Journal of King Saud University – Computer and Information Sciences 34 (2022) 10273–10292

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Deep Learning-based Multi-stage segmentation method using ultrasound images for breast cancer diagnosis

Se Woon Cho, Na Rae Baek, Kang Ryoung Park *

Division of Electronics and Electrical Engineering, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, South Korea

ARTICLE INFO

Article history:
Received 12 July 2022
Revised 19 October 2022
Accepted 20 October 2022
Available online 27 October 2022

Keywords:
Breast cancer
Ultrasound image
Breast tumor segmentation
BTEC-Net
RFS-UNet

ABSTRACT

Globally, breast cancer occurs frequently in women and has the highest mortality rate. Owing to the increased need for a rapid and reliable initial diagnosis of breast cancer, several breast tumor segmentation methods based on ultrasound images have attracted research attention. Most conventional methods use a single network and demonstrate high performance by accurately classifying tumor-containing and normal image pixels. However, tests performed using normal images have revealed the occurrence of many false-positive errors. To address this limitation, this study proposes a multistage-based breast tumor segmentation technique based on the classification and segmentation of ultrasound images. In our method, a breast tumor ensemble classification network (BTEC-Net) is designed to classify whether an ultrasound image contains breast tumors or not. In the segmentation stage, a residual feature selection UNet (RFS-UNet) is used to exclusively segment images classified as abnormal by the BTEC-Net. The proposed multistage segmentation method can be adopted as a fully automated diagnosis system because it can classify images as tumor-containing or normal and effectively specify the breast tumor regions.

1. Introduction

Deep learning-based cancer diagnosis systems have been frequently considered in the medical imaging field (Xue et al., 2021). Among the various cancers, breast cancer frequently occurs in women, accounting for approximately 25 % of the cancers that occur in women and has the highest mortality rate. Further, the incidence of breast cancer in women is increasing every year. Breast cancer occurs when cells in the breast grow uncontrollably, thereby forming a tumor.

Tumors can be categorized as either benign or malignant. Noncancerous benign tumors are not generally aggressive toward the surrounding tissues and do not spread outside the breast. In contrast, malignant tumors are cancerous and aggressive because they spread and damage the surrounding tissues Notably, the patient survival rate increases significantly in cases in which breast cancer tumors are diagnosed early. Therefore, techniques that facilitate the timely initial diagnosis of breast cancer have attracted significant research attention. A variety of medical imaging methods have been used for breast cancer diagnosis, and various treatment methods based on these imaging techniques have been studied. Breast ultrasonography-based techniques are inexpensive. Moreover, compared with other medical imaging techniques, they facilitate detailed breast tumor diagnoses for each part of the breast (Wang and Yao, 2022). Ultrasound images are commonly adopted in conventional medical imaging, and many computer-aided diagnostic (CAD) methods have been developed. Furthermore, various convolutional neural network (CNN)-based segmentation models have been actively studied. However, the accurate segmentation of breast tumor cells with ultrasound images remains challenging owing to speckle noise, low contrast, unclear boundaries, and different tumor sizes and shapes (Wang and Yao, 2022). Furthermore, most conventional methods trained a single network for tumor pixel segmentation. This significantly reduces the segmentation performance during testing because of the large occurrence of false-positive pixels in normal images that do not contain tumors.

Therefore, we propose a multistage-based breast tumor segmentation technique using ultrasound images. Our method can be used to perform both classification and segmentation. In the first stage (classification stage), a breast tumor ensemble classification network (BTEC-Net) is used to perform binary classification using the breast ultrasound images of patients to determine

* Corresponding author.
E-mail address: parkgr@dongguk.edu (K.R. Park).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.10.020
1319-1578/

presence of tumors. The BTEC-Net uses two sub-networks as feature extractors: multi-scale squeeze and excitation (MSSE) DenseNet121 (MSSE-DenseNet121) and ResNet101 (MSSE-ResNet101). High-quality features are extracted via the squeeze and excitation (SE) blocks and multi-scale feature connections. The BTEC-Net shows the best performance compared to existing methods through the learning-based feature-level fusion. Notably, these features of our proposed network are novel and have not been reported by previous studies. In the second stage (segmentation stage), a residual feature selection UNet (RFS-UNet) is used to perform segmentation on the images classified as abnormal. Unlike in various U-Net-based models, the performance of RFS-UNet improves when an RFS module is used. The RFS module can be used in the residual connection part of the network to learn the ratio of adding residual features. Thus, the proposed method identifies normal images in advance, significantly reducing the occurrence of false positive errors. The proposed method can be used as a practical, fully automated CAD system because radiologists need not classify the image as tumor-containing or normal in advance or specify the region of interest (ROI) of lesions. Furthermore, it facilitates personalized medication, which could assist in cancer diagnosis and treatment customized according to the severity, type, and size of the tumor for each patient.

- Compared with previous methods, the proposed multi-stage segmentation method significantly reduces the occurrence of segmentation (false-positive) errors in normal-case images.
- The proposed MSSE-DenseNet121 and MSSE-ResNet101 networks are designed by applying the SE block to conventional DenseNet121 and ResNet101 models. In addition, this study presents a BTEC-Net that combines two subnetworks through learning-based feature-level fusion.
- We propose an RFS-UNet for breast tumor segmentation. The RFS module, which could be applied to the residual connection, is designed to improve breast tumor segmentation performance.
- Our trained CNN models and algorithms are publicly available upon request, allowing other researchers to use them.

2. Related work

Conventional methods of segmenting breast ultrasound images have been studied as CAD systems to assist radiologists. However, they have the following limitations: low contrast, speckle noise, unclear tumor boundaries, and different shapes and sizes of tumor regions in ultrasound images. Therefore, discriminating normal tissues from breast tumors accurately at the pixel level is challenging.

Handcrafted-feature-based methods extract unique breast tumor features or features that are distinguished from the background in an ultrasound image to segment the breast tumor region. A previous study detected lesion boundaries using marker-controlled watershed transformation and applied the function of an average radial derivative to locate lesion contours (Gómez et al., 2010). Log-Gabor filters were applied to attain texture features with the classifier of linear discriminant analysis to discriminate the tumor in each local region and perform segmentation (Gómez-Flores and Ruiz-Ortega, 2016). To leverage both domains, a study has proposed multi-domain features that comprise the phase features in a frequency domain, as well as the features of a radial distance and texture in a spatial domain (Shan et al., 2012a). A neutrosophic l-means clustering algorithm (Shan et al., 2012b) has been proposed based on the traditional clustering and neutrosophy. Moreover, the Accuracy was enhanced by using phase features. In a vector quantization-based method (Kekre and Shrinath, 2013), a probability image was generated according to the appearance frequency of the pixel values, and the tumor region was segmented using a clustering algorithm. Robust graph-based segmentation algorithms (Huang et al., 2012; Huang et al., 2014) were proposed and used to improve segmentation performance. These methods are used in various medical fields and have been applied to mental health studies. In superpixel-based methods (Huang et al., 2020), a simple algorithm of linear and iterative clustering is adopted for the decomposition of a breast ultrasound image into superpixels. Subsequently, the superpixels were classified as tumor-containing or normal to perform segmentation. Handcrafted feature-based methods use different filters and algorithms to extract tumor features and perform segmentation. However, the use of high-dimensional features in these methods is challenging, because these filters or algorithms are designed by humans. Furthermore, optimal parameters must be adjusted manually according to the environment (Xue et al., 2021).

