Data Acquisition and Processing of Breast Cancer Assisted Diagnosis Based on Ultrasound Imaging

Gu Yunchao
School of Computer Science and Engineer, Beihang University
No.37 Xueyuan Road, Haidian District, Beijing, China
008613401186670
guyunchao@buaa.edu.cn

Shi Faqiang
School of Computer Science and Engineer, Beihang University
No.37 Xueyuan Road, Haidian District, Beijing, China
008618293106300
14061115@buaa.edu.cn

ABSTRACT
Breast disease is a common disease in women. The analysis and judgment of B-mode ultrasound images by doctors depend heavily on the operation experience and technical level of doctors. Computer image processing technologies such as natural image classification, target detection and semantic segmentation, represented by deep learning, have been relatively mature, and have been widely used successfully in automatic driving, security, finance and other fields. In this paper, through consultation and cooperation with medical institutions, a large mammary ultrasound image data set is constructed, which basically meets the needs of deep neural network training and validation testing. It is used to develop and validate algorithms for subsequent sub-tasks of ultrasound image analysis.

CCS Concepts
• **Computing methodologies→ Modeling and simulation→ Simulation types and techniques**

Keywords
Breast cancer diagnosis, Data acquisition, Neural network training, Ultrasound image analysis

1. INTRODUCTION
Cancer has always been the enemy of human health. In recent years, the incidence of cancer is increasing gradually around the word. Like most countries, breast cancer is the most frequently diagnosed cancer, and the sixth leading cause of cancer-related deaths among Chinese women[1], which seriously endangers women's lives and health.

The practice shows that the "three early principles" of early detection, early diagnosis and early treatment are the key to improve the survival rate of patients with malignant tumors. Although the diagnostic and therapeutic techniques of breast cancer have made great progress and new technologies and drugs have been developed continuously, there is no effective method to prevent breast cancer so far. The etiology and mechanism of adenocarcinoma are still unclear [2]. In the early diagnosis and treatment of breast cancer, the use of mammography is still controversial among women under 50 years old. However, 57% of Chinese breast cancer patients are within this age range [3]. According to statistics, breast masses in people under 20 years old are benign, breast cancer cases in people under 30 years old are rare, and the incidence of breast cancer in people over 30 years old is rising, especially in people over 40 years old, who are at the highest risk of breast cancer. The incidence of breast cancer in breast masses increases with age. There is no national breast cancer screening program in China Cancer Registry so far. The commonly used mammography and X-ray technology in western countries cannot achieve convincing results in the face of China's dispersed and huge population, and the price is high and the efficiency is low.

Ultrasound has become a routine method for the diagnosis of breast diseases because of its simplicity and practicability. In the general process of breast cancer diagnosis, doctors mainly analyze and judge B-mode ultrasound images through vision. However, the results of B-mode ultrasonography are related to the level of proficiency of doctors, poor image quality, benign manifestations of malignant lesions and visual fatigue or negligence of observers. For a large number of ultrasound mammary images, if observed manually, there will be great defects. Lesions that should be carefully identified but not detected by radiologists result in a high rate of misdiagnosis [4].

In recent years, big data technology and statistical learning technology have promoted the development of artificial intelligence. Especially with the gradual maturity of deep learning technology, it has led to the rapid development of many scientific fields. It has been applied in automatic driving, security, education, medical treatment, finance, e-commerce and retail, in order to improve the automation ability. It provides important tools and methods to reduce human errors, and the cost is low. In the medical field, Google uses machine learning, predictive analysis and pattern recognition to diagnose breast cancer. Its accuracy may reach or even surpass that of human pathologists. This means that the application of machine learning can not only predict results more reliably and improve diagnosis, but also replace most of the work of pathologists [5]. Therefore, in the study of computer-aided diagnosis of breast cancer, advanced convolutional neural network technology and target recognition detection technology in deep learning are used to quickly, efficiently and accurately classify, detect and segment the lesion areas in breast ultrasound images, and to provide real-time diagnostic reference for physicians in the actual medical

ICBDE'19, March 30-April 1, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6186-6/19/03…$15.00
DOI: https://doi.org/10.1145/3322134.3322148

production environment. It has become a research hotspot and the trend of technology development at home and abroad.

