Analysis of the effects of image quality differences on CAD performance in AI-based benign-malignant discrimination processing of breast masses

Kazuya Abe, Soma Kudo, Hideya Takeo, Yuichi Nagai, Shigeru Nawano

Kazuya Abe, Soma Kudo, Hideya Takeo, Yuichi Nagai, Shigeru Nawano, "Analysis of the effects of image quality differences on CAD performance in AI-based benign-malignant discrimination processing of breast masses," Proc. SPIE 12286, 16th International Workshop on Breast Imaging (IWBI2022), 122860R (13 July 2022); doi: 10.1117/12.2623398

SPIE.

Event: Sixteenth International Workshop on Breast Imaging, 2022, Leuven, Belgium

Analysis of the Effects of Image Quality Differences on CAD Performance in AI-Based Benign-Malignant Discrimination Processing of Breast Masses

Kazuya Abe[a], Soma Kudo[a], Hideya Takeo[a], Yuichi Nagai[b], Shigeru Nawano[c]
[a] Kanagawa Institute of Technology, Electrical and Electronic Engineering Dept.
[b] National Cancer Center Hospital East
[c] Shinmatsudo Central General Hospital

ABSTRACT

In recent years, the amount of images to be read has increased due to the higher resolution of diagnostic imaging devices, and the burden on doctors has also increased. To solve this problem, the improvement of CAD (computer-aided diagnosis) performance has been studied. In this study, we developed an AI-based system for discriminating benign and malignant breast cancer tumors using transfer learning, one of the deep learning methods of AI, and analyzed what factors are necessary to improve the diagnostic accuracy of the system. Classification of benign and malignant diseases using diagnostic images showed an accuracy of 90%, which was equivalent to physician's discrimination, but the accuracy for medical checkup images was low at 85%, and image comparison revealed that this was due to noise and low contrast. We analyzed that these improvements are necessary for the construction of a more accurate CAD system for medical checkup images.
Keywords: Breast Masses, Screening, Diagnosis, CAD, Benign and Malignant classification

1. INTRODUCTION

Recent years have seen an increase in the amount of medical images generated in medical applications due to resolution improvements to CT, MRI, and other imaging devices. Because the number of images taken per patient may reach into the hundreds, the load on medical practitioners interpreting these images is rising sharply. Computer-Aided Diagnosis/Detection (CAD) is therefore viewed with considerable importance as a means of alleviating this situation.

In the field of breast cancer benign-malignant discrimination, past CAD research results have included the development of various CAD systems. These include the method by Kenichi Inoue[1], by which an Alexnet-based, high-performance CAD system that achieves a 96% proper diagnosis rate with diagnostic images was built; the serration shape region detection method of Nakagawa et. al.[2], a system capable of giving medical practitioners clear instructions about the basis of the results of discrimination using mass border shape recognition; and the method of Fukuoka et. al.[3], a system that, utilizing images of Japanese people as an AI training database, quantifies the diagnostic criteria used in discrimination performed by medical practitioners as features for an artificial neural network (ANN).

Mammography, an essential tool in the detection of breast cancer, comprises x-ray imaging. Research by Takeshi Iinuma[4], however, has shown that x-ray imaging incurs a higher risk of exposure and reduced lifespan to young people.

The ability to perform breast cancer benign-malignant discrimination at the screening mammography stage would make it possible to go immediately to biopsy without having to conduct diagnostic mammography bi-directional imaging for benign-malignant discrimination.

With a view towards realizing this, in this study, we analyze the effects on CAD performance of image quality differences in breast cancer examination images. We then identifying points for improvement, and based on them derive the type of image processing necessary for making possible AI-based breast cancer benign-malignant discrimination that utilizes breast cancer screening mammographic images.

16th International Workshop on Breast Imaging (IWBI2022), edited by Hilde Bosmans, Nicholas Marshall, Chantal Van Ongeval, Proc. of SPIE Vol. 12286, 122860R

2. DIFFERENCES BETWEEN SCREENING AND DIAGNOSTIC MAMMOGRAPHIC IMAGES

2.1 Evaluating the image quality of x-ray images used for diagnosis

X-ray images used for diagnosis are evaluated according to image contrast, which is the difference or ratio between bright and dark areas; sharpness, a physical measurement that indicates image detail clarity; and graininess, which is random granular texture caused by unevenness of film density.

X-ray imaging characteristics include the loss of image contrast due to the effects of radiation scattering and the loss of graininess due to noise.

In image processing, graininess is negatively correlated with clarity and contrast. Sharpness and contrast are positively correlated.