Considering the issues raised by handcrafted feature-based methods, several investigations concerning the use of deep learning have recently been studied. In deep-feature-based methods, CNN models with various structures are designed to facilitate the efficient extraction of high-dimensional features. Several high-performance automatic segmentation methods have been proposed. In particular, residual units and dilation convolution have been widely used to reduce information loss and focus on learning tumor pixels. A selective kernel U-Net has been proposed to replace the convolutional blocks of a conventional U-Net with SK blocks. These SK blocks use both dilated and conventional convolutions, whereby the reflecting ratio of each output feature map can be determined based on attention coefficients (Byra et al., 2020). In another study, a U-Net-based backbone network was adopted, supplemented by an ASP module (Zhu et al., 2020). In this combination network, the first branch performs convolutions for all the feature map regions. The second branch divides the feature maps into four subregions to perform independent convolutions. This sequence of operations ensures the efficient use of the global and local features of an input image. In the improved attention UNet (Abraham and Khan, 2019), the authors used a focal Tversky loss function and applied multiscale inputs along with a deep supervision technique. A boundary detection (BD) module and a global guidance (GG) block have been proposed to increase the performance. The GG block extracts feature maps from previous convolutional blocks. Subsequently, these feature maps can be used in combination with those corresponding to the last layer. The BD module can be connected to each convolutional layer and used to identify the boundaries of the tumor region (Xue et al., 2021). In segmentation-based preprocessing method (Wang and Yao, 2022), authors proposed to perform both segmentation and detection. First, tumor pixels are found by applying the improved UNet to the preprocessed image. By using the binary image obtained by performing segmentation, the contrast of the original image is further increased. In the detection stage, the location of the tumor is found using an anchor-free detection network. In this method, the contrast of the tumor area was increased through the segmentation technique. In the cascaded CNN method (Chen et al., 2022), a C-Net composed of multiple networks in a cascade method was proposed. This model consists of three networks and is trained by applying the deep supervision technique. The first model, UNet, generates and delivers saliency maps. After that, it goes through the attention network and refinement network to output the segmentation result. Through this structure, a deeper model can be designed, and the function of each sub-network can be subdivided. In transformer-based method (Shen et al., 2022), the author proposed a dilated transformer (DT) and compared it with existing CNN models. The encoder of the DT model consists of transformer layers, and the residual connection is applied to the axial attention technique. In addition, dilated convolution was used to compress global features

and expand the receptive field. Through this, it showed better performance than the existing CNN or medical transformer model.

A previous study closely related to self-supervised learning (CR-SSL) proposed a method that utilizes a small amount of annotated labeling data when training a model. Through UNet-based segmentation, it was proven that this method showed better performance when using CR-SSL (Mishra et al., 2022). LAEDNet (Zhou et al., 2022) was designed as a relatively light model, and exhibited good performance through an encoder–decoder structure based on EfficientNet. In addition, the object segmentation performance was more accurate when the SE block was used in the decoder. In a multilevel context refinement network (MCRNet) (Lou et al., 2022), the authors improved the segmentation performance through blocks that reduced feature gaps between encoders and decoders and improved context correlation. Previous segmentation methods achieved high performance using CNN models.

However, because they involve the training of a single segmentation network to differentiate tumor pixels from normal pixels, false positive errors can be increased significantly when the test datasets include normal images that do not contain any tumors. Therefore, this study proposed a multi-stage breast tumor segmentation method based on segmentation after the classification of breast tumor images. In the first stage, BTEC-Net classified an input image as abnormal if it contained tumor cells and normal if it did not. In the output results of the images classified as normal, all the pixels were processed as normal and assigned a value of zero. In the second stage, segmentation using RFS-UNet was performed exclusively on images classified as abnormal. This enabled us to distinguish whether the breast ultrasound image was normal or abnormal in advance, thereby significantly improving segmentation performance by reducing the occurrence of false-positive errors when operating on datasets that include normal images.

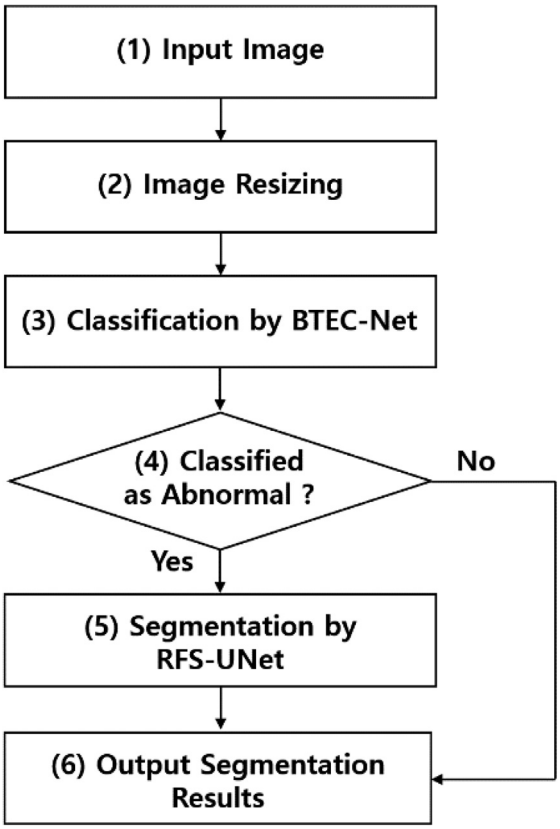

Fig. 1. Flowchart of proposed algorithm.

3. Proposed method

3.1. Overall procedure

We describe the procedure of the proposed algorithm in detail (Fig. 1). In step (1), a breast ultrasound image is captured as the input image. In step (2), considering the computational cost, the input image is resized by 224 × 224 pixels. In step (3), binary classification is performed on the input breast ultrasound image.

By using our BTEC-Net, the input image was classified as either normal (0) or abnormal (1), depending on whether it contained any breast tumor cells. In step (4), normal images are excluded based on the classification result, whereas abnormal images are provided as inputs to the next stage. As shown in step (5), segmentation is performed on the abnormal images using the RFS-UNet. Finally, in step (6), our proposed model outputs a 1-channel segmentation map, wherein the tumor and normal pixels are segmented as 1 and 0, respectively. For normal images, the segmentation output assigns a value of zero to all the pixels.

In the various steps of our method, most operations, which can be expressed as floating-point operations (FLOPs), are occupied by the BTEC-Net and RFS-UNet. The overall computational complexity of the BTEC-Net is 10.0705 Giga FLOPs (GFLOPs), and most of the computations are in the feature extractor part. In the case of the RFS-UNet, the computational complexity is 17.7097 GFLOPs, which requires more calculations than the BTEC-Net.

