Computer Methods and Programs in Biomedicine 190 (2020) 105361

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks

Woo Kyung Moon[a], Yan-Wei Lee[b], Hao-Hsiang Ke[b], Su Hyun Lee[a], Chiun-Sheng Huang[c], Ruey-Feng Chang[b,d,e,f,*]

[a] Department of Radiology, Seoul National University Hospital and Seoul National University College of Medicine, Seoul 110-744, South Korea
[b] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, ROC
[c] Department of Surgery, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan, ROC
[d] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan, ROC
[e] Graduate Institute of Network and Multimedia, National Taiwan University, Taipei, Taiwan, ROC
[f] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan, ROC

ARTICLE INFO

Article history:
Received 12 May 2019
Revised 14 January 2020
Accepted 24 January 2020

Keywords:
Breast cancer
Breast ultrasound
Computer-aided diagnosis
Deep learning
Convolutional neural network
Ensemble learning

ABSTRACT

Breast ultrasound and computer aided diagnosis (CAD) has been used to classify tumors into benignancy or malignancy. However, conventional CAD software has some problems (such as handcrafted features are hard to design; conventional CAD systems are difficult to confirm overfitting problems, etc.). In our study, we propose a CAD system for tumor diagnosis using an image fusion method combined with different image content representations and ensemble different CNN architectures on US images. The CNN-based method proposed in this study includes VGGNet, ResNet, and DenseNet. In our private dataset, there was a total of 1687 tumors that including 953 benign and 734 malignant tumors. The accuracy, sensitivity, specificity, precision, F1 score and the AUC of the proposed method were 91.10%, 85.14%, 95.77%, 94.03%, 89.36%, and 0.9697 respectively. In the open dataset (BUSI), there was a total of 697 tumors that including 437 benign lesions, 210 malignant tumors, and 133 normal images. The accuracy, sensitivity, specificity, precision, F1 score, and the AUC of the proposed method were 94.62%, 92.31%, 95.60%, 90%, 91.14%, and 0.9711. In conclusion, the results indicated different image content representations that affect the prediction performance of the CAD system, more image information improves the prediction performance, and the tumor shape feature can improve the diagnostic effect.

1. Introduction

Ultrasound (US) is a useful way for the detection and diagnosis of breast cancer [1] because they are non-invasive, non-radioactive, real-time imaging, and high image resolution. However, to read US images requires well-trained and experienced radiologists. Even a well-trained expert might have a high inter-observer variation rate on tumor diagnosis [2]. Hence, computer-aided diagnosis (CAD) could be used to assist radiologists in breast cancer classification and detection [3–6]. Recently, several studies [7–10] have discussed the automatic breast cancer diagnosis method to classify benign and malignant tumor in US images.

Convolutional neural network (CNN) approaches have been proven to be very effective in a wide range of computer vision applications [11–14]. In addition, CNN can recognize visual patterns directly from pixel images with minimal preprocessing and automate the whole feature extraction process. Furthermore, CNN has been employed broadly in medical image analysis, such as segmentation [15], classification [16], and detection [17]. In recent years, the usages of CNN models in ultrasound of breast cancers are shown significant development. Byra et al. [18] proposed a color conversion method that transfers the grayscale ultrasound images to 3-channel (RGB) images, which enhanced the classification performance. Yap et al. [19] proposed an end-to-end deep learning model in automated breast ultrasound lesions recognition; they are the first to implement semantic segmentation on BUS images and compared the performance between different CNN models. Yap et al. [20] proposed an automatic detection system of breast ultrasound lesions using CNN models, which compared three different CNN models of CAD systems that reduced the operator-dependent problem. Even if the CNN method was widely used in medical image fields for segmentation and diagnosis, but we want to under-

* Corresponding author at: Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan, ROC.
E-mail address: rfchang@csie.ntu.edu.tw (R.-F. Chang).

https://doi.org/10.1016/j.cmpb.2020.105361
0169-2607/

Table 1
Detailed transducers information of different US machines used in this study.

Machine Model	Description
Siemens	8–15 MHz linear-array
Acu-	52-mm ultrasound probe
Philips	6–12 MHz linear-array
Sequoia	50-mm ultrasound probe
HDI	3–11 MHz linear-array
5000	37.4-mm ultrasound probe
GE	4.5–14 MHz linear-array
LOGIQ	40-mm ultrasound probe
GE	4.5–14 MHz linear-array
LOGIQ	40-mm ultrasound probe
Kretz	24–31 MHz linear-array
Vo-	38-mm ultrasound probe
Medison	27–32 MHz linear-array
Voluson	38-mm ultrasound probe
Philips	27–42 MHz linear-array
iU22	38-mm ultrasound probe

stand the impact of different image content descriptions and CNN architectures on the various US diagnostic system.

In our study, we propose a CAD system for tumor diagnosis using an image fusion method combined with different image content representations and ensemble different CNN architectures on US images. First, we manually extract the region of interest (ROI), which covers the whole tumor and the ROI boundary close to the tumor margin. Then, the expert manually extracts the tumor region and the tumor shape image (TSI). In addition, we employed an image fusion method to enhance the diagnostic performance of our CAD system. Finally, we employed the ensemble method to combine multiple CNN results.

2. Material

In this study, we used two datasets of breast ultrasound images: Seoul National University Hospital (SNUH, Korean) collected the private dataset (dataset SNUH); the public dataset (dataset BUSI [21]) was collected by Baheya Hospital for Early Detection & Treatment of Women's Cancer (Cairo, Egypt) (https://doi.org/10.1016/j.dib.2019.104863).

2.1. SNUH dataset

In our private dataset, the breast US images were obtained from 8 different US machines. Table 1 provides detailed descriptions of the transducer information of each device. Due to different parameter settings and models of the US machines, the acquired images are shown in Fig. 1(a)–(h).

The study was permitted by the local ethics committee and informed consent was obtained from all of the included patients. The US images used in this research were acquired from Seoul National University Hospital (SNUH) from June 2000 to January 2018. A total of 1225 patients (mean age: 45.58 ± 9.62 years; range: 17–85 years) with 1687 tumors (mean diameter: 14.51 ± 8.03 mm; range: (2.3–45 mm)) with biopsy-proven diagnosis were included in this study. Malignant tumors are slightly larger than benign lesions, the mean size of benign lesions was 11.58 ± 5.96 mm(range, 2.3–35.5 mm), the mean size of malignant tumors was 20.86 ± 8.27 mm(range, 4.4–45.0 mm). Moreover, the number of benign lesions was 503 in fibroadenoma (FA), 404 in fibrocystic change (FC), 17 in papilloma, and 29 in other lesions. The number of malignant tumors was 663 in infiltrating ductal carcinoma (IDC), 40 in ductal carcinoma in situ (DCIS), 12 in invasive tubular carcinoma (ITC), and 19 in other lesions.

2.2. BUSI dataset

Dataset BUSI is an open dataset, as shown in Fig. 1(i)–(l), collected in 2018 from Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo, Egypt. A total of 780 tumor images from 600 female patients (25–75 years old), including 133 normal images without masses, 437 images with cancer masses, and 210 images with benign masses were included in the dataset. All images were collected by the LOGIQ E9 ultrasound system and LOGIQ E9 Agile ultrasound system; the transducers are 1–5 MHz on ML6-15-D Matrix linear probe. Dataset BUSI also provides the ground truth image for each corresponding tumor ultrasound image, which freehand segmentation used the Matlab software. The BUSI dataset is suitable for examining the robustness in our system; hence, for reproducibility, we test this open dataset, including the 437 cancer masses, 210 benign masses, and 133 normal images using the proposed method.

3. Method

In our study, we proposed a CAD system for tumor diagnosis by using different CNN architectures with the ensemble method on US images. Fig. 2 shows the flow chart of our CAD system.

3.1. Images

The CNN architecture is commonly used to classify images in CAD systems, but we are interested in the information of the image affecting classification results. Therefore, we used the different image information of the same tumor image to compare in our study, which includes the original ROI image, tumor image, extracted tumor shape image (TSI), and fused image.

ROIimage. The ROI of B-mode US image is manually cropped by an expert-defined, the ROI region covers the whole tumor area and the ROI boundary close to the tumor margin. In our study, the ground truth image of tumor contour is handmade by an experienced radiologist. The sketch map of ROI extraction is shown in Fig. 3.

Tumor image. For correct separation between the tumor region and normal tissues, a precise segmentation of a suspected tumor region from the B-mode US image is necessary. For each case, the boundary of the tumor was manually delineated by an expert as our segmentation results. Fig. 4 shows the result images after the expert delineated.