By using artificial intelligence technology of computer vision and deep learning, collecting and establishing a unified database of breast ultrasound images, analysis and training, the realization of automated method for computer-aided analysis of breast cancer B-ultrasound images can quickly enhance the correct diagnostic rate of front-line medical staff in our country, reduce the difference of operation level between urban and rural doctors, realize accurate auxiliary diagnosis of breast cancer earlier, faster, simpler and lower cost, and save people's lives and property safety. This will benefit the vast majority of urban and rural residents, with obvious medical needs and great social significance.

2. DATA COLLECTION

Aiming at the problem that the existing publicly accessible data sets in the field of automatic medical image analysis cannot meet the research needs, firstly, through consultation and cooperation with medical institutions, a large data set of mammary ultrasound images is constructed, which basically meets the needs of deep neural network training and validation testing. It is used to develop and validate algorithms for subsequent sub-tasks of ultrasound image analysis.

2.1 Data Sources

Based on the accumulated data of breast ultrasound in Hunan Xiangya Hospital over the years, 2673 mammograms were collected in three batches, of which 1425 were normal mammograms and 1248 were pathological mammograms from 405 patients. These lesions also include patient information and 14 subtypes of breast cancer and bounding-box labeling and mask labeling of lesions, such as specific intraductal cancer, invasive cancer, non-specific invasive cancer, high-grade intraductal cancer, invasive ductal cancer, invasive lobular cancer, invasive breast ductal cancer, right breast high-grade intraductal cancer, acne-type, left chest wall invasive poorly differentiated cancer, intraductal papillary cancer, intraductal papillary papillary cancer, high-grade intraductal carcinoma of left breast, intraductal carcinoma of left breast tending to breast, as shown in Figure 2. Another 118 breast ultrasound images without labeling information were used as independent test sets. The breascancer Breast Ultrasound (BUS) data set proposed by Yap et al. [6] is also used, including 163 breast ultrasound images of lesions, all with mask labeling bounding-box labeling. Based on these data, after desensitization operation, this paper constructs an original database containing nearly 3000 ultrasound images and their labeling information, using these data which no longer contain the patient's personal privacy, as shown in Table 1. This database contains all kinds of mammary ultrasound images from different regions, hospitals, ages and equipments. It has the generality and representativeness in general sense, and meets the basic requirement of independent and identical distribution of original data in machine learning tasks in theory.

Figure 1 Examples of breast ultrasound images

Table 1 Statistics of raw data

Source	Type	Quantity
Xiangya hospital	Normal	1425
	Pathological changes	1248
	Independent Test Set	118
BUS	Pathological changes	163
Total		2954

Figure 2 An example of lesion area mask

2.2 Data Annotation

The part of this data set from BUS [6] has labeled data, while 2673 original images from hospitals need to obtain labeled data first. In order to realize the task of classification, detection and segmentation of breast ultrasound images, a graphical labeling tool was developed for doctors based on LabelMe open source system [7]. Doctors were coordinated to label the specific canceration types and masks of each ultrasound image, and then the bounding-box was extracted for the training and verification of further learning model.

(1) Classification and labeling

Firstly, the type of each image is identified by the file name, which renames each image as a-b-c.png, in which a represents the unique id number of each image, ranging from 1 to 2673; b represents the disease id, of which 0 represents normal and no lesion, while 1 to 14 correspond to the 14 common subtypes of breast cancer, i.e. intraductal cancer, invasive cancer, Non-specific invasive carcinoma, high-grade intraductal carcinoma, invasive ductal carcinoma, invasive lobular carcinoma, invasive ductal carcinoma, high-grade intraductal carcinoma of the right breast, acne-type, low-differentiated invasive carcinoma of the left chest wall, intraductal papillary tumor, intraductal papillary carcinoma, high-grade intraductal carcinoma of left breast and intraductal carcinoma of left breast; c stands for patient id, 0 for missing information, and 1 to n for data from n different patients, which will help to evaluate the similarity of each patient's own data and the difference between different patient's data. In the process of labeling this part of data, this paper uses the pathological gold standard of pathological histological examination results of patients as the accurate basis, which is also the highest standard in the international medical field. That is, doctors will observe the patient's tissue slices under a microscope until the cancer cells are found.