3.2. Stage 1: Breast tumor image classification

3.2.1. Breast tumor ensemble classification network (BTEC-Net)

First, we describe the proposed breast tumor classification network, BTEC-Net. This network is implemented by improving the original DenseNet121 (Huang et al., 2017) and ResNet101 (He et al., 2016) classification networks and combining them in an ensemble manner. The input to this stage comprises breast ultrasound images resized to 224 × 224 pixels. Each of the four blocks of the two networks is connected to the SE block and global average pooling (GAP). Then, the feature maps are combined in channel direction. Because the combined feature vectors of each network have different lengths, the number of dimensions is matched by applying a 1024-dimensional fully connected (FC) layer.

In a feature-level fusion module (FFM), both extracted feature vectors are combined in the channel direction before performing feature-level fusion. The FFM module comprises an FC layer with one output node and a sigmoid layer. The FC layer learns the weights of the 2048-dimensional feature vector and outputs the probabilities of the class through the sigmoid. Subsequently, the images containing benign and malignant tumors are classified as abnormal (1), whereas those not containing any tumor cells are classified as normal (0). More detailed descriptions of the components of the BTEC-Net are provided in Sections 3.2.2 to 3.2.5.

Fig. 2 depicts the proposed classification workflow. The BTEC-Net is an ensemble model that combines the features of two subnetworks. Both networks are relatively large models often used as baselines. When both models are trained simultaneously, the hyperparameter setting becomes challenging. Moreover, the model may fall into the local minimum during the training process. Therefore, the MSSE-DenseNet121 and MSSE-ResNet101 are pre-trained independently using training data for the classification of abnormal and normal breast ultrasound images. Each subnetwork is used to obtain optimal features. By pre-training each model individually, different features for the same input can be extracted. When combining features using the FFM, more diverse and highly distinguishing features can be utilized.

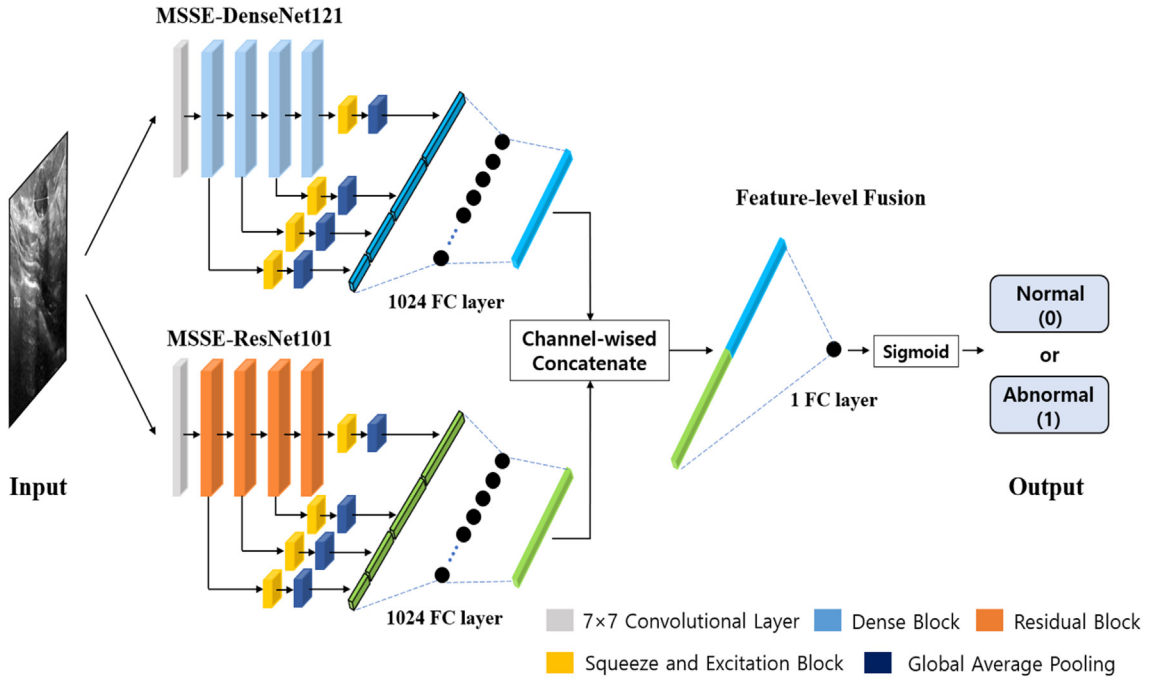

Fig. 2. Image classification workflow of proposed BTEC-Net.

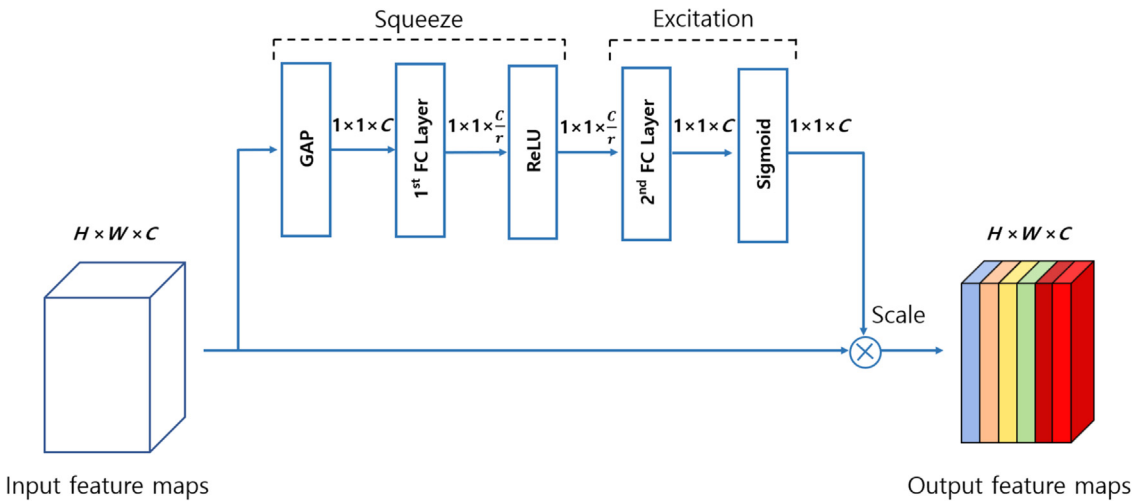

Fig. 3. SE block architecture.

3.2.2. Feature recalibration using SE block

The SE block (Hu et al., 2018) comprises two parts: one squeezes input feature maps, and the other performs excitation (Fig. 3). The channel-wise dependencies of the feature maps are calculated, followed by the scaling of the importance of each channel. In the squeezing part, the GAP is used to compress the spatial information as a representative value of each channel (C). Subsequently, channel counts reduce from C to C/r through the 1st FC layer. The parameter r denotes the squeeze ratio of the feature maps, and its value is set to 16. In the excitation part, the second FC layer is used to expand the channel counts to C, and the scale value (between 0 and 1) is calculated using the sigmoid function. Finally, feature recalibration is performed by multiplying the scale value of each channel of the input. The SE block extracts important breast tumor features from each network and uses multi-scale information. In the proposed BTEC-Net, the SE block is applied to the last layer of each block of both MSSE-DenseNet121 and MSSE-ResNet101.