Tumor shape image (TSI). In some previous studies [22–24], that indicates the shape of the breast tumor has a high correlation with the tumor diagnostic result. Rangayyan et al. [24] have demonstrated the importance of tumor shape for tumor classification. In most of the breast US image cases, benign lesions usually have smooth edges, rounded or elliptical shapes with an obvious border; however, malignant tumors often have spiked edges or branches with blurred borders [25,26]. Hence, the segmented map has sufficiently shape information to distinguishing between benign lesions and malignant tumors. After tumor segmentation, we obtained a segmented map image that contains tumor contour information and we called the TSI. In this study, we added a TSI to assist the diagnosis system to classify tumors into benign or malignant. The examples of B-mode tumor images and the corresponding TSI images are shown in Fig. 5.

Fused image. In some previous studies [27–31], several image fusion methods were used to enhance the content representation, which improves the performance of target segmentation, object detection, target recognition, etc. In our study, we both used the original tumor image and TSI to provide complementary information which includes the details of the texture, the intensity of the pixel, the degree of tissue variation within the tumor, and the

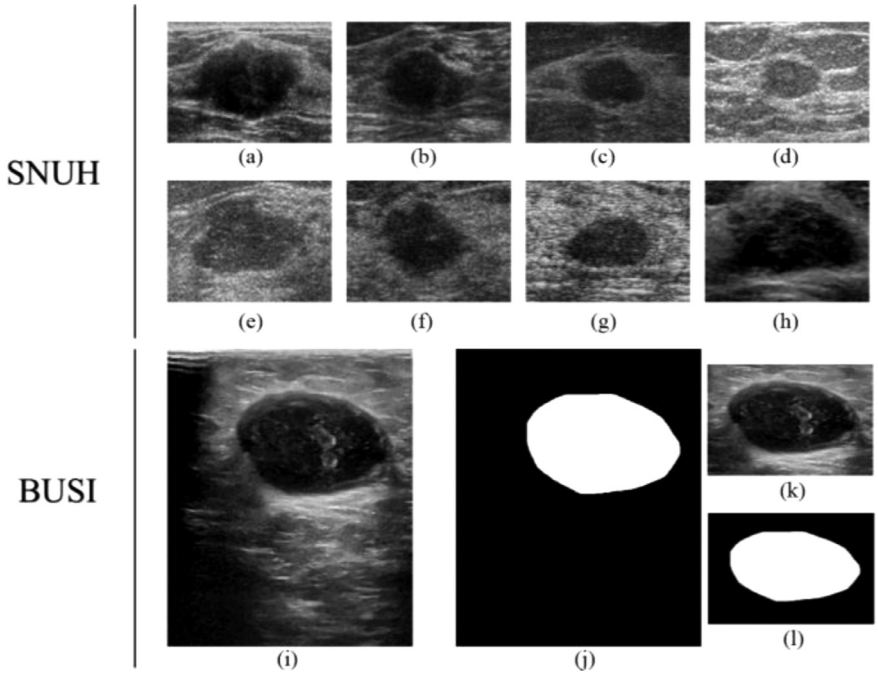

Fig. 1. Two datasets are used in this study. In the private dataset (SNUH), the image data acquired from 8 different US machines. Each machine has distinctive characteristics, which challenging the robustness of the proposed systems. (a) Siemens Acuson Sequoia, (b) Philips ATL HDI 5000, (c) GE Expert, (d) GE LOGIQ V7, (e) GE LOGIQ, (f) Kretz Voluson 730, (g) Medison Voluson, (h) Philips iU22. The public dataset (BUSI) collected by the LOGIQ E9 ultrasound system includes the breast ultrasound images and the related tumor boundary ground truth image. (i) a benign case image, (j) the lesion boundary ground truth image of (i), (k) the cropped tumor image, (l) the cropped tumor boundary ground truth image.

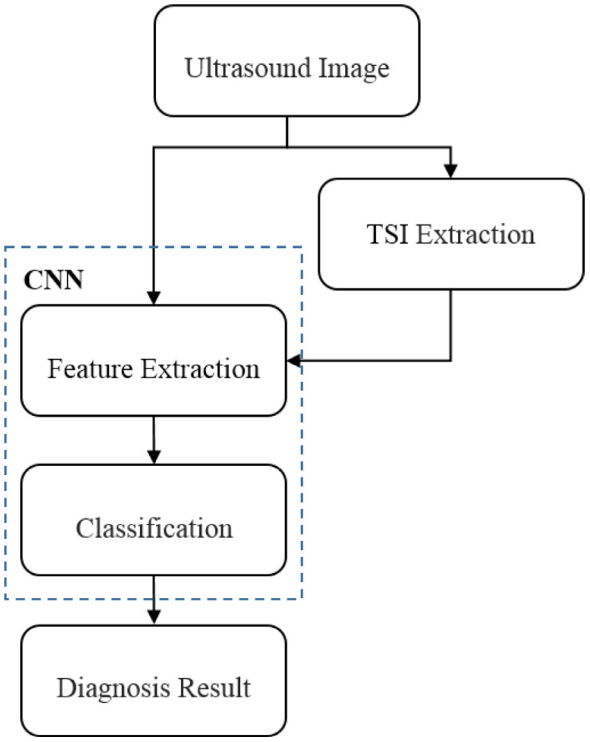

Fig. 2. The flow chart of the tumor diagnosis system.

tumor shape information. Furthermore, we used an image fusion method to combine multiple images into one single image, the multi-channel image could provide different descriptions of a tumor, it might further enhance the performance of tumor diagnosis, and the image fusion process is illustrated in Fig. 6. In the image

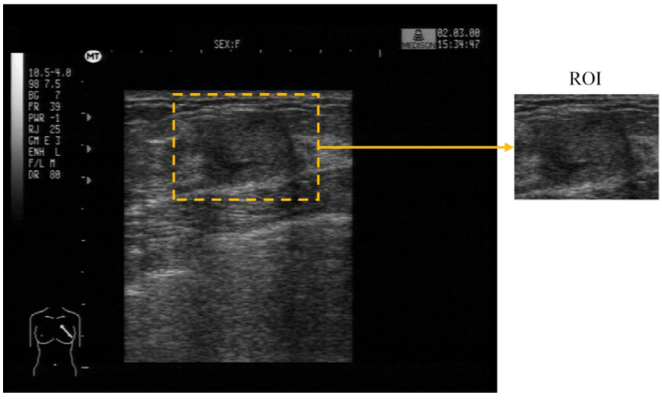

Fig. 3. The region of interest (ROI) on B-mode US image.

fusion procedure, three types of images (original tumor, segmented tumor, and TSI) were concatenated into a 3-channel image (RGB-style), where the original tumor image placed in the red channel, the segmented tumor image placed in the green channel, and the TSI placed in the blue channel, and the schematic diagram is illustrated in Fig. 6(b). Then, we obtained a 3-channel image that contains different information extracted from the tumor.

3.2. CNN architectures

Different CNN architectures affect the performance of classification under a deep learning framework. In our study, we want to understand the impact of different CNN architectures on the diagnostic system. In this section, we provided a description of three different CNN models which include VGG, ResNet, and DenseNet. In addition, all CNN models were trained from scratch on each image type and used 80% and 20% of auto-shuffled data to the train-

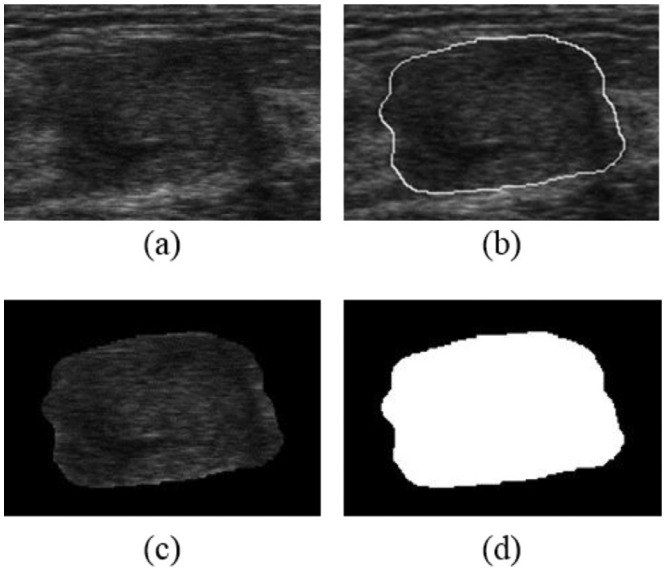

Fig. 4. The tumor region extraction process. (a) The original image of a breast tumor; (b) the tumor contour overlapped on the original image (after manual segmentation); (c) the tumor region image; (d) the segmented map (also called the tumor shape image) of the original image.