(2) Mask annotation

For the mask part of the lesion area, the polygon is formed by point-by-point links between coordinate points of the lesion edge to confirm whether it is the lesion area pixel by pixel, and the rest of the pixels are all grouped into the background pixel set. At the same time, the method of setting seed points for regional growth is used to help doctors to screen the lesion area roughly, so as to reduce the workload of doctors. In the process of labeling, because of the lack of pathological gold standard data, the method of voting for each pixel with the same part labeled by three doctors is adopted. That is to say, for each pixel, at least two doctors label together before they think that it is really the pathological part, and then return it to the doctor for manual verification, which becomes ground-truth labeling in the database as a real labeling. The annotation process is shown in Figure 3. After labeling, the binary image with mask labeling data is saved in PNG image format. This is because compared with JPEG format, PNG adopts lossless compression algorithm, which can retain more picture information without significantly increasing storage space consumption, and is convenient for the use of the subsequent deep learning model.

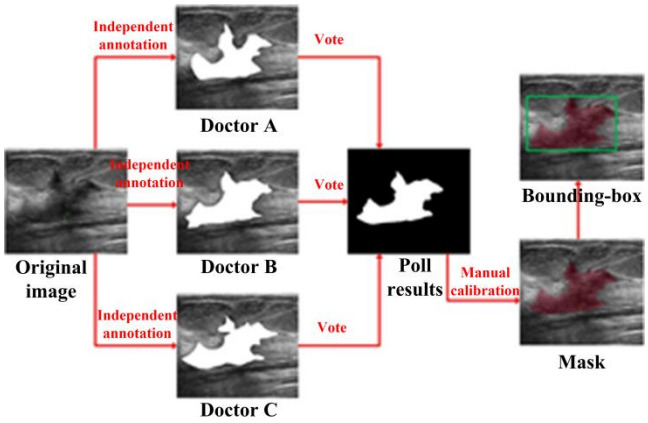

Figure 3 Schematic diagram of lesion area labeling algorithm

(3) Extraction of bounding-box

In order to further reduce the workload of doctors and improve the level of automation, after obtaining the source code labeling of mask on each picture, the bounding-box data is no longer labeled manually, but the bounding-box boundary box can be obtained by simply calculating the minimum positive outer rectangle of the mask area, and then the final proofreading can be carried out by the doctor. The minimum positive outer rectangle of the mask region is used because it avoids the problem that the minimum outer rectangle of the mask region exceeds the image region. At the same time, regular positive rectangle is helpful to reduce the computational cost of predicting bounding-box in target detection.

The difficulty is that when there are multiple lesion targets in each image, the problem of multi-target overlap is easy to occur. At the same time, there are noise interference problems in non-labeled areas, as shown in Figure 4.

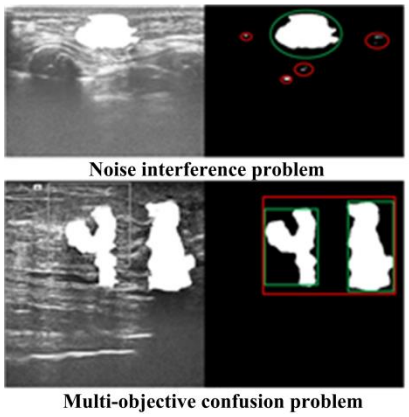

Figure 4 A sketch of the difficulty of bounding-box automatic labeling: where the green area is the correct labeling required, and the red part is the error interference

(4) Multi-objective overlap problem: In this paper, non-maximum suppression (NMS) algorithm is used for filtering and fusion to avoid multi-objective overlap. Non-maximum suppression, as its name implies, is to search for local maxima by suppressing elements that are not maxima. Thus, most of the non-maximum suppression targets are filtered out and different targets are detected in different local domains.

(5) Noise interference problem: This paper uses Gauss filter in OpenCV open source library to solve the problem. This is a linear smoothing filtering algorithm, which is suitable for eliminating Gauss noise and is widely used in image processing. Generally speaking, Gauss filtering is the process of weighted averaging of the whole image. The value of each pixel is obtained by weighted averaging of its own and other pixel values in its neighborhood. Generally, there are two ways to realize it: discrete window sliding convolutional and Fourier transform.

After the boundary box extraction is completed, it is transformed into PASCAL VOC annotation format. That is, a picture corresponding to a standard XML format file is used to describe the lesion area target box, in which each target as an object contains its target type name and bounding-box. The type names here include 14 subtypes of breast cancer and 15 background types, while the bounding-box is determined by the upper left coordinate and the lower right coordinate.

3. DATA PREPROCESSING

3.1 Key Region Extraction

As shown in Figure 1, the original data contains interference information such as hospital identification, ultrasound equipment manufacturer information, and there is a clear boundary segmentation between these information and the main image information needed. Therefore, traditional edge detection algorithms such as Laplace operator extraction can be used to extract the key ROI (Region of Interest) regions, and histogram equalization is used to preprocess the images to normalize the spatial distribution of the pixels, which is more conducive to the training of deep learning model, as shown in Figure 5.