3.2.3. Feature extraction using MSSE-ResNet101

The MSSE-ResNet101 in Fig. 4 comprises the first feature extractor incorporated into the proposed BTEC-Net. In the original ResNet101, a residual connection method is applied to maintain low-level features that can be lost when an input image passes through its layers. Because the tumor edge or shape needs to be efficiently obtained, ResNet101 is considered the backbone of the proposed network. The ResNet101 model is pretrained using the ImageNet dataset, and the four SE block and GAP layers applied to the last layers of the four residual blocks.

The SE block assigns a weight to each channel in the input feature maps, and a feature vector of size 1 × 1 × C is generated through the GAP layer. The multi-scale feature maps from Residual

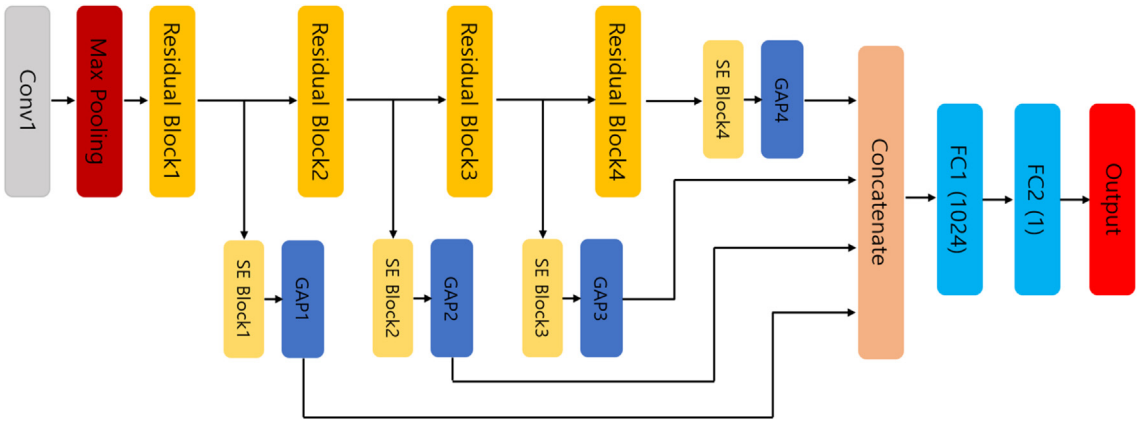

Fig. 4. Proposed MSSE-ResNet101 architecture.

Block 1–4 pass through the corresponding SE blocks, GAP layers, and output feature vectors. The number of channels contained in these feature vectors equals 256, 512, 1024, and 2048. The feature vectors extracted from the residual blocks are connected in the channel direction. These connected feature vectors finally pass through the FC layer before being reduced to 1,024 dimensions. The training of the proposed MSSE-ResNet101 as a single model involved the inclusion of a one-dimensional FC and output (sigmoid) layers, as shown in Fig. 4.

3.2.4. Feature extraction using MSSE-DenseNet121

The MSSE-DenseNet121 is the second feature extractor used in the proposed method. Unlike other conventional classification networks, the original DenseNet121 model can preserve low-level features. This is because it connects all the previous feature maps within the dense block. Accordingly, the DenseNet121 is considered the second backbone model of the proposed network for facilitating the efficient extraction of breast tumor features.

The structure of this network comprises an input layer (Conv1), four dense blocks, and transition layers. The filter size of Conv1 is 7×7, and the feature map size is reduced through a 3×3-max pooling layer. Dense Blocks 1–4 comprise 6, 12, 24, and 16 bottleneck layers, respectively. An average pooling layer is included in the transition layer, thereby reducing the input feature map size and channel count by half. Three transition layers connect each block between the dense blocks. The MSSE-DenseNet121 shown in Fig. 5 comprises SE blocks and GAP layers connected to the last layers of each of the four dense blocks to extract feature vectors of sizes $1 \times 1 \times C$. The multi-scale feature maps obtained from Dense Blocks 1–4 pass through the SE blocks and GAP layers, and feature vectors comprising 256, 512, 1024, and 1024 channels, respectively, are obtained as outputs. The four feature vectors extracted from each dense block are connected in the channel direction and finally reduced to 1024 dimensions through an FC layer of equivalent dimensions. The proposed MSSE-DenseNet121 is pretrained using each breast ultrasound dataset by adding a one-dimensional FC layer and an output layer (sigmoid), as depicted in Fig. 5.

3.2.5. Feature-Level fusion module

Because it is difficult to classify all data using only a single network, we used an FFM to aggregate the extracted feature vectors acquired from the sub-networks and performed the final classification. Notably, previous studies on medical image classification have also frequently used an ensemble method that combines several networks. However, the proposed BTEC-Net does not simply combine the results obtained from each network. Instead, it applies a learning-based feature-level fusion technique using FC layers. Traditional ensemble methods either directly set the weight parameters or use only the decision values (0 and 1) of each model to obtain the final output. In our FFM, we extracted and used the feature vectors and not the decision value of each subnetwork. In addition, by learning the nonlinear FC layer, the optimal weights for the feature vectors of the two subnetworks could be determined.

In the FFM, the two 1024-dimensional feature vectors extracted from the MSSE-DenseNet121 and MSSE-ResNet101 were first combined in the channel direction. When performing feature-level

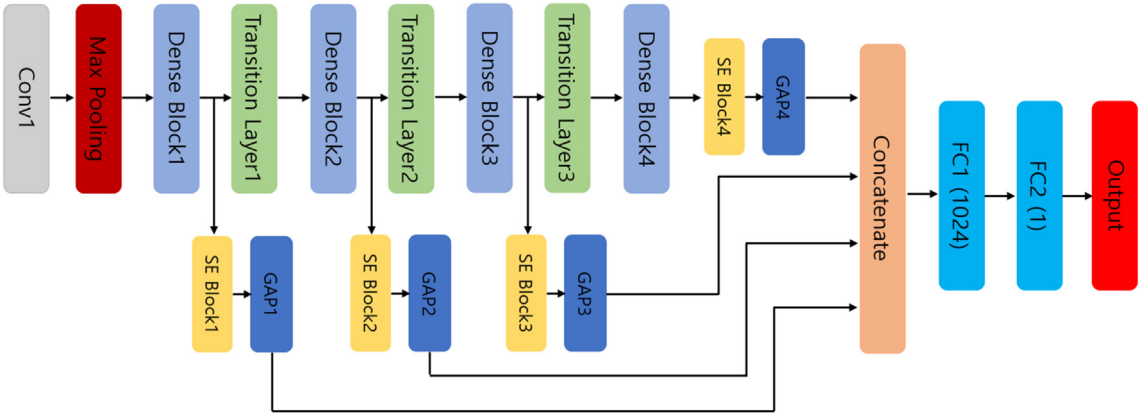

Fig. 5. Proposed MSSE-DenseNet121 architecture.

fusion, the feature vector dimensions were matched identically by applying the 1024-dimensional FC layer to each feature extractor in advance to balance the two models. Finally, the combined 2048-dimensional feature vectors passed through the one-dimensional FC and sigmoid layers, thereby classifying the input images as normal or abnormal. Table 1 presents the FFM architecture and the constituent layers.