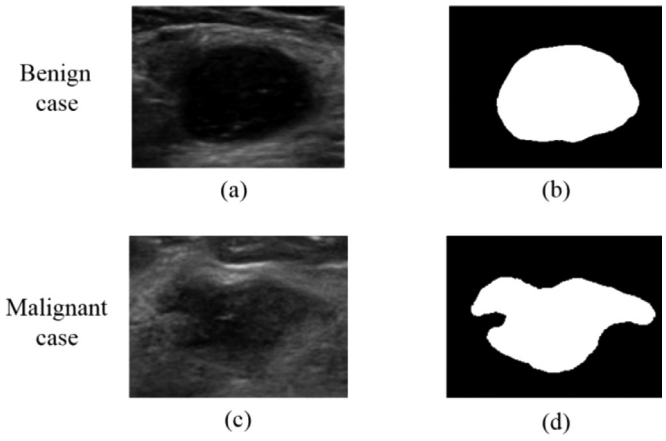

Fig. 5. The examples of B-mode tumor images and the corresponding TSI images. (a) A benign lesion case. (b) The corresponding TSI image from (a). (c) A malignant tumor case. (d) The corresponding TSI image from (c).

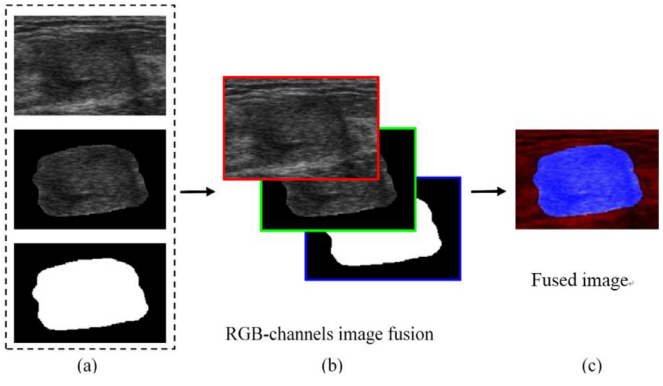

Fig. 6. Image fusion process. (a) The original tumor (upper), segmented tumor (middle), and TSI (down). (b) The concatenated 3-channel (RGB-style) image, the original tumor image placed in the red channel, the segmented tumor image placed in the green channel, and the TSI placed in the blue channel (c) The fusion image that produces from (b).

ing process and testing process respectively (20% as the validation set that extracted from the training set).

VGG-16 and VGG-like. The VGGNet model was proposed by Simonyan [32], which regarded as an extended CNN architecture of AlexNet. VGGNet model extracts more information by using more hidden layers (16 or 19 layers in general) because the smallest filter size (3×3) has a smaller receptive field that helps to acquire more detailed information from an image. However, since VGG-16 is deeper than the other typical CNN architecture, it has more hyper-parameters and more susceptible to the vanishing gradient problem [33–36]. Hence, in order to improve time consuming and memory required during the training process, we proposed a modified VGG-16 architecture, called VGG-Like. The architectures of VGG-16 and VGG-Like are illustrated in Table 2.

ResNet. The Residual Neural Network (ResNet) proposed by Kaiming He et al. [37], which imported a novel architecture with a shortcut connection path that skips one or more layers in the network. ResNet used the residual structure by adding an identity mapping to convert the original function to $F(x) + x$. The input and output of a residual block were linked by a shortcut connection. Based on the benefits of the residual block method, ResNet with 152 layers still has a lower computation complexity than VGGNet. Therefore, ResNet could show a good performance for image classification [38], object detection [39], semantic segmentation [40], etc. We used three ResNet models with different layers 18, 50, and 101 layers in our study.

DenseNet. DenseNet is a densely connected convolutional neural network proposed by Huang [41], and the main idea was there exists a direct connection between any two layers. The input of each layer obtains additional input from all previous layers and the feature maps learned by the layer are also passed directly to all afterward layers as input. The advantages of the DenseNet architecture include reducing the vanishing-gradient problem, enhancing the transmission of the feature, using features more effectively and reducing the number of parameters to make the network faster to be trained [41]. We used the DenseNet model with different layers 40, 121, and 161 in our study.

3.3. Ensemble method

In supervised learning of machine learning, our goal is to learn a stable model that performs well in all aspects of our defined problems. The ensemble method combines multiple models in order to get a better and more comprehensive generalized model [42], and the main idea was even if a weak classifier got a wrong prediction, the whole ensemble classifiers (strong classifier) could correct the error back yet. In addition, the ensemble method could reduce the variance [43] and bias [44] of all model predictions to achieve more accurate predictions. Therefore, we used ensemble methods to enhance the performance and generalization of our CAD.

Base machine selection. In our study, we want to know the effect of different image types that training on different base machines (CNN architecture). Hence, we trained four base machines in each image type (16 in total). We pick the one with the highest diagnostic accuracy from the base machine trained in each image type as our base machine chosen strategy. Then, we could get four base machines that performed best on each image type. Finally, we could get different results through the combination of the four base machines with different combining strategies.

Combining strategy. The ensemble method combined multiple models with different combining strategies to reduce the variance of prediction by different base machine, then enhances the final classification result. The conceptual diagram of the ensemble was illustrated in Fig. 7. In our experiments, we used the unweighted average, weighted average, weighted voting, and stacking as our

Table 2
The VGG-16 and VGG-Like architecture proposed in our paper.

(a) VGG-16			(b) VGG-Like		
Filter size	Layer type	No. of filters	Filter size	Layer type	No. of filters
Input			Input		
3 × 3	Conv	64	3 × 3	Conv	32
3 × 3	Conv	64	3 × 3	Conv	64
2 × 2 Max pooling			3 × 3 Max pooling		
3 × 3	Conv	128	3 × 3	Conv	128
3 × 3	Conv	128	3 × 3	Conv	256
2 × 2 Max pooling			3 × 3 Max pooling		
3 × 3	Conv	256	FC-256		
3 × 3	Conv	256	FC-256		
3 × 3	Conv	256	FC-1		
2 × 2 Max pooling			Sigmoid		
3 × 3	Conv	512			
3 × 3	Conv	512			
3 × 3	Conv	512			
2 × 2 Max pooling					
3 × 3	Conv	512			
3 × 3	Conv	512			
3 × 3	Conv	512			
2 × 2 Max pooling					
FC-4096					
FC-4096					
FC-1					
Sigmoid					

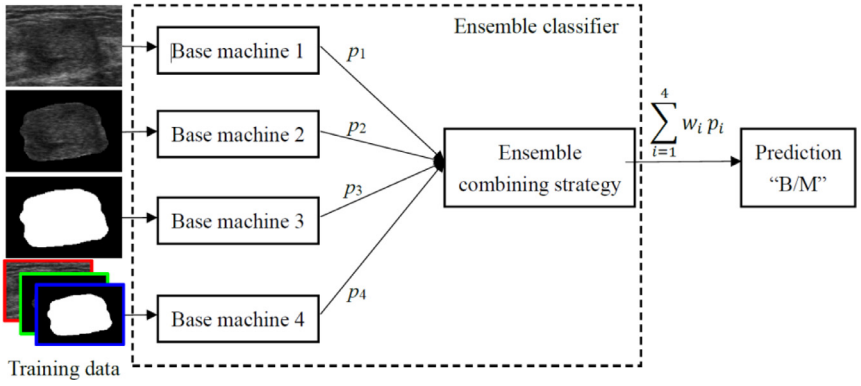

Fig. 7. The concept diagram of the ensemble method proposed in our system.

ensemble method. In addition, we set double weights on the base machine which has the highest accuracy in the single classifier experiment. Finally, the comparison of results on the effects of different combine strategies.

4. Experiment results

In this section, we compared the diagnostic performance of all CNN architectures, including: VGG-Like, VGG-16, ResNet-18, ResNet-50, ResNet-101, DenseNet-40, DenseNet-121, and DenseNet-161. Furthermore, we list the base machine which achieved the best performance on the test set and compares performance between using different ensemble methods.

