tan, H. Kim, T. Aoki, Classification of osteoporosis from phalanges CR images based on DCNN, in: Proceedings of the International Conference on Control, Automation and Systems, ICCAS, 2017, pp. 1593–1596.

[7] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[8] Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M.R. Harowicz, M.A. Mazurowski, Deep learning for identifying radiogenomic associations in breast cancer, Comput. Biol. Med. 109 (2019) 85–90.

[9] S.M. Mathews, C. Kambhamettu, K.E. Barner, A novel application of deep learning for single-lead ecg classification, Comput. Biol. Med. 99 (2018) 53–62.

[10] L.B. Holder, M.M. Haque, M.K. Skinner, Machine learning for epigenetics and future medical applications, Epigenetics 12 (7) (2017) 505–514.

[11] K. Lan, D.-t. Wang, S. Fong, L.-s. Liu, K.K. Wong, N. Dey, A survey of data mining and deep learning in bioinformatics, J. Med. Syst. 42 (8) (2018) 139.

[12] R. Togo, N. Yamamichi, K. Mabe, Y. Takahashi, C. Takeuchi, M. Kato, N. Sakamoto, K. Ishihara, T. Ogawa, M. Haseyama, Detection of gastritis by a deep convolutional neural network from double-contrast upper gastrointestinal barium x-ray radiography, J. Gastroenterol. 54 (4) (2019) 321–329.

[13] M. Kanai, R. Togo, T. Ogawa, M. Haseyama, Gastritis detection from gastric x-ray images via fine-tuning of patch-based deep convolutional neural network, in: Proceedings of the IEEE International Conference on Image Processing, ICIP, 2019, pp. 1371–1375.

[14] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (2009) 1–58.

[15] B. Zong, Q. Song, M.R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding gaussian mixture model for unsupervised anomaly detection, in: Proceedings of the International Conference on Learning Representations, ICLR, 2018, pp. 1–19.

[16] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[17] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2017, http://archive.ics.uci.edu/ml.

[18] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-Encoders for hierarchical feature extraction, in: Proceedings of the International Conrefence on Artificial Neural Networks and Machine Learning, ANN, 2011, pp. 52–59.

[19] Y. Chen, X.S. Zhou, T.S. Huang, One-class SVM for learning in image retrieval, in: Proceedings of the International Conference on Image Processing, ICIP, 2001, pp. 34–37.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Processings of the IEEE International Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 2818–2826.

[21] T. Schlegl, P. Seeböck, S. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: Proceedings of the International Conference on Information Processing in Medical Imaging, IPMI, 2017, pp. 146–157.

[22] I. Jolliffe, Principal Component Analysis, Springer Verlag, 1986.

[23] S. Günter, N.N. Schraudolph, S.V.N. Vishwanathan, Fast iterative kernel principal component analysis, J. Mach. Learn. Res. 8 (2007) 1893–1918.

[24] P.J. Huber, Robust Statistics, vol. 523, John Wiley & Sons, 2004.

[25] C. Zhou, R.C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 665–674.

[26] S. Zhai, Y. Cheng, W. Lu, Z. Zhang, Deep structured energy based models for anomaly detection, in: Proceedings of the International Conference on International Conference on Machine Learning, ICML, 2016, pp. 1100–1109.

[27] V. Barnett, T. Lewis, Outliers In Statistical Data, second ed., John Wiley & Sons Ltd., 1978.

[28] A. Zimek, E. Schubert, H.-P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min. 5 (5) (2012) 363–387.

[29] J. Kim, C.D. Scott, Robust kernel density estimation, J. Mach. Learn. Res. 13 (Sep) (2012) 2529–2565.

[30] L. Xiong, B. Póczos, J. Schneider, Group anomaly detection using flexible genre models, in: Proceedings of the International Conference on Neural Information Processing Systems, NeurIPS, 2011, pp. 1071–1079.

[31] Q. Song, W. Hu, W. Xie, Robust support vector machine with bullet hole image classification, IEEE Trans. Syst. Man Cybern. C 32 (4) (2002) 440–448.

[32] G. Williams, R. Baxter, H. He, S. Hawkins, L. Gu, A comparative study of RNN for outlier detection in data mining, in: Proceedings of the IEEE International Conference on Data Mining, ICDM, 2002, pp. 709–712.

[33] K. Gupta, A. Bhavsar, A.K. Sao, Detecting mitotic cells in HEp-2 images as anomalies via one class classifier, Comput. Biol. Med. 111 (2019) 103328.

[34] N. Yamamichi, C. Hirano, Y. Takahashi, C. Minatsuki, C. Nakayama, R. Matsuda, T. Shimamoto, C. Takeuchi, S. Kodashima, S. Ono, et al., Comparative analysis of upper gastrointestinal endoscopy, double-contrast upper gastrointestinal barium x-ray radiography, and the titer of serum anti-helicobacter pylori igg focusing on the diagnosis of atrophic gastritis, Gastric Cancer 19 (2) (2016) 670–675.

[35] K. Sugano, Screening of gastric cancer in asia, Best Pract. Res. Clin. Gastroenterol. 29 (6) (2015) 895–905.

[36] W.G. e Gonçalves, M.H.d.P. dos Santos, F.M.F. Lobato, Â. Ribeiro-dos Santos, G.S. de Araújo, Deep learning in gastric tissue diseases: a systematic review, BMJ Open Gastroenterol. 7 (1) (2020) e000371.

Ren Togo: received the B.S. degree in health sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is also a Radiological Technologist. He is currently a Specially Appointed Assistant Professor with the Education and Research Center for Mathematical and Data Science, Hokkaido University. His research interests include machine learning and its applications. He is a member of the IEEE.

Haruna Watanabe received her B.S. degree from the Faculty of Engineering, Iwate University, Japan in 2018. She received her M.S. degree from the Graduate School of Information Science and Technology, Hokkaido University, in 2020. Her research interests are medical image analysis and machine learning techniques. She is currently working at NS Solutions Corporation.

Takahiro Ogawa: received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include AI, the IoT, and big data analysis for multimedia signal processing and its applications. He is a member of ACM, IEICE, and ITE. He was a Special Session Chair of the IEEE ISCE2009, a Doctoral Symposium Chair of ACM ICMR2018, an Organized Session Chair of the IEEE GCCE 2017–2019, a TPC Vice Chair of the IEEE GCCE2018, a Conference Chair of the IEEE GCCE2019, and so on. He has also been an Associate Editor of ITE Transactions on Media Technology and Applications.

Miki Haseyama: received the 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 with Washington University, USA, from 1995 to 1996. She is currently a Professor with 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 is a member of the IEICE, the Institute of Image Information and Television Engineers (ITE), and the Acoustical Society of Japan (ASJ). She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), the 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). N-14, W-9, Kita-ku, Sapporo, 060-0814, JAPAN