3.3. Stage 2: Breast tumor image segmentation

3.3.1. Residual feature selection UNet (RFS-UNet)

Stage 2 of our method performs exclusive segmentation of the breast tumor images classified as abnormal. Fig. 6 shows the RFS-UNet architecture (Table 2), which can be divided into the encoder and decoder parts. These parts are connected via skip connections at layer positions where the feature maps have the same size. The encoder part comprises five encoder blocks connected in series, which acquire the high-level features. Each encoder block comprises a convolutional layer and residual convolutional block that double the channel count of the feature maps and maintain low-level features, respectively. Subsequently, the encoder blocks are connected to the RFS module, and the weight is learned for each channel from the feature maps corresponding to the previous layer and current output using a weighted sum method. The first four encoder blocks reduce the feature map size by half through the max pooling layer. The decoder comprises four decoder blocks to increase the feature map size. Each decoder block comprises a transpose convolutional layer, which doubles the previous-layer feature map size and reduces the channel count by half. The transpose convolutional layer was connected to the encoder block via two feature maps of the same sizes. Thereafter, the structure of the decoder blocks resembled that of its encoder counterpart, that is, a convolutional layer for reducing the channel count, a residual convolutional block, and an RFS module. The structure of the skip connection is identical to that observed in the original attention UNet. Furthermore, a spatial attention (SA) module was adopted to rescale and deliver the encoder block feature maps to the decoder block. The SA module maintains the information of the previous layer by connecting the encoder and decoder blocks and learns by focusing on various shapes of the target.

Table 1
Description of the FFM.

Layers	Input size	Output size	Iterations
Feature Input [f1, f2]	1 × 1 × 1024 (f1) 1 × 1 × 1024 (f2)	1 × 1 × 1024 (f1) 1 × 1 × 1024 (f2)	–
Concatenate	1 × 1 × 1024 (f1) 1 × 1 × 1024 (f2)	1 × 1 × 2048	1
FC	1 × 1 × 2048	1 × 1 × 1	1
Output Layer (sigmoid)	1 × 1 × 1	1	1

3.3.2. Residual block and RFS module

To enhance the segmentation accuracies for breast tumors, we replaced the convolutional layers used in the existing UNet (Ronneberger et al., 2015) with the residual blocks shown in Fig. 6 and used it together with the newly proposed RFS module.

The residual block comprises three convolutional layers with filter sizes of 3 × 3. Instance normalization, rectified linear units (ReLUs), and residual connection were applied to all the convolution layers. A residual connection is applied to the entire residual block, and the resulting output contains the input and output feature maps of the block. In the RFS-UNet, a residual block is used in combination with the RFS module. In previous studies, residual and

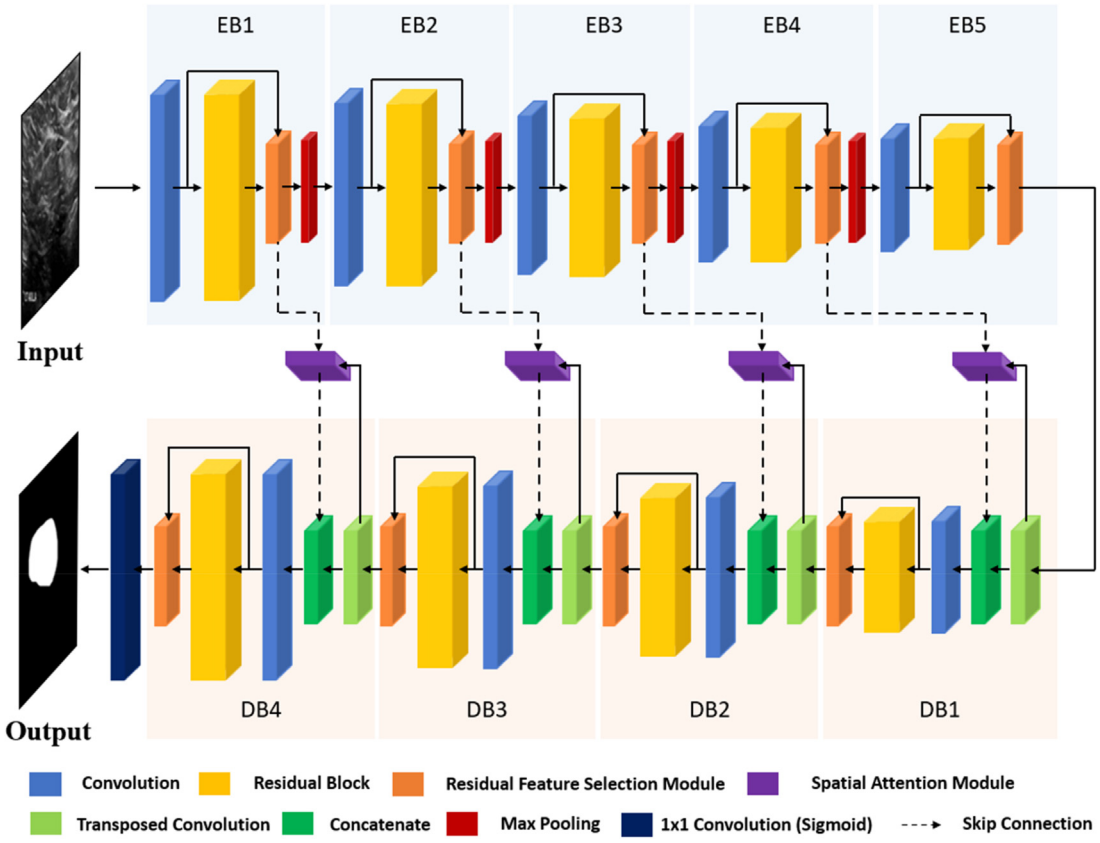

Fig. 6. Architecture of proposed RFS-UNet.

Table 2
Description of the proposed RFS-UNET.