2.2 Differences between screening and diagnostic mammographic images

In mammography, a distinction between screening and diagnostic is required. Screening mammography is performed during breast cancer screening and diagnostic mammography is performed at detailed-examination clinics or outpatient clinics that specialize in mammary glands. The reason for this distinction is as follows:

In screening mammography, single- or bi-directional mammographic imaging is performed. It comprises the examination of images in order to determine the management of whether or not a detailed examination is needed by looking for findings indicative of the possibility of breast cancer in asymptomatic individuals. Diagnostic mammography, however, comprises the examination of diagnostic breast images in order to determine the management of the necessity of performing a biopsy and of follow-up based on the results of bi-directional imaging and special roentgenography (such as enlarged spot) performed on symptomatic patients and those determined by screening to require a detailed examination.[5]

Because of differences between these imaging methods, in the case of diagnostic images, more mammographic images of different types will be taken for the same patient than in the case of screening images, thus enabling more accurate benign-malignant discrimination.

3. ISSUE AND PURPOSE

3.1 Issue

By serving as a second opinion, CAD holds the promise of decreasing the number of medical accidents and lightening the burden on medical practitioners by reducing physician oversights and cutting the time required for diagnosis. A pre-condition for this, however, is that the trustworthiness of CAD must be established by raising its accuracy.

3.2 Purpose

In this study, we develop breast cancer benign-malignant discrimination processing using two databases, one of screening images and one of diagnostic images, and verify discrimination performance. Our purpose is to compare screening and diagnostic (i.e. second phase detailed examination) detection performance and analyze the diagnostic and screening images for causes of any differences that are found.

4. TEST METHOD

4.1 Overview

In this study, we build a breast cancer mass benign-malignant discrimination processing system with two types of databases, one comprising breast cancer mass screening images and the other breast cancer mass diagnostic images, and compare proper diagnosis rates. MATLAB was used to build the CAD system.

Next, we use image processing to compare the contrast and graininess of the screening and diagnostic images in order to find possible causes for differences in CAD performance.

4.2 Detailed procedure

As AI training images, we prepared 250 images each of benign tumors and malignant tumors with the tumor borders cut out. Tests were then conducted using the following method. The nth degree was set to 5.

The following 11 types of AI-based breast cancer tumor benign-malignant discrimination processing systems were used:

1. Perceptron and 3-layer NN learning method (basic machine learning methods)
2. CNN with 1 to 6 middle layers (deep learning methods)
3. VGG16, VGG19, Alexnet (three types of transfer learning that apply a model trained in one area to another area)

In addition, the screening image and diagnostic image databases were further divided into two using the image extensions described below and learning conducted.

1. Methods without image extensions, which use unmodified images for learning.
2. Images with previously-adjusted rotation, brightness, and contrast were added to the learning data and learning conducted (image extensions present).

The following describes the method for comparing screening image and diagnostic images.

4.2.1 Detailed procedure

Fourier transformation was applied to images with a mass region of 5×5 cm cut out (Figure 1, left).

Low frequencies were removed by setting to 0 the signal for the center of transformed images (Figure 1, center).

The amount of noise was determined by subjecting this image (Figure 1, right) to inverse Fourier transformation and examining the signal histogram.

The circle in the center is half of the length and width of the image.

Regarding the quality of image graininess when using this method, the amount of image noise can be said to decrease as the standard deviation of the above histogram (variation in the high frequency regions of the original image) decreases.

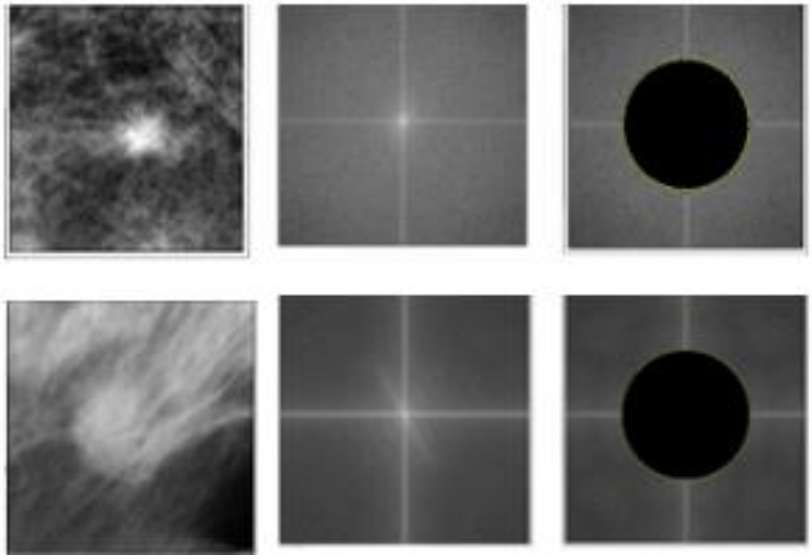


Figure 1 Effects of radiation scattering

4.2.2 Measurement of contrast (estimation of amount of radiation scattering)

The image density histogram was examined and contrast determined according to the size of the standard deviation value.