In statistical analysis, six quantitative indicators were used to evaluate the diagnostic performance: accuracy (ACC), sensitivity (SEN), specificity (SPEC), precision, recall, and F1 score. Also, we used the receiver operating characteristic (ROC) curve and the area under the receiver ROC curve (AUC) to assess the diagnostic performance between different CNN architectures. The ROC curve was plotted by using the ROCKIT software (University of Chicago, Chicago, IL, USA). Also, *Precision* and *Recall* are common indicators for evaluating classification performance [45]. However, sometimes *Precision* and *Recall* are contradictory, so we employed the *F1 Score* [46] for comprehensive consideration.

The equations about the ACC, SEN, SPEC, Precision, Recall and F1 Score are defined as below:

$$\text{Accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Sensitivity (SEN)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity (SPEC)} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1 Score} = \frac{2{*}\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where truth positive is the TP, false positive is FP, true negative is TN, and false negative is FN.

And, the detailed implementation hyper-parameters are depicted in Table 3, the recorded hyper-parameters generated the best performance in our experiments.

Table 3
The hyper-parameters of all architectures.

	VGG-Like	VGG-16	Res-18	Res-50	Res-101	Dense-40	Dense-121	Dense-161
Optimizer	Adam	Adam	SGD	Adam	SGD	SGD	SGD	SGD
Loss function	Binary crossentropy	Binary crossentropy	Binary crossentropy	Binary crossentropy	Binary crossentropy	Binary crossentropy	Binary crossentropy	Binary crossentropy
Batch size	256	64	256	256	256	16	36	32
Epoch	200	200	200	200	200	200	200	200
Learning rate	4e-06	3e-06	1e-02	3e-02	4e-02	3e-02	8e-03	8e-02

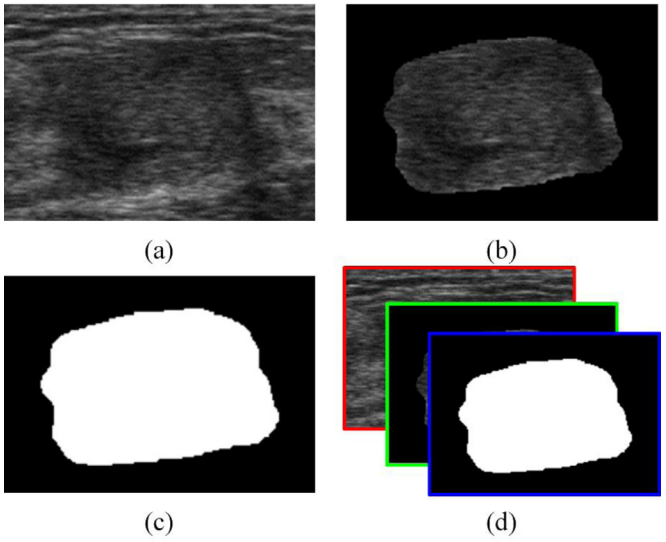

Fig. 8. Four kinds of image types used for our diagnosis system. (a) The original image of a breast tumor; (b) the segmented tumor image, which we use the segmented map of original image to capture tumor image part; (c) the TSI, which contains the tumor shape information; (d) the fused image, that we directly arranged (a), (b) and (c) in three channels.

4.1. Comparison of different image type (SNUH dataset)

In our study, we used four different images extracting from the US image, which includes original tumor, segmented tumor, tumor mask, and 3-channel image (illustrated in Fig. 8), and then we employed different CNN architectures to compare the performance. After the above processing, we obtained four types of images from B-mode images that were used in our proposed CNN-based diagnosis system and compared the performance between different image types.

Image type 1 – The original tumor image (ROI). First, the ACC, SEN, SPEC, Precision, F1 score and AUC of VGG-Like based on Image type 1 (the original tumor image) were 84.57%, 73.65%, 93.12%, 89.34%, 80.74% and 0.9198, respectively, when using VGG-16 were 84.57%, 73.64%, 93.12%, 89.34%, 80.74% and 0.9322, respectively. The performance of VGG-Like is similar to VGG-16.

Second, the ACC, SEN, SPEC, Precision, F1 score and AUC of ResNet-18 based on Image type 1 were 81.60%, 86.49%, 77.77%, 75.29%, 80.50% and 0.9185, respectively, when using ResNet-50 were 81.60%, 75.68%, 86.24%, 81.16%, 78.32% and 0.8883, respectively, when using ResNet-101 were 84.57%, 75.00%, 92.06%, 88.10%, 81.02% and 0.9104, respectively. We could find that the ResNet models have a higher sensitivity than models in Image type 1.

Third, the ACC, SEN, SPEC, Precision, F1 score and AUC of DenseNet-40 based on Image type 1 were 85.46%, 79.05%, 90.48%, 86.67%, 82.69% and 0.9352, respectively, when using DenseNet-121 were 86.35%, 77.70%, 93.12%, 89.84%, 83.33% and 0.9248, respectively, when using DenseNet-161 were 83.09%, 69.59%, 93.65%, 89.57%, 78.33% and 0.8918, respectively. We could find that the deeper DenseNet models have a higher specificity in Image type 1.

Moreover, the comparison of all CNN architectures using Image type 1, are listed in Table 4. Fig. 9 illustrated the ROC curves and AUC value for all CNN architectures using Image type 1. According to Table 4, DenseNet-40 has the highest AUC value.

Image type 2 – The segmented tumor image (Tumor image). First, the ACC, SEN, SPEC, Precision, F1 score and AUC of VGG-Like based on Image type 2 (the segmented tumor image) were 87.24%, 85.14%, 88.89%, 85.71%, 85.42% and 0.9423, respectively, when using VGG-16 were 85.16%, 81.08%, 88.36%, 84.51%, 82.76%

Table 4
The results of all CNN architectures using Image type 1. Bold indicates the best results training and testing on this image type.

Method	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
VGG-Like	84.57	73.65	93.12	89.34	80.74	0.9198
VGG-16	84.57	73.64	93.12	89.34	80.74	0.9322
ResNet-18	81.60	**86.49**	77.77	75.29	80.50	0.9185
ResNet-50	81.60	75.68	86.24	81.16	78.32	0.8883
ResNet-101	84.57	75.00	92.06	88.10	81.02	0.9104
DenseNet-40	85.46	79.05	90.48	86.67	82.69	**0.9352**
DenseNet-121	**86.35**	77.70	93.12	**89.84**	**83.33**	0.9248
DenseNet-161	83.09	69.59	**93.65**	89.57	78.33	0.8918

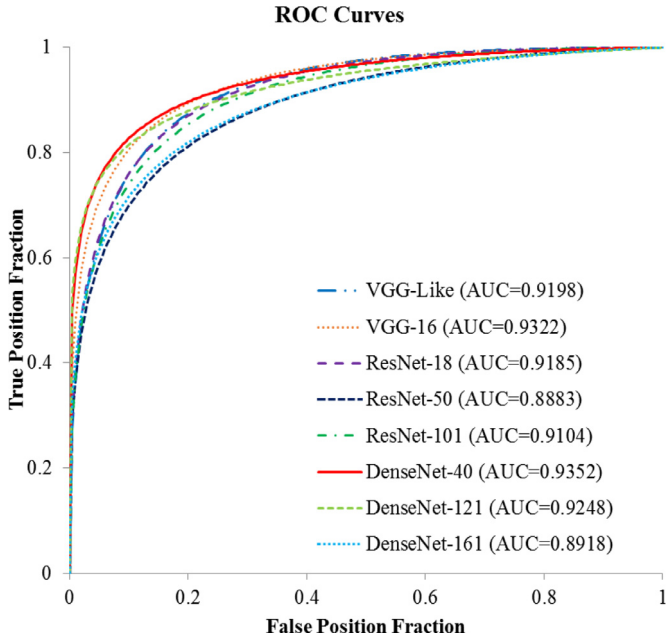

Fig. 9. The ROC Curves of all CNN architectures using Image type 1. The DenseNet-40 has the highest AUC value (0.9352) than other CNN structures.