Layer names		Input size	Output size	Kernels	Iterations
Image Input		224 × 224 × 3	224 × 224 × 3	–	–
EB1	Conv1	224 × 224 × 3	224 × 224 × 32	3 × 3	1
	Residual Block1	224 × 224 × 32	224 × 224 × 32	3 × 3	3
	RFS Module1	224 × 224 × 32 224 × 224 × 32	224 × 224 × 32		1
	MaxPool1	224 × 224 × 32	112 × 112 × 32	2 × 2	1
SA Module1		224 × 224 × 32	224 × 224 × 32	3 × 3 3 × 3 1 × 1	1
EB2	Conv2	112 × 112 × 32	112 × 112 × 64	3 × 3	1
	Residual Block2	112 × 112 × 64	112 × 112 × 64	3 × 3	3
	RFS Module2	112 × 112 × 64	112 × 112 × 64	–	1
	MaxPool2	112 × 112 × 64	56 × 56 × 64	2 × 2	1
SA Module2		112 × 112 × 64	112 × 112 × 64	3 × 3 3 × 3 1 × 1	1
EB3	Conv3	56 × 56 × 64	56 × 56 × 128	3 × 3	1
	Residual Block3	56 × 56 × 128	56 × 56 × 128	3 × 3	3
	RFS Module3	56 × 56 × 128	56 × 56 × 128	–	1
	MaxPool3	56 × 56 × 128	28 × 28 × 128	2 × 2	1
SA Module3		56 × 56 × 128	56 × 56 × 128	3 × 3 3 × 3 1 × 1	1
EB4	Conv4	28 × 28 × 128	28 × 28 × 256	3 × 3	1
	Residual Block4	28 × 28 × 256	28 × 28 × 256	3 × 3	3
	RFS Module4	28 × 28 × 256	28 × 28 × 256	–	1
	MaxPool4	28 × 28 × 256	14 × 14 × 256	2 × 2	1
SA Module4		28 × 28 × 256	28 × 28 × 256	3 × 3 3 × 3 1 × 1	1
EB5	Conv5	14 × 14 × 256	14 × 14 × 512	3 × 3	1
	Residual Block5	14 × 14 × 512	14 × 14 × 512	3 × 3	3
	RFS Module5	14 × 14 × 512	14 × 14 × 512	–	1
DB1	Trans Conv1	14 × 14 × 512	28 × 28 × 256	2 × 2	1
	Concat1	28 × 28 × 256	28 × 28 × 512	–	1
	Conv6	28 × 28 × 512	28 × 28 × 256	3 × 3	1
	Residual Block6	28 × 28 × 256	28 × 28 × 256	3 × 3	3
	RFS Module6	28 × 28 × 256	28 × 28 × 256	–	1
DB2	Trans Conv2	28 × 28 × 256	56 × 56 × 128	2 × 2	1
	Concat2	56 × 56 × 128	56 × 56 × 256	–	1
	Conv7	56 × 56 × 256	56 × 56 × 128	3 × 3	1
	Residual Block7	56 × 56 × 128	56 × 56 × 128	3 × 3	3
	RFS Module7	56 × 56 × 128	56 × 56 × 128	–	1
DB3	Trans Conv3	56 × 56 × 128	112 × 112 × 64	2 × 2	1
	Concat3	112 × 112 × 64	112 × 112 × 128	–	1
	Conv8	112 × 112 × 128	112 × 112 × 64	3 × 3	1
	Residual Block8	112 × 112 × 64	112 × 112 × 64	3 × 3	3
	RFS Module8	112 × 112 × 64	112 × 112 × 64	–	1
DB4	Trans Conv4	112 × 112 × 64	224 × 224 × 32	2 × 2	1
	Concat4	224 × 224 × 32	224 × 224 × 64	–	1
	Conv9	224 × 224 × 64	224 × 224 × 32	3 × 3	1
	Residual Block9	224 × 224 × 32	224 × 224 × 32	3 × 3	3
	RFS Module9	224 × 224 × 32	224 × 224 × 32	–	1
Output Layer (sigmoid)		224 × 224 × 32	224 × 224 × 1	1 × 1	1

attention blocks were used separately. However, in the RFS-UNet, selective attention is applied to the residual connections. A residual connection is a structure that adds residual (previous) feature maps to current output feature maps. Selective attention applies different weights for each channel to two different feature maps. It learns which feature to use by multiplying a value between 0 and 1 for each channel among the two feature maps. Thus, selectively focusing on important features for each channel is termed selective attention.

The scaling value α_i denotes the importance of the feature maps for each channel, and it is multiplied for each channel of the input feature maps (X_1) transferred through the residual connection. The second input feature map (X_2) represents the residual block output, and the value of 1-α_i is multiplied for each channel. Notably, α_i is multiplied by the two feature maps to apply weights in a weighted sum method. Applying α_i and 1- α_i as ratio values for each channel to the two feature maps amplifies the more important features and suppresses the less important features at the same channel location.

In the RFS-UNet, the RFS module is applied to the input and output feature maps of the residual block. It learns the importance of the residual connection at each layer level by itself and can selectively utilize the kernel of the input or output feature maps. The RFS module controls the contribution of feature maps added during the residual operation process. The RFS module consists of a simple layer that can easily be applied to the residual connection part. This means that the RFS module has high scalability and compatibility with those of other studies.

The inputs to the RFS module comprise feature maps before and after passing through the residual block. Here, the importance of these feature maps is calculated in terms of their dimensions, and feature maps of the same dimension are added using the

weighted sum method. This self-attention method enables the network to concentrate on the most important features of each channel and learn what should be added. Fig. 7 shows the combination of two different input features to form a single feature map, which is divided into two parallel branches to perform operations. In the first branch, the GAP layer is applied to focus on global features. In the second branch, the global max-pooling (GMP) layer is adopted to focus on local features. the outputs of the GAP and GMP layers are $1 \times 1 \times C$. These are followed by shared FC layers. The feature-map dimensions are reduced by half and increased to C in the first and second FC layers, respectively. The FC layers applied to each branch share the same weights to perform their relevant operations. The output of each branch that passes through the shared FC layer is added again, and attention weights (α_i) with values between 0 and 1 are output through the sigmoid layer. Finally, the attention weights are applied and the scaled feature maps X1 and X2 are added again, and the feature map (Y) of the original size is obtained as the output. Conventional residual connections simply serve to add input features. However, our RFS module eliminates unnecessary features from the previous layer and selectively uses good features.

3.3.3. Spatial attention module

In this study, certain parts of the convolutional layer in the conventional attention module (Oktay et al., 2018) were modified before its application to the proposed RFS-UNet. The SA module shown in Fig. 8 was adopted for the skip connection for performance enhancement. It used the encoder and decoder feature maps as inputs and applied weights by channel to the feature maps transferred from the encoder. Furthermore, the SA module, which refers to the decoder's feature maps, learns the most prominent features of the encoder and multiplies them with higher attention values. Thus, meaningful and diverse features are delivered to the end as the network deepens. Through this self-attention method, the network learns important features by itself.

Fig. 8 represents the proposed SA module. The inputs of the SA module are encoder and decoder block-feature maps of the same sizes. These maps are summated after passing them through the convolutional layer. In a previous attention module (Oktay et al., 2018), 1×1 convolution was adopted for the two input feature maps. However, in the proposed SA module, the filter size is modified to 3×3. Therefore, a large receptive field, which is designed to focus more on spatial information, is used in the proposed network. The added feature maps pass through the ReLU and 1×1 convolutional layer that contains a single filter and sigmoid layer, which outputs the attention weight (α_w). Finally, the encoder block feature maps are multiplied by α_w for each channel and delivered to the decoder block.