The contrast is regarded as good to the extent that the standard deviation of the image density histogram (pixel variation in the original image) is high.

5. TEST RESULTS

5.1 Test results of AI training using screening images

Table 1 below shows the test results for mass benign-malignant discrimination processing when using screening images.

5.2 Test results of AI training using diagnostic images

Table 2 shows the test results when using diagnostic images

The training network that used VGG16 and images with extensions exhibited the highest proper diagnosis rate. This indicates the successful development of AI that can produce a proper diagnosis rate of 80 to 90%, which is the same range as that of medical specialists.

Table 3 shows the results with highest accuracy among screening and diagnostic images with and without image extensions.

At 85%, Alexnet with image extensions exhibited the maximum accuracy for screening images. At 90%, VGG16 with image extensions exhibited the maximum accuracy for diagnostic images. Diagnostic images had 5% higher accuracy than screening images.

Table 1 Results of breast cancer mass benign-malignant discrimination processing using screening images

	Accuracy (%)	
	Without extensions	With extensions
Perceptron	54	59
3 layer NN	53	64
CNN (middle 1)	57	68
CNN (middle 2)	55	69
CNN (middle 3)	60	69
CNN (middle 4)	60	71
CNN (middle 5)	59	71
CNN (middle 6)	60	73
VGG16	63	80
VGG19	64	83
Alexnet	64	85

Table 2 Results of breast cancer mass benign-malignant discrimination processing using diagnostic images

	Accuracy (%)	
	Without extensions	With extensions
Perceptron	69	60
3 layer NN	75	68
CNN (middle 1)	71	70
CNN (middle 2)	66	65
CNN (middle 3)	73	70
CNN (middle 4)	73	78
CNN (middle 5)	77	82
CNN (middle 6)	68	80
VGG16	73	90
VGG19	69	80
Alexnet	69	83

Table 3 Results of AI-based benign-malignant discrimination processing that exhibit maximum accuracy

Maximum accuracy for screening images		Maximum accuracy for diagnostic images	
No extensions	With extensions	No extensions	With extensions
64% VGG19 Alexnet	85% Alexnet	77% CNN (middle 5)	90% VGG16

5.3 Screening image and diagnostic image comparison results

5.3.1 Comparison results of graininess

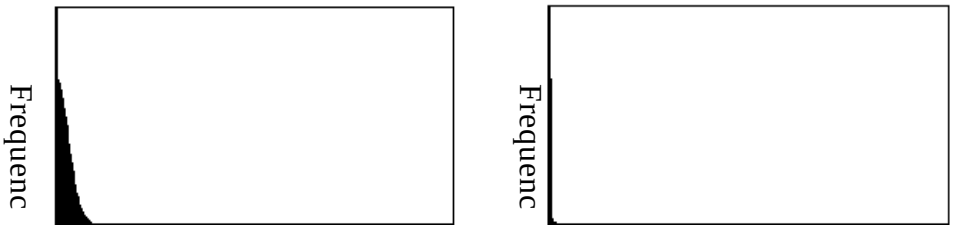
Using the method described in 4.2a), we compared the graininess of four cases for screening and diagnostic images. Figure 2 shows histograms that record signal level and its frequency for two cases, one for screening images and one for diagnostic images. The average standard deviation was 2.7225 for screening images and 0.5855 or diagnostic images.

The standard deviation of screening images was about four times that of diagnostic images, indicating significantly greater amounts of image noise.

5.3.2 Comparison results of contrast (amount of radiation scattering)

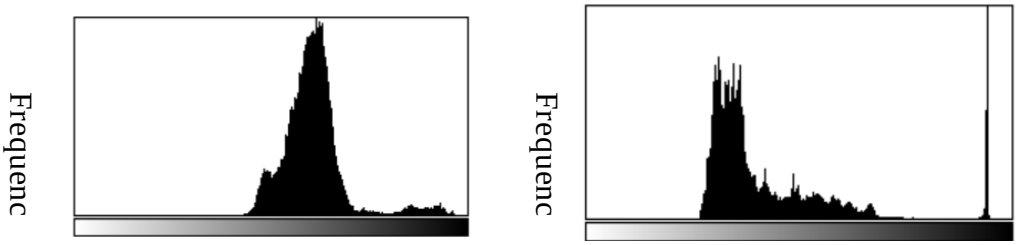
Using the method described in 4.2b), we compared the contrast of four cases for screening and diagnostic images. Figure 3 shows histograms that record density level and its frequency for two cases, one for screening images and one for diagnostic images. The average standard deviation was 17.485 for screening images and 38.20075 for diagnostic images.