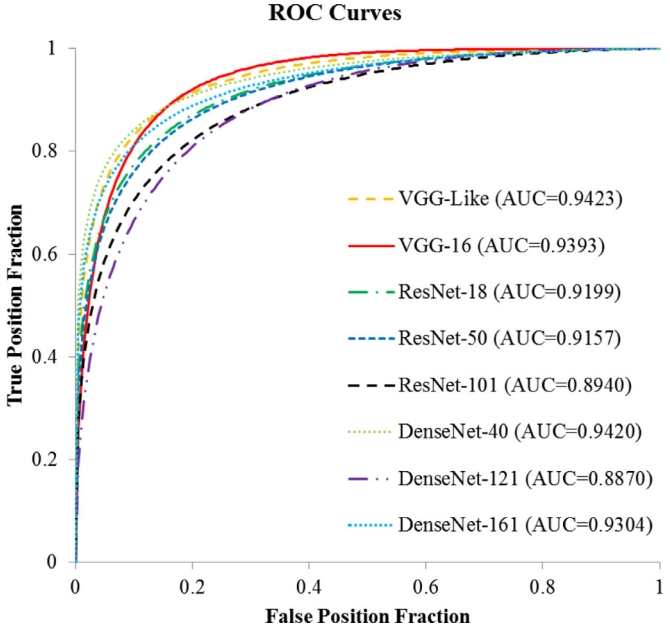

Fig. 10. The ROC Curves of all CNN architectures testing on Image type 2. The VGG-Like has the highest AUC value (0.9452) than other CNN structures.

and 0.9393, respectively. We could find that the VGGNet models have a higher F1 score than others in Image type 2.

Second, the ACC, SEN, SPEC, Precision, F1 score and AUC of ResNet-18 based on Image type 2 were 83.68%, 74.32%, 91.01%, 86.61%, 80.00% and 0.9199, respectively, when using ResNet-50 were 84.27%, 81.76%, 86.24%, 82.31%, 82.03% and 0.9157, respectively, when using ResNet-101 were 85.76%, 76.35%, 93.12%, 77.50%, 80.52% and 0.8940, respectively. We could find that the ResNet models have a higher specificity than others in Image type 2.

Third, the ACC, SEN, SPEC, Precision, F1 score and AUC of DenseNet-40 based on Image type 2 were 87.24%, 83.11%, 90.48%, 87.23%, 85.12% and 0.9420, respectively, when using DenseNet-121 were 86.35%, 84.46%, 87.83%, 84.46%, 84.46% and 0.8870, respectively, when using DenseNet-161 were 86.05%, 78.38%, 92.06%, 88.55%, 83.15% and 0.9304, respectively. We could find that the DenseNet models in Image type 2 have a higher precision.

Table 5 shows that VGG-Like gets the highest AUC value and DenseNet-121 gets the lowest one. Fig. 10 illustrated the ROC curves and AUC value for all CNN architectures testing on Image type 2.

Image type 3 – The tumor shape image (TSI). First, the ACC, SEN, SPEC, Precision, F1 score and AUC of VGG-Like based on Image type 3 (the TSI) were 84.27%, 81.08%, 86.77%, 82.76%, 81.91% and 0.8989, respectively, when using VGG-16 were 83.97%, 85.14%, 83.07%, 79.75%, 82.35% and 0.9076, respectively. We could find that the VGGNet models have a higher F1 score and AUC than others in Image type 3.

Second, the ACC, SEN, SPEC, Precision, F1 score and AUC of ResNet-18 based on Image type 3 were 81.60%, 86.49%, 77.77%, 76.58%, 79.08% and 0.8709, respectively, when using ResNet-50 were 81.01%, 81.76%, 80.42%, 77.03%, 77.03% and 0.8680, respectively, when using ResNet-101 were 79.82%, 77.03%, 82.01%, 84.16%, 68.27% and 0.8549, respectively.

Third, the ACC, SEN, SPEC, Precision, F1 score and AUC of DenseNet-40 based on Image type 3 were 76.56%, 57.43%, 91.53%, 78.57%, 80.13% and 0.8932, respectively, when using DenseNet-121 were 82.19%, 81.76%, 82.54%, 77.63%, 78.67% and 0.8872, respectively, when using DenseNet-161 were 81.01%, 79.73%, 82.01%, 75.54%, 73.17% and 0.8533, respectively. From the results of the ResNet and DenseNet models could indicate the deeper networks have inversely proportional to performance, in Image type 3.

Table 6 shows that VGG-16 has the highest AUC value and DenseNet-161 gets the lowest one. Fig. 11 illustrated the ROC curves and AUC value for all CNN architectures testing on Image type 3.

Image type 4 – The fused image. First, the ACC, SEN, SPEC, Precision, F1 score and AUC of VGG-Like based on Image type 4 (the 3-channel image) were 86.05%, 80.41%, 90.48%, 86.86%, 83.51% and 0.9421, respectively, when using VGG-16 were 88.72%, 83.78%, 92.59%, 89.86%, 86.71% and 0.9556, respectively. We could find that

Table 5
The results of all CNN architectures using Image type 2. Bold indicates the best results training and testing on this image type.

Method	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
VGG-Like	87.24	**85.14**	88.89	85.71	**85.42**	**0.9423**
VGG-16	85.16	81.08	88.36	84.51	82.76	0.9393
ResNet-18	83.68	74.32	91.01	86.61	80.00	0.9199
ResNet-50	84.27	81.76	86.24	82.31	82.03	0.9157
ResNet-101	85.76	76.35	**93.12**	77.50	80.52	0.8940
DenseNet-40	**87.24**	83.11	90.48	**87.23**	85.12	0.9420
DenseNet-121	86.35	84.4	87.83	84.46	84.46	0.8870
DenseNet-161	86.05	78.38	92.06	88.55	83.15	0.9304

Table 6
The results of all CNN architectures using Image type 3. Bold indicates the best results training and testing on this image type.

Method	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
VGG-Like	**84.27**	81.08	86.77	82.76	81.91	0.8989
VGG-16	83.97	85.14	83.07	79.75	**82.35**	**0.9076**
ResNet-18	81.60	**86.49**	77.77	76.58	79.08	0.8709
ResNet-50	81.01	81.76	80.42	77.03	77.03	0.8680
ResNet-101	79.82	77.03	82.01	**84.16**	68.27	0.8549
DenseNet-40	76.56	57.43	**91.53**	78.57	80.13	0.8932
DenseNet-121	82.19	81.76	82.54	77.63	78.67	0.8872
DenseNet-161	81.01	79.73	82.01	75.54	73.17	0.8533

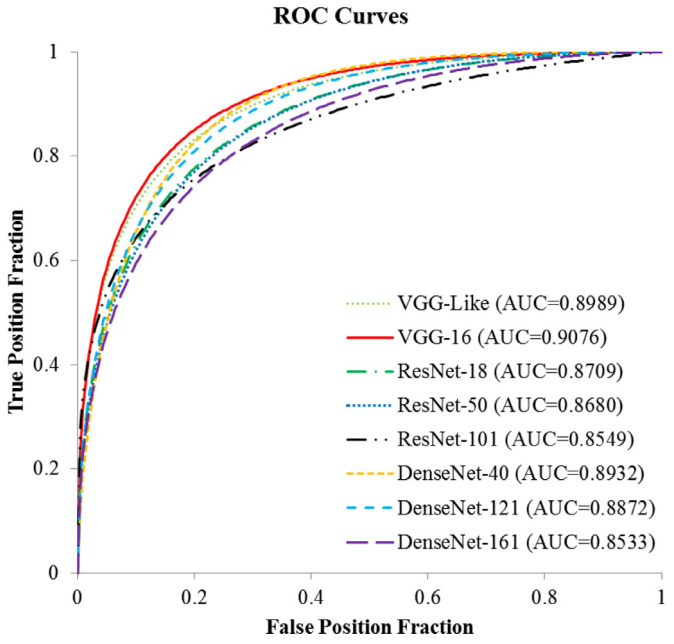

Fig. 11. The ROC Curves of all CNN architectures testing on Image type 3. The VGG-16 has the highest AUC value (0.9076) than other CNN structures.

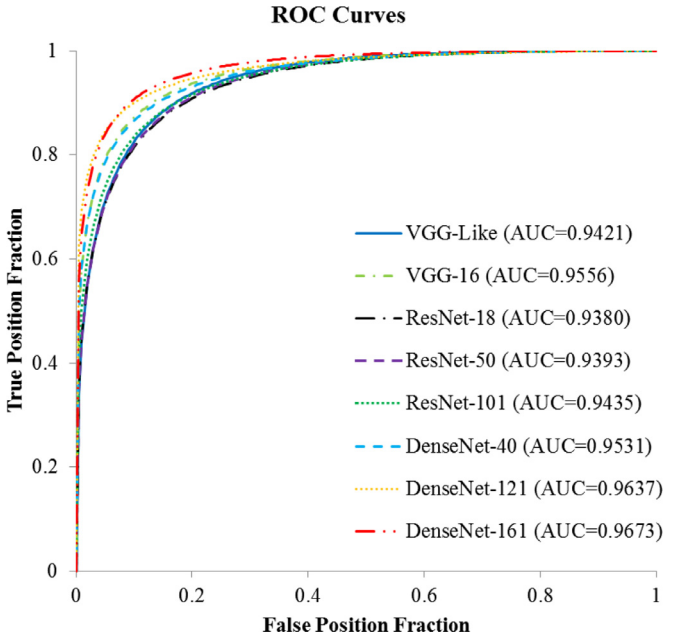

Fig. 12. The ROC Curves of all CNN architectures testing on Image type 4. The DenseNet-161 has the highest AUC value (0.9673) than other CNN structures.

the VGGNet models have a higher SPEC than others in Image type 4.