3.3.4. Loss function

The multi-stage segmentation method presented in this study performs both classification and segmentation. In Stage 1, the MSSE-DenseNet121, MSSE-ResNet101, and BTEC-Net were used to perform breast tumor image classification. At this stage, the binary cross-entropy (BCE) was applied to train each model. The BCE loss equation is expressed as follows:

$$\text{BCE loss} = -\frac{1}{NUM}\sum_{i=1}^{NUM}[TC_i \cdot \log(PC_i) + (1 - TC_i) \cdot \log(1 - PC_i)] \quad (1)$$

where NUM represents the number of images, TC_i represents the actual target class label of the i^{th} image (i.e., 0 and 1 for the normal and abnormal classes, respectively), and PC_i denotes the predicted class for the i^{th} training image and assumes a value between 0 and 1 after passing through the sigmoid layer. When TC_i is 0, the BCE loss becomes $\log(1-PC_i)$, and when TC_i is 1, the BCE loss becomes $-\log(PC_i)$. The former term $-\log(PC_i)$ is monotonically decreasing, and the latter term $-\log(1 - PC_i)$ is monotonically

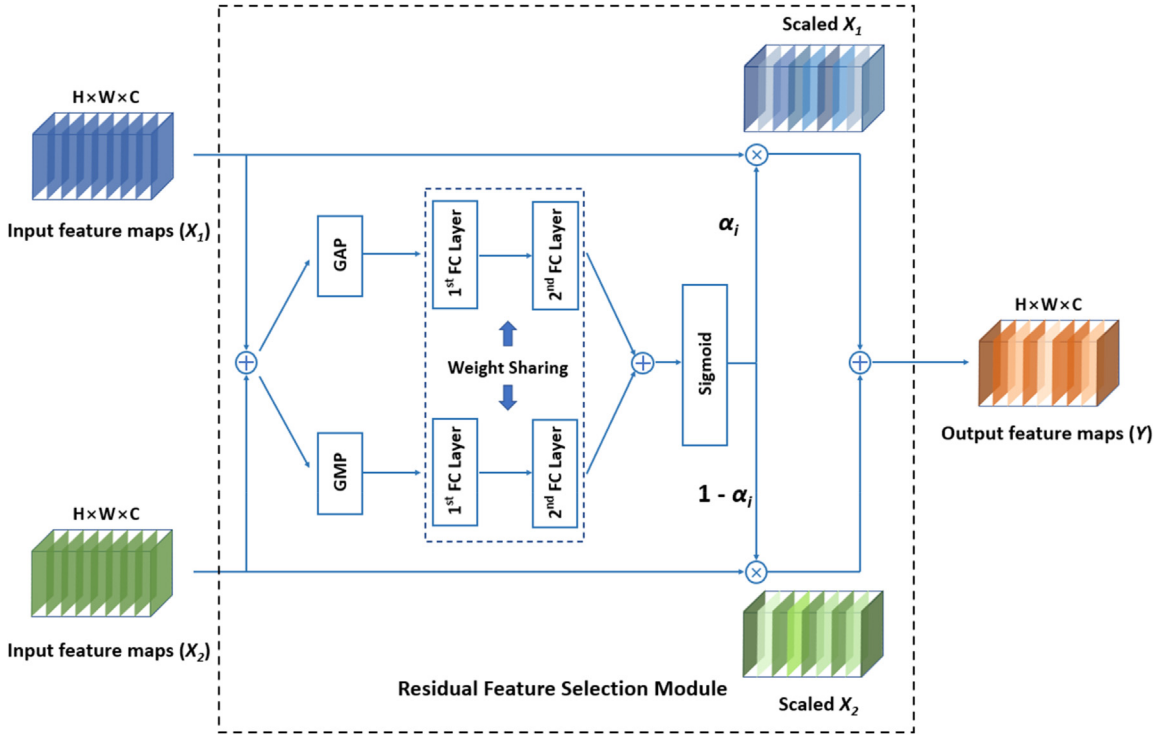

Fig. 7. Architecture of proposed RFS module used in RFS-UNet.

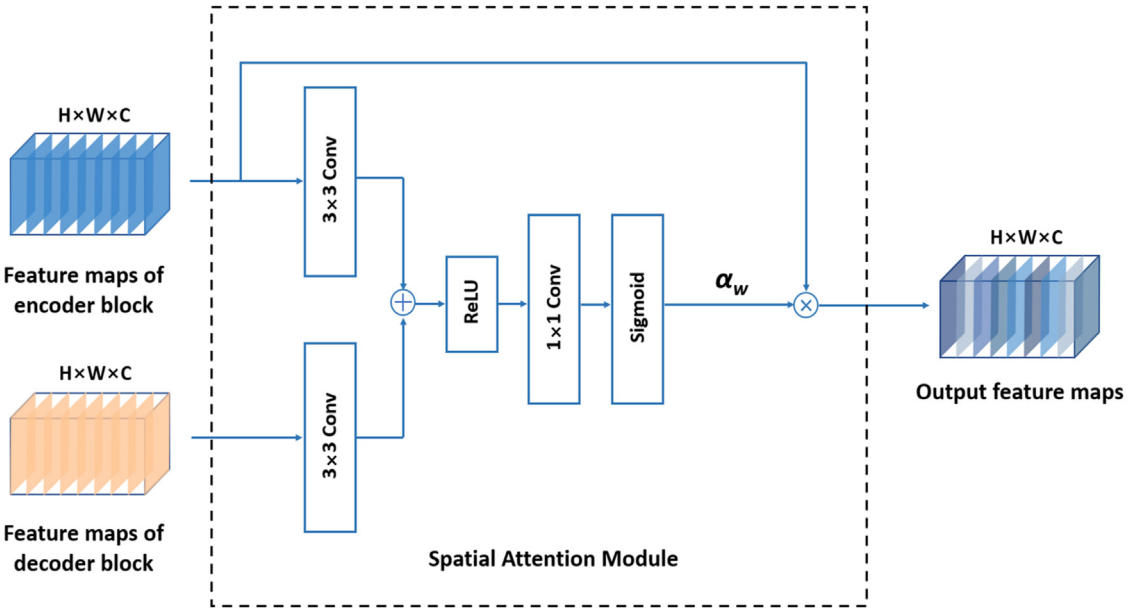

Fig. 8. Architecture of SA module used in proposed RFS-UNet. It is applied to four skip connections of RFS-UNet.

increases with respect to PC_i. Because the value of TC_i (0 or 1) is multiplied by both terms, the loss value is calculated only for the corresponding label value. Therefore, the BCE loss is not monotonic with respect to PC_i.

In the segmentation stage, the proposed RFS-UNet is used to segment images classified as abnormal, and the dice coefficient (DC) loss is adopted for training.

$$\mathrm{DC} = \frac{1}{NUM}\sum_{i=1}^{NUM}\frac{2\times(GR_i \cap PR_i)}{(GR_i + PR_i)} \tag{2}$$

$$\mathrm{DC\,loss} = 1 - \frac{1}{NUM}\sum_{i=1}^{NUM}\frac{2\times(GR_i \cap PR_i)}{(GR_i + PR_i)} \tag{3}$$

GR_i denotes the tumor region, which has a label value of 1 in the ground truth. PR_i denotes the tumor region predicted by the network during the segmentation stage. The denominator expression in the above-mentioned DC equations represents the sum of the tumor pixels contained within the network outputs and ground-truth images. The numerator denotes the intersection of the tumor pixels in the two images. In this study, the DC loss was applied as 1-DC to learn our proposed network in the direction of increasing DC and decreasing loss values.