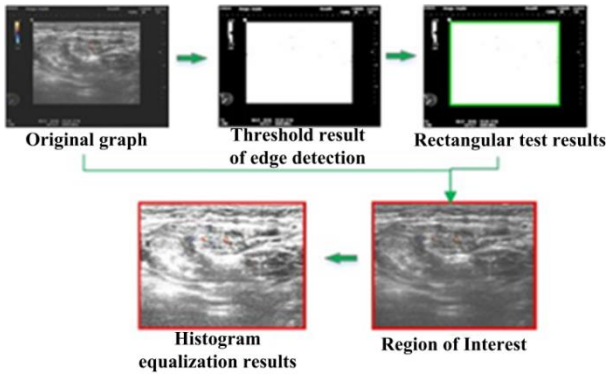

Figure 5 A schematic diagram of ROI extraction process for ultrasound images

3.2 Data Enhancement

For deep convolutional neural networks, in order to give full play to their own strong expression ability, we must use massive data such as ImageNet data set to drive model training and validation, otherwise, it will appear over-fitting phenomenon because of its strong feature expression ability However, the data volume of more than 3000 pictures in this paper is still smaller than that of deep convolutional neural network, so this paper uses a variety of data augmentation methods to expand the existing data sets.

Flipping: that is to say, each picture is randomly flipped horizontally and vertically. Our ultrasound images are detected from the skin surface down by the ultrasound instrument, so they have the characteristics of the upper part of the image is the skin surface, and the lower part is the breast deep layer. Therefore, it is not suitable for vertical flipping, but can only do all horizontal flipping to double the amount of data.

Random clipping: Because the convolutional neural network usually uses the image with the aspect ratio of 1:1, that is, the square as the input image of the network, so it uses larger square, such as 0.8 times to 0.9 times the size of the original image, and picks the image randomly on the original map to form new images directly, so the number of picks on each image determines the multiples of data expansion. In this way, the direct use of squares to extract will avoid image distortion caused by stretching, compression and other transformations in subsequent operations.

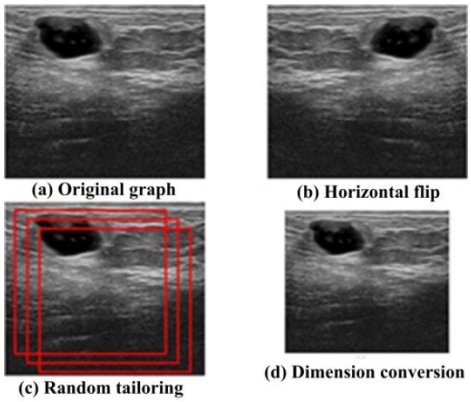

Figure 6 Schematic diagram of data enhancement method

Color jitter: This is a slight disturbance to the RGB color distribution of the original image in RGB color space, or fine-tuning the brightness and saturation of the original image in HSV color space, or even disturbing the hue of the original image in a small range.

Rotating transformation: This method is to rotate the original map to -30 degrees, -15 degrees, 15 degrees, 30 degrees and other certain angles, directly as a new map.

Scaling: This method directly transforms the image resolution to 0.8, 0.9, 1.1, 1.2, 1.3 multiples of the original image to obtain new images. It can be seen that scaling and rotation operations will increase the stability of generalization performance of deep convolutional neural network model in scale and direction.

Noise disturbance: This is a method of randomly adding a few noise elements to the original image to produce a new one. Finally, more than 20,000 data sets containing labeled ultrasound images were obtained for training. Practice shows that as the first step of training deep convolutional neural network, reasonable and effective data enhancement operation can not only increase the number of training samples, but also rapidly improve the diversity of training samples, which will bring double benefits of avoiding over-fitting and improving the performance of the model.

3.3 Category Balance Processing

Among the data we obtained, 1425 normal and 1248 sick photos were taken respectively. There was an imbalance between the two types of data. In the diseased images, the data distribution gap of 14 subtypes is larger (Figure 7), so it is easy to neglect the part of less data in the training process of neural network. In this paper, data balance processing is done, that is, data sampling method is used to make the whole data sample tend to balance between classes.