Second, the ACC, SEN, SPEC, Precision, F1 score and AUC of ResNet-18 based on Image type 4 were 86.65%, 81.76%, 90.48%, 87.05%, 84.32% and 0.9380, respectively, when using ResNet-50 were 86.05%, 86.49%, 85.71%, 82.58%, 84.49% and 0.9393, respectively, when using ResNet-101 were 86.05%, 84.46%, 87.30%, 83.89%, 84.18% and 0.9435, respectively.

Third, the ACC, SEN, SPEC, Precision, F1 score and AUC of DenseNet-40 based on Image type 4 were 89.32%, 85.81%, 92.06%, 89.44%, 87.59% and 0.9531, respectively, when using DenseNet-121 were 89.32%, 87.84%, 90.48%, 87.84%, 87.84% were 0.9637, respectively, when using DenseNet-161 were 90.80%, 89.86%, 91.53%, 89.26%, 89.56% and 0.9673, respectively. We could find that the DenseNet family has a higher F1 Score than others in Image type 4.

Table 7 shows that DenseNet-161 has the highest AUC value and ResNet-18 gets the lowest one. Fig. 12 illustrated the ROC curves and AUC value for all CNN architectures testing on Image type 4.

4.2. Comparison of different ensemble method strategy (SNUH dataset)

According to the above results, we obtained four machines that provided the best performance based on different image types. Hence, we used these four machines as our base machines in the ensemble method. In addition, the combining strategy in the ensemble method also affects the final predict results. Hence, we

Table 7
The results of all CNN architectures using Image type 4. Bold indicates the best results training and testing on this image type.

Method	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
VGG-Like	86.05	80.41	90.48	86.86	83.51	0.9421
VGG-16	88.72	83.78	**92.59**	**89.86**	86.71	0.9556
ResNet-18	86.65	81.76	90.48	87.05	84.32	0.9380
ResNet-50	86.05	86.49	85.71	82.58	84.49	0.9393
ResNet-101	86.05	84.46	87.30	83.89	84.18	0.9435
DenseNet-40	89.32	85.81	92.06	89.44	87.59	0.9531
DenseNet-121	89.32	87.84	90.48	87.84	87.84	0.9637
DenseNet-161	**90.80**	**89.86**	91.53	89.26	**89.56**	**0.9673**

Table 8
The highest accurate CNN method from each image type.

Method	Image type	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
DenseNet-121	1	86.35	77.70	93.12	89.84	83.33	0.9248
DenseNet-40	2	87.24	83.11	90.48	87.23	85.12	0.9420
VGG-Like	3	84.27	81.08	86.77	82.76	81.91	0.9423
DenseNet-161	4	90.80	89.86	91.53	89.26	89.56	0.9673

used the different combining strategies in our ensemble method, including the unweighted average (UA), weighted average (WA), weighted voting (WV) and stacking (S) method, and we also analyze and compare the results between different combine strategies. The UA strategy averages the predict results from each base machine, the WA strategy averages the predict results from each base machine with weights (in this study, the better base machine has the larger weight value). The WV strategy is the voting method with weights, the S strategy combines results from base machines via a regression process.

Results of base machine selection. Table 8 shows the highest diagnostic accuracy from the CNN method trained in each image type and we select these CNN methods as our base machine for the ensemble method. The DenseNet-121 has the best performance presentation in Image type 1 and the ACC, SEN, SPEC, Precision and F1 score were 86.35%, 77.70%, 93.12%, 89.84%, and 83.33%, respectively. The DenseNet-40 has the best performance presentation in Image type 2 and the ACC, SEN, SPEC, Precision and F1 score were 87.24%, 83.11%, 90.48%, 87.23%, and 85.12%, respectively. The VGG-Like has the best performance presentation in Image type 3 and the ACC, SEN, SPEC, Precision and F1 score were 84.27%, 81.08%, 86.77%, 82.76%, and 81.91%, respectively. The DenseNet-161 has the best performance presentation in Image type 4 and the ACC, SEN, SPEC, Precision and F1 score were 90.80%, 89.86%, 91.53%, 89.26%, and 89.56%, respectively.

Combining strategy. In this section, we showed all the combining strategy results including UA, WA, WV, and S. The ACC, SEN, SPEC, Precision, F1 score, and AUC of the UA strategy were 88.13%, 79.05%, 95.24%, 92.86%, 85.40%, and 0.9662, respectively. The ACC, SEN, SPEC, Precision, F1 score and AUC of the WA strategy were 91.10%, 85.14%, 95.77%, 94.03%, 89.36%, and 0.9697, respectively. The ACC, SEN, SPEC, Precision, F1 score and AUC of the WV strategy were 90.20%, 87.84%, 92.06%, 89.66%, 88.74%, and 0.8995, respectively. The ACC, SEN, SPEC, Precision, F1 score and AUC of the S strategy were 89.02%, 83.78%, 93.12%, 90.51%, 87.02%, and 0.9619, respectively. Because base machine 4 has the highest accuracy in the single classifier experiment, so we set the highest weight value for base machine 4, the weighting table used by the WA and WV as illustrated in Table 9.

The comparison with performances corresponding to different combine strategies are listed in Table 10, and the WA strategy obtained the highest AUC value (0.9697) and S strategy has the worst one (0.9619). Fig. 13 illustrated the ROC curves and AUC values for all combine strategies. The *p*-values using the Student's t-test between all of the ensemble methods with different combine strategies, including the UA, WA, WV, and S, are listed in Table 11.

Table 9
The weighting table used by the WA and WV combining strategy.

Method	Image type	Weight value by WA	Weight value by WV
DenseNet-121	1	0.13	1
DenseNet-40	2	0.35	1
VGG-Like	3	0.02	1
DenseNet-161	4	0.5	2

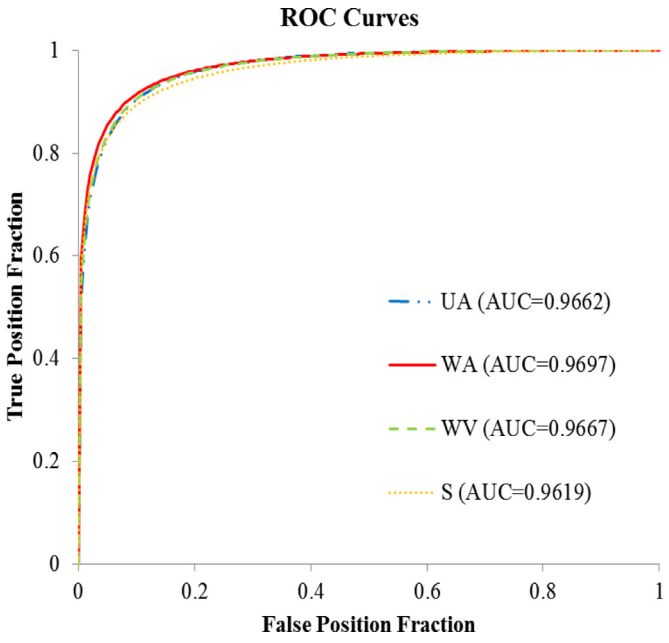

Fig. 13. The ROC Curves of different combining strategies, including WA, UA, S, and WV. WA has the highest AUC value (0.9697) than other combining strategies.

Performance of the proposed method. According to the above results, the ensemble method with the WA combining strategy obtained the best diagnostic accuracy. In our study, the ensemble method consisted of four base machines, which including the DenseNet-121 model using the original tumor image (as base machine 1), the DenseNet-40 model using the segmented tumor image (as base machine 2), the VGG-Like model using TSI (as base machine 3), and the DenseNet-161 model using 3-channel images

Table 10
The results of using all base machines with different combine strategies, including unweighted average (UA), weighted average (WA), weighted voting (WV) and stacking (S) method. Bold indicates the best ensemble method results.