The average standard deviation for diagnostic images was about twice that for screening images, indicating that screening images had lower contrast than diagnostic images. This suggests that screening images had greater amounts of radiation scattering.



（a）Screening image **（b）Diagnostic image**

Figure 2 Graininess test results cases



（a）Screening image **（b）Diagnostic image**

Figure 3 Contrast test results cases

6. CONSIDERATIONS

Regarding causes affecting the accuracy of the AI-based benign-malignant discrimination processing of screening and diagnostic images, the results of our analysis of images showed that the average standard deviation (the quality of image graininess) obtained in the noise testing of screening and diagnostic images indicated that screening images had about four times the noise of diagnostic images and half the contrast. This combination of a high level of noise and poor contrast is therefore a likely factor negatively affecting training accuracy.

These differences are attributable to the differences in image capture equipment used in screening and diagnosis. Image processing and advances in equipment are required to improve this situation. However, just eliminating noise through image processing in order to improve graininess is expected to reduce contrast and thereby cause image blurring.

In addition, many incorrectly-judged images had calcifications in them. For such cases, combined use with an AI learning network for breast cancer calcifications has the potential to deliver benign-malignant discrimination processing with higher accuracy.

7. CONCLUSIONS AND ISSUES

In this study, in order to indicate the effects of differences between screening images and diagnostic images on CAD systems as a value, i.e. a proper diagnosis rate, rather than configure a system that, based on the serration shape region detection method of Nakagawa et. al.[2], is capable of clearly instructing medical practitioners on the basis of the results of discrimination that relies on mass border shape recognition, or build a CAD system that, based on the method of Fukuoka et. al.[3], visualizes features, we instead built a CAD system that generates high performance through transfer learning that is based on the method by Kenichi Inoue.[2]

As a result, we succeeded in deriving proper diagnosis rate values that are near those obtained in image interpretation by medical practitioners at 85% for screening images and 90% for diagnostic images. We also succeeded in analyzing the differences in the proper diagnosis rates, analyzing the differences between screening and diagnostic images through measurements of radiation scattering and graininess that used image spatial frequency Fourier transform-based harmonic elimination.

However, in order to correct the difference with the diagnostic images, it will be necessary to combine the use of image processing with AI learning networks for calcifications, which will be a future issue.

As discussed in 6. Considerations, as graininess and contrast are negatively correlated in image processing, it is necessary to use image processing that can balance graininess and contrast.

Solving this problem will require performing image processing that supports the simultaneous improvement of radiation scattering and graininess and then performing training using images thus generated.

One such image processing method is Virtual Grid, which was developed by Fujifilm. Virtual Grid technology consists of contrast improvement processing and graininess improvement technology for improving both image contrast loss and graininess loss caused by x-ray scattering within the image subject.[6] To solve the issues identified in this study, we are currently studying image quality improvement processing while referencing this technology.

REFERENCES

[1] Kenichi Inoue, Autodetection of Mammography Using Convolutional Neural Network, 6th JAMI & JSAI AIM Joint Research Meeting, 2018

[2] Toshiaki Nakagawa, Hiroyuki Sakurai, Takeshi Hara, et al, Development of Automatic Classification System for Mammographic Masses, Transactions of the Japanese Society for Medical and Biological Engineering :BME 43(3) , 2005, p 437-446.

[3] Daisuke FUKUOKA, Takeshi HARA, Hiroshi FUJITA, et al, Development of an Automated Method for Classification of Masses in Mammogram, Japanese journal of medical electronics and biological engineering 39-1, 2001, p24-29.

[4] Takeshi IINUMA, Risk-benefit analysis of 2-year interval mammography screening, 14th Annual Meeting of the Japanese Association of Breast Tumor Screening, 2004

[5] The Japanese Breast Cancer Society ed, Manual for preparing a detailed breast cancer screening report based on screening categories and diagnostic categories, KANEHARA&CO., LTD, 2019

[6] Kawamura, Takahiro, Naito, Satoshi, Okano, Kayo, et al, Improvement in Image Quality and Workflow of X-Ray Examinations using a new Image Processing Method, ”Virtual Grid Technology”, FUJIFILM RESEARCH & DEVELOPMENT No.60, 2015, p21-27