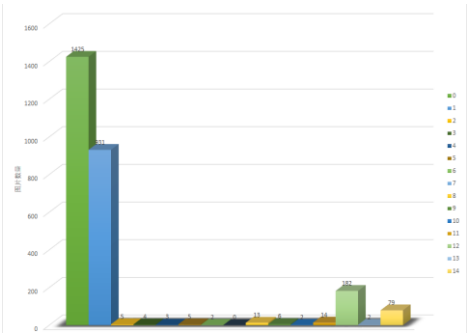

Figure 7 Data distribution diagram

For example, data over-sampling, that is to say, for a class with fewer samples, it is simple to duplicate repeated sampling until it is consistent with the largest sample size. In practice, we use the data enhancement method to replace simple duplication. Data under-sampling is not the direct random discarding of some images in the sample classes with large amount of data, because it will waste less image data due to reducing the amount of training data, further affecting the generalization prediction ability of the model.

In practice, the correct under-sampling method is to use the training method of Mini-batch stochastic gradient descent (MBGD) in the training process of the model. Then, in each training batch (Mini-batch), a small number of class samples are strictly sent into the training, while only a part of the larger number of classes are randomly sampled into the training. The purpose of sample balance is not to waste existing data. Although this data sampling method achieves the goal of making the whole data sample tend to be balanced between classes, even if the number of samples of each type is basically the same, it is easy to lead to the over-fitting of the deep convolutional neural network model due to repeated and continuous use of the same samples.

Even with the combination of data over-sampling and data under-sampling, the risk of over-fitting is still high.

Therefore, on the basis of the combination of data over-sampling and data down-sampling, this paper constructs a method to generate the class-balanced data sample list, which can quickly generate the training sample data list with discontinuous repetition. It improves the problem of model over-fitting caused by continuous repetition of data in the process of data re-sampling, completes the task of data re-sampling with uniform sampling, and greatly reduces the risk of data over-fitting.

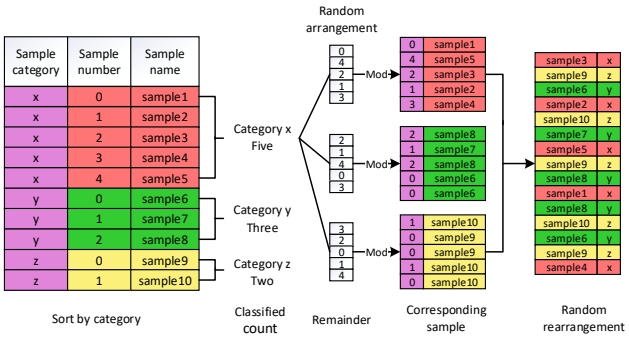

Figure 8 Samples of Class Balanced Data Sample List Generation Algorithms

The steps of this method are as follows (Figure 8): Assuming that there are n categories in the sample (a_1, a_2,..., a_n), we first count the number of samples of each category in the original sample (s_1, s_2,..., s_n), and then find out the number of samples of the type with the largest number of samples, that is, $MaxValue = max$ (s_1, s_2,..., s_n). Next, according to the maximum of *MaxValue* sample number, a list of random numbers is generated for each class of samples, that is, for *n* class samples, *n* random permutations (0, 1, 2,..., *MaxValue*-1) are generated. Next, for each class of samples, the remainder of each class of samples is calculated by using the random number in the random number list of such samples, and the corresponding index values are obtained. For example, for class *i* data, there are s_i samples, for which the list of random numbers (x_1, x_2,..., $x_{MaxValue}$) is generated. Then an index list ($index_1$, $index_2$,..., $index_{MaxValue}$) can be obtained, where $index_j = x_j \% s_i$, $1<j<MaxValue$. Then, according to the index list, the image is extracted from the image of this kind, and the random list of this kind of image is generated. In the last step, the random lists of all categories are stitched together, and the sample list of data sets can be obtained by randomly disrupting the order again. It is not difficult to prove that the number of samples of each category is the same.

In practice, the process of establishing data samples by this method can be embedded in the data loading process of the model. After only providing a list of original data, all operations can be completed in memory without tedious manual operation. For example, for the training set, after completing the training once, it only needs to repeat the above process to continue training, which is very easy to implement, and automatically completes the process of multi-fold cross-validation.

4. DATA SET PARTITIONING

After completing the preparation of data set *D* containing *n* sample data, it needs to be divided into training data and testing data. Because the data set is relatively small, after dividing the data set into two parts by using common leave-out method, a part of the sample data will be retained because of the need of model testing, which will reduce the sample size of training data, and then lead to estimation bias because of the different size of training samples, affecting the experimental results. Specializing the leave-out method to only one sample for testing leave-one method will incur huge computational overhead due to the need to train *m* models and integrate them. In fact, it cannot be realized. So bootstrapping is used to partition data sets.