Method	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
UA	88.13	79.05	95.24	92.86	85.40	0.9662
WA	**91.10**	85.14	**95.77**	**94.03**	**89.36**	**0.9697**
WV	90.20	**87.84**	92.06	89.66	88.74	0.9667
S	89.02	83.78	93.12	90.51	87.02	0.9619

Table 11
The *p-values* for all the ensemble combining strategy including the unweighted average (UA), weighted average (WA), weighted voting (WV), and stacking (S) method.

Method			ACC	SEN	SPEC	AUC
UA	vs	WA	0.207	0.172	0.804	0.3988
UA	vs	S	0.716	0.296	0.380	0.0371*
UA	vs	WV	0.386	0.042*	0.206	0.4522
WA	vs	S	0.368	0.748	0.262	0.6
WA	vs	WV	0.691	0.496	0.132	0.5079
S	vs	WV	0.614	0.318	0.694	0.1130

* stands for statistically significant.

(as base machine 4). From the experiment results, base machine 4 has the best diagnostic performance. Finally, the base machines used the WA combine strategy has excellent diagnostic performance and the ACC, SEN, SPEC, Precision, F1 score and AUC were 91.10%, 85.14%, 95.77%, 94.03%, 89.36% and 0.9697, respectively.

4.3. Comparison of prediction performance in BUSI dataset

For comparison of the prediction performance, we also used the open dataset (dataset BUSI) to examine in our proposed method. First, the ACC, SEN, SPEC, Precision, F1 score and AUC on Image type 1 (the original tumor image) were 88.46%, 83.78%, 90.32%, 77.50%, 80.52%, and 0.9052, respectively. Second, the ACC, SEN, SPEC, Precision, F1 score and AUC on Image type 2 (the segmented tumor image) were 90.77%, 88.89%, 91.49%, 80.00%, 84.21%and 0.8857, respectively. Third, the ACC, SEN, SPEC, Precision, F1 score and AUC on Image type 3 (the TSI) were 85.38%, 76.92%, 89.00%, 75.00%, 75.95%, and 0.8660, respectively. Forth, the ACC, SEN, SPEC, Precision, F1 score and AUC on Image type 4 (the fused image) were 94.62%, 92.31%, 95.60%, 90.00%, 91.14%, and 0.9711, respectively. Finally, we also used the ensemble method with the weighted average to combine the predict results from above base machines, the ACC, SEN, SPEC, Precision, F1 score and AUC were 90.77%, 96.67%, 89.00%, 72.50%, 82.86%, and 0.9489, respectively. The predict results of BUSI dataset using our proposed method are listed in Table 12, and the corresponding ROC curves with AUC values are shown in Fig. 14.

5. Conclusion and discussion

In recent years, several studies [7–10] had been published breast cancer diagnosis methods to classify benign and malignant tumors in US images. In recent researches, several studies [11,36,47–49] had been published to diagnosis breast cancer based on a convolution neural network. Many studies [11,36,48] enhance the diagnosis performance by extending the model capability through designed different CNN architectures or different machine learning methods. Furthermore, some researches [50,51] were combined with conventional handheld features or add more tumor information (margin tissue, tumor hardness, etc.) to their diagnosis method. Hence, we develop a CAD system to diagnosis the breast tumor using the ensemble method with reinforced image content representation information.

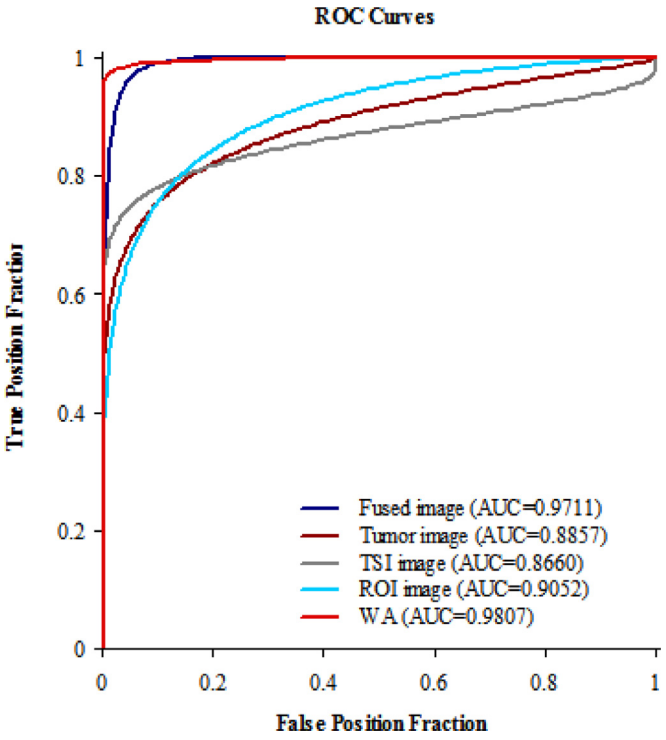

Fig 14. The ROC curves of the BUSI dataset using our proposed method.

In our study, we propose a CAD system for tumor diagnosis using an image fusion method combined with different image content representations and ensemble different CNN architectures on US images. First, the operator extracts the ROI area which covers the whole tumor and the ROI boundary is close to the tumor margin. Then, we extract the tumor region image and the TSI. In addition, we generated 3-channels fused image from different images, which enhance the tumor features further to increase the diagnostic performance of our CAD system. Then, we trained CNN models to learn benign and malignant tumor features from all types of images.

We employed many CNN models on different data sets, including original tumor image, segmented tumor images, tumor mask, and fused image, to compare the diagnostic results. Our method not only used different CNN models (including VGG-16, VGG-Like, ResNet and DenseNet) to focus on different data sets but also find out which CNN model is more suitable for dealing with a specific type of tumor image. Finally, we selected the base machines from different CNN models, which provide the highest predictive performance in each image type. Furthermore, we used the ensemble method to integrate the results produced by different base machines. In our experiment results, the diagnostic performance of the ensemble method used the WA combined strategy in the SNUH dataset showed the accuracy, sensitivity, specificity, and AUC are 91.10%, 85.14%, 95.77%, and 0.9697, respectively. The diagnostic performance of the ensemble method used the WA combined strategy

Table 12
The results of BUSI dataset using our proposed method.

Image type	Method	ACC (%)	SEN (recall) (%)	SPEC (%)	Precision (%)	*F*1 score (%)	AUC
1	DenseNet-121	88.46	83.78	90.32	77.50	80.52	0.9052
2	DenseNet-40	90.77	88.89	91.49	80.00	84.21	0.8857
3	VGG-Like	85.38	76.92	89.00	75.00	75.95	0.8660
4	DenseNet-161	94.62	92.31	95.60	90.00	91.14	0.9711
Ensemble (WA)		90.77	96.67	89.00	72.50	82.86	0.9489

in the open BUSI dataset showed the accuracy, sensitivity, specificity, and AUC is 90.77%, 96.67%, 89.00%, and 0.9489, respectively.

The results in both datasets showed the fused image has more image information that helps the deeper network architecture to provide better performance. We concatenate three types of images into a 3-channel image (RGB-style) included three different information, and the results showed that combining more useful information by image fusion method has improved the performance of CAD diagnostics significantly. On the other hand, the deeper network architecture cannot improve performance on the TSI due to the TSI only contains the image information indicating the background in black and the tumor in white, and the features that can be learned by CNN are fewer than other data sets. From the TSI results, we observed that the single features of the tumor shape are not enough for diagnosis, but it still helps for assisting diagnosis. According to the results of the ensemble method and four base machines, the diagnostic performance of the ensemble method was better than other models which using single CNN architecture. Hence, the ensemble method with the weighted average can reduce the variability of the diagnostic results and achieve the best diagnosis.

In addition, the performance of VGG-Like being very similar to VGG-16. The reasons might include: (1) the deeper conventional CNN architectures might not be able to extract more detailed features from a smaller image. (2) Using deeper conventional CNN with more parameters could easily have overfitting problems on small data sets. Therefore, ResNet and DenseNet used skip connection to achieve message rectification, avoiding inter-layer transmission loss and resolving gradient disappearance problem, because the skip connection technique could enhance the deep CNN architecture to learn more feature effectively. In our experimental materials, although the SNUH dataset was acquired from 8 different US machines with distinctive characteristics, the results of our study showed the robustness of our proposed system. The results show that the open BSUI dataset using our CAD system for breast tumor diagnosis also performing well; the ROC curve shows that the ensemble method could reduce the false positive rate.