The basic principle of self-help method is to put back and resample. Firstly, an empty set D_1 is established. Then, on our existing data set *D* containing *n* samples, one sample is randomly sampled at a time and put into D_1. This operation is repeated *n* times, so that D_1 contains *n* data samples and can be directly used as a training set. Because this is a replayed repeated sampling, some samples in *D* appear repeatedly in D_1 and some samples do not appear, so this part of the samples that do not appear will be directly used as tester D_2. According to the knowledge of probability theory, it can be simply estimated that the probability that a sample will never be sampled is $(1-\frac{1}{n})^n$ when repeated sampling is performed in data set *D* with *n* samples. It is assumed that the sampling can be repeated indefinitely, and the limit is as follows:

$$\lim(1-\frac{1}{n})^n = \frac{1}{n} \approx 0.368 \tag{1}$$

Therefore, the test set D2 contains about 36.8% of the total data, which also conforms to the general rule that the ratio of training set to test set is 7:3.

In some machine learning tasks, test sets need to be further divided into validation sets and test sets. In this paper, we use the same method to continue the self-help operation with replay and resampling on the basis of the existing test set D_2, and divide it into test set D_3 and verification set D_4. In this way, we divide the data set into training set D_1, verification set D_4 and test set D_3, and meet the requirements of machine learning algorithm.

$$D_1 \cap D_3 \cap D_4 = \Phi \tag{2}$$

$$D_1 \cup D_3 \cup D_4 = D \tag{3}$$

After partitioning the data set, the mean reduction operation is performed on the training set, the verification set and the test set respectively. That is to say, for each set, the pixel mean of all images is calculated, and then the new image pixel value is obtained by subtracting this mean from each image. The principle of mean subtraction operation is that we default that natural images are a kind of stationary data distribution, that is to say, these data obey the same statistical distribution law in each dimension, so that after subtracting the statistical average of data from each sample, the same part of the sample can be removed from other samples. Furthermore, the individual differences of each sample are highlighted, which is beneficial to the feature learning of deep convolutional neural network.

Here we first divide the training set, verification set and test set, and then subtract the mean on each set separately, instead of calculating the whole mean directly on the undivided whole data set. This is because there is a basic logic in machine

learning that can and can only get information from training data but not from test data in the process of model training.

5. CONCLUSION

This paper briefly summarizes the process of collecting mammary ultrasound image data in cooperation with medical institutions. Under the golden standard of pathological examination, a data labeling tool is developed to meet the data characteristics and labeling requirements, which helps doctors complete the data labeling work. Then data cleaning and preprocessing such as Key Region Extraction, Data Enhancement and Class Balance Processing Algorithms in image were conducted, and the partition of data sets is completed.

This paper is still in further consultation with medical institutions to make this data set public, to help more people train their own relevant algorithm models, and to promote the development and progress of the global medical image assisted analysis industry.

6. REFERENCES

[1] Ferlay J. GLOBOCAN 2008, cancer incidence and mortality worldwide: IARC Cancer-Base No. 10[J]. http://globocan. iarc. fr, 2010.

[2] Zhang P., Verma B., Kumar K. Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection[J]. Pattern Recognition Letters, 2005, 26(7):909-919.

[3] Li J., Zhang B. N., Fan J. H., et al. A Nation-Wide multicenter 10-year (1999-2008) retrospective clinical epidemiological study of female breast cancer in china[J]. BMC Cancer, 2011, 11(1):364.

[4] Jackson V. P., Hendrick R. E., Feig S. A., et al. Imaging of the radio graphically dense breast.[J]. Radiology, 1993, 188(2):297-301.

[5] Z O., EJ E. Predicting the Future -Big Data, Machine Learning, and Clinical Medicine[J]. N Engl J Med, 2016, 375(13):1216-1219.

[6] Yap M. H., Pons G., Mart íJ., et al. Automated Breast Ultrasound Lesions Detection using Convolutional Neural Networks[J]. IEEE Journal of Biomedical & Health Informatics, 2017, PP(99):1-1.

[7] Russell B. C., Torralba A., Murphy K. P., et al. LabelMe: A Database and Web-Based Tool for Image Annotation[J]. International Journal of Computer Vision, 2008, 77(1-3): 157-173.