Currently, there are still some limitations in this study, the ROI region and tumor contour of B-mode US image are cropped by the experts-defined, and it may result in a different ROI region and tumor contour from different operators. Therefore, we will develop an automatic detection and segmentation method to find out the tumor locations and extract the tumor automatically to reduce the effect of human interventions. In the future, we will consider the effects of the surrounding tissue or tumor environment for our CAD system, and the surrounding tissue or tumor environment might provide useful information to help the diagnostic performance.

In conclusion, we proposed a CAD system to diagnose the breast tumor using original tumor images, segmented tumor images, tumor masks, and fused images in the US. Our study demonstrated the CAD system based on CNN architecture combining multiple tumor features could provide a precise result to diagnose a tumor in the patient with breast cancer.

Declaration of Competing Interest

The authors declare that they have no financial and personal relationships with other people or organizations that could inappropriately influence their work.

Acknowledgment

The authors thank the Ministry of Science and Technology of Taiwan (MOST 107-2634-F-002-013,MOST 108-2634-F-002-010, and MOST 109-2634-F-002-026) for financial support.

References

[1] T. Tan, B. Platel, H. Huisman, C.I. Sánchez, R. Mus, N. Karssemeijer, Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation, IEEE Trans. Med Imaging 31 (2012) 1034–1042.
[2] H.-D. Cheng, J. Shan, W. Ju, Y. Guo, L. Zhang, Automated breast cancer detection and classification using ultrasound images: a survey, Pattern Recognit. 43 (2010) 299–317.
[3] M. Samulski, R. Hupse, C. Boetes, R.D. Mus, G.J. den Heeten, N. Karssemeijer, Using computer-aided detection in mammography as a decision support, Eur Radiol 20 (2010) 2323–2330.
[4] L.A. Meinel, A.H. Stolpen, K.S. Berbaum, L.L. Fajardo, J.M. Reinhardt, Breast MRI lesion classification: improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system, J. Magnetic Resonance Imaging 25 (2007) 89–95.
[5] B. Sahiner, H.-P. Chan, M.A. Roubidoux, L.M. Hadjiiski, M.A. Helvie, C. Paramagul, et al., Malignant and benign breast masses on 3D US volumetric images: effect of computer-aided diagnosis on radiologist accuracy, Radiology 242 (2007) 716–724.
[6] H.-P. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, et al., Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study, Radiology 212 (1999) 817–827.
[7] J. Shan, S.K. Alam, B. Garra, Y. Zhang, T. Ahmed, Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods, Ultrasound Med. Biol. 42 (2016) 980–988.
[8] D.-R. Chen, R.-F. Chang, C.-J. Chen, M.-F. Ho, S.-J. Kuo, S.-T. Chen, et al., Classification of breast ultrasound images using fractal feature, Clin. Imaging 29 (2005) 235–245.
[9] H.-W. Lee, B.-D. Liu, K.-C. Hung, S.-F. Lei, P.-C. Wang, T.-L. Yang, Breast tumor classification of ultrasound images using wavelet-based channel energy and imageJ, IEEE J. Sel. Top. Signal Process 3 (2009) 81–93.
[10] P.-H. Tsui, Y.-Y. Liao, C.-C. Chang, W.-H. Kuo, K.-J. Chang, C.-K. Yeh, Classification of benign and malignant breast tumors by 2-D analysis based on contour description and scatterer characterization, IEEE Trans. Med. Imaging 29 (2010) 513–522.
[11] F. Hu, G.-S. Xia, J. Hu, L. Zhang, Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery, Remote Sens. (Basel) 7 (2015) 14680–14707.
[12] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 142–158.
[13] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (2014) 580–587.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Processing Syst. (2012) 1097–1105.
[15] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, et al., Deep convolutional neural networks for multi-modality isointense infant brain image segmentation, Neuroimage. 108 (2015) 214–224.
[16] W.K. Moon, Y.-W. Lee, Y.-S. Huang, S.H. Lee, M.S. Bae, A. Yi, et al., Computer-aided prediction of axillary lymph node status in breast cancer using tumor surrounding tissue features in ultrasound images, Comput. Methods Programs Biomed. 146 (2017) 143–150.

[17] H.R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, et al., Improving computer-aided detection using convolutional neural networks and random view aggregation, IEEE Trans. Med. Imaging 35 (2016) 1170–1181.
[18] M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, et al., Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion, Med. Phys. 46 (2019) 746–755.
[19] M.H. Yap, M. Goyal, F. Osman, E. Ahmad, R. Martí, E. Denton, et al., End-to-end breast ultrasound lesions recognition with a deep learning approach, Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, 2018.
[20] M.H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, et al., Automated breast ultrasound lesions detection using convolutional neural networks, IEEE J. Biomed. Health Inform. 22 (2017) 1218–1226.
[21] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief. (2019) 104863.
[22] W.-J. Wu, W.K. Moon, Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features, Acad. Radiol. 15 (2008) 873–880.
[23] Y.L. Huang, D.R. Chen, Y.R. Jiang, S.J. Kuo, H.K. Wu, W. Moon, Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound, Ultrasound Obstetrics Gynecol. 32 (2008) 565–572.
[24] R.M. Rangayyan, N.M. El-Faramawy, J.L. Desautels, O.A. Alim, Measures of acutance and shape for classification of breast tumors, IEEE Trans. Med. Imaging 16 (1997) 799–810.
[25] A.T. Stavros, D. Thickman, C.L. Rapp, M.A. Dennis, S.H. Parker, G.A. Sisney, Solid breast nodules: use of sonography to distinguish between benign and malignant lesions, Radiology 196 (1995) 123–134.
[26] C. Sohn, J.-U. Blohmer, U. Hamper, Breast Ultrasound: A Systematic Approach to Technique and Image Interpretation, Thieme, 1999.
[27] S. Liu and Z. Liu, "Multi-channel CNN-based object detection for enhanced situation awareness," *arXiv preprint arXiv:1712.00075*, 2017.
[28] M.A. Smeelen, P.B. Schwering, A. Toet, M. Loog, Semi-hidden target recognition in gated viewer images fused with thermal IR images, Inf. Fusion 18 (2014) 131–147.
[29] J. Han, B. Bhanu, Fusion of color and infrared video for moving human detection, Pattern Recognit. 40 (2007) 1771–1784.
[30] Y. Niu, S. Xu, L. Wu, W. Hu, Airborne infrared and visible image fusion for target perception based on target region segmentation and discrete wavelet transform, Math. Prob. Eng. 2012 (2012).
[31] G. Bhatnagar, Z. Liu, A novel image fusion framework for night-vision navigation and surveillance, Signal Image Video Process 9 (2015) 165–175.
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[33] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (1994) 157–166.
[34] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[35] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertain., Fuzziness Knowl.-Based Syst. 6 (1998) 107–116.
[36] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (2016) 1285–1298.
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[38] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI, 2017, p. 12.
[39] J. Dai, Y. Li, K. He, J. Sun, in: R-fcn: object detection via region-based fully convolutional networks, 2016, pp. 379–387.
[40] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 834–848.
[41] G. Huang, Z. Liu, K.Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, p. 3.
[42] Y. Liu, X. Yao, Ensemble learning via negative correlation, Neural netw. 12 (1999) 1399–1404.
[43] Y. Ganjisaffar, R. Caruana, C.V. Lopes, Bagging gradient-boosted trees for high precision, low variance ranking models, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 85–94.
[44] Z.-H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artif. Intell. 137 (2002) 239–263.
[45] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874.
[46] C. Van Rijsbergen, "Information retrieval. dept. of computer science, university of glasgow," *URL:* citeseer.ist.psu.edu/vanrijsbergen79information.html, vol. 14, 1979.
[47] B. Huynh, K. Drukker, M. Giger, MO-DE-207B-06: computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks, Med. Phys. 43 (2016) 3705-3705.
[48] T. Xiao, L. Liu, K. Li, W. Qin, S. Yu, Z. Li, Comparison of transferred deep neural networks in ultrasonic breast masses discrimination, Biomed. Res. Int. 2018 (2018).
[49] X. Xie, F. Shi, J. Niu, X. Tang, Breast ultrasound image classification and segmentation using convolutional neural networks, in: Pacific Rim Conference on Multimedia, 2018, pp. 200–211.
[50] Q. Zhang, Y. Xiao, W. Dai, J. Suo, C. Wang, J. Shi, et al., Deep learning based classification of breast tumors with shear-wave elastography, Ultrasonics 72 (2016) 150–157.
[51] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, et al., A deep learning framework for supporting the classification of breast lesions in ultrasound images, Phys. Med. Biol. 62 (2017) 7714.