4. Experimental results

4.1. Experimental databases and setup

We used two publicly available datasets, breast ultrasound images (BUSI) (Al-Dhabyani et al., 2020) and UDIAT (Yap et al., 2018), as illustrated in Fig. 9. The BUSI dataset comprised 780 ultrasound images acquired from females whose ages ranged from 25 to 75 years. Three categories were observed: normal (133), benign (437), and malignant (210), respectively. All the images were grayscale. In the experiments, we separated the BUSI dataset into training (80 %) and testing (20 %) subsets to demonstrate the generalized performance. The training and testing of all the networks were executed based on 5-fold cross-validation. When training the BUSI dataset, 10 % of the training set was used for validation to verify whether the model training was well performed. Because the number of images was small, part of the training set was used for validation without using a separate validation set.

The second public dataset, UDIAT, comprised 163 breast ultrasound images: 110 benign and 53 malignant cases. All the images were grayscale, and the ground-truth images were labeled by radiologists. As previous studies using the UDIAT dataset (Byra et al., 2020; Lee et al., 2020) did not use validation data, and the number of UDIAT dataset was extremely small, we also performed the experiments in the same way.

In the experiments, benign and malignant images containing breast tumors were categorized as abnormal, and images without any breast tumors were classified as normal. Because the UDIAT dataset does not have normal images, only abnormal images were used to measure the performance. Detailed descriptions of the datasets are presented in Tables 3 and 4.

4.2. Training of the proposed network

During the classification stage, the MSSE-ResNet101 and MSSE-DenseNet121 were trained using the BUSI dataset. Considering the computational cost, the sizes of the original images were changed to 224 × 224 pixels and used as the inputs. The batch size was 16, and the network was trained for 300 epochs. We applied adaptive moment estimation (Adam) optimizer and set 0.001 as the learning rate. The networks were trained by reducing the learning rate to half at 50-epoch intervals. The BTEC-Net uses the pretrained MSSE-ResNet101 and MSSE-DenseNet121 networks as feature extractors and applies the FFM to combine the two models. During the BTEC-Net training, we used batch sizes and epochs of 12 and 20, respectively. Furthermore, the learning rate was 0.0001, and end-to-end learning was performed by reducing it to half at 5-epoch intervals. The BCE loss function was applied during the classification stage.

The RFS-UNet was trained from the beginning to segment the tumor pixels within the abnormal images. Its inputs comprised images resized to 224 × 224 pixels, and it was trained considering a batch size and epoch count of 12 and 100, respectively. The Adam optimizer was used. In addition, the learning rate was 0.001, and

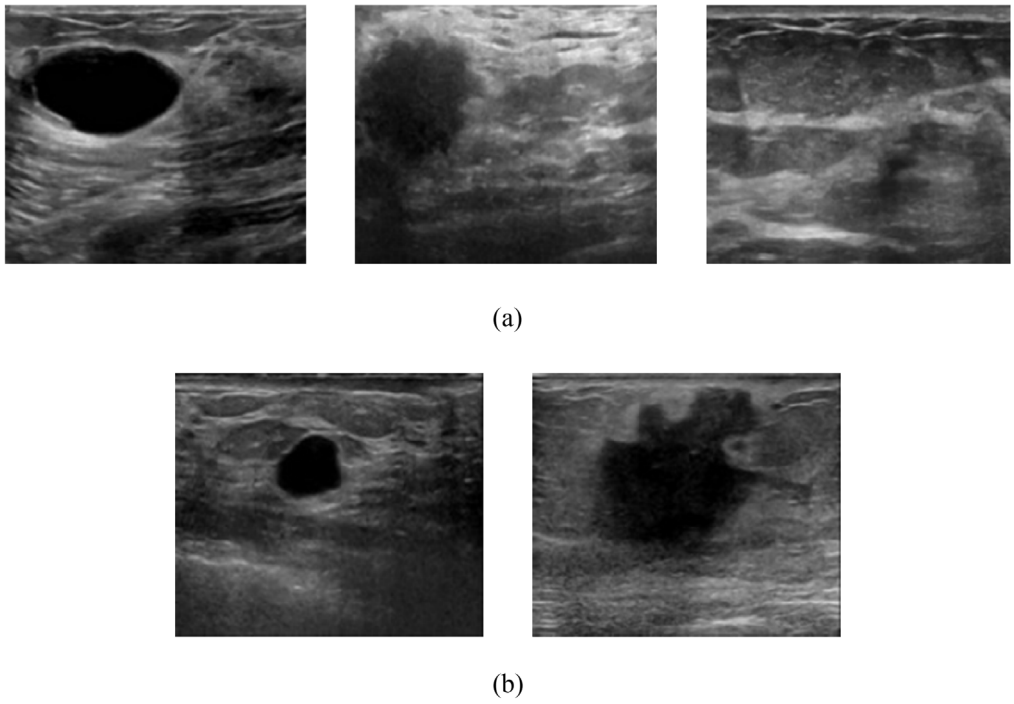

Fig. 9. Sample images from (a) BUSI and (b) UDIAT datasets. The images from left to right depict benign tumor, malignant tumor, and normal cases, respectively. In case of UDIAT dataset, normal case images are not included.

Table 3
Data distribution of the BUSI dataset. (unit: image).

Fold	Training (Validation)			Testing		
	Abnormal class		Normal class	Abnormal Class		Normal class
	Benign	Malignant	Normal	Benign	Malignant	Normal
1st fold	350 (35)	168 (17)	106 (11)	87	42	27
2nd fold	350 (35)	168 (17)	106 (11)	87	42	27
3rd fold	350 (35)	168 (17)	106 (11)	87	42	27
4th fold	349 (35)	168 (17)	107 (11)	88	42	26
5th fold	349 (35)	168 (17)	107 (11)	88	42	26

Table 4
Data distribution of the UDIAT dataset. (unit: image).

Fold	Training			Testing		
	Abnormal class		Normal class	Abnormal Class		Normal class
	Benign	Malignant	Normal	Benign	Malignant	Normal
1st fold	88	42	–	22	11	–
2nd fold	88	42	–	22	11	–
3rd fold	88	42	–	22	11	–
4th fold	88	43	–	22	10	–
5th fold	88	43	–	22	10	–

the network was trained by reducing it by half at 30-epoch intervals. The DC loss function was applied for training. All our experiments were performed with 5-fold cross-validation.

Fig. 10 depicts that training losses converge reliably and that the proposed network is optimized. As shown in Fig. 10(c) and (d), the validation losses converge to sufficiently low values, and the validation accuracies converge to high values according to the increase in epochs. This result confirmed that our model did not overfit the training data.

All training and testing were executed on a desktop (Intel ® Core™ i7-7700 CPU) installed with a Windows 10 and RTX 3080 (NVIDIA GeForce RTX 3080, 2021) with 10 GB of graphics memory. The codes of our CNN models were executed within the TensorFlow 2.5.0 framework (nightly version) (TensorFlow, 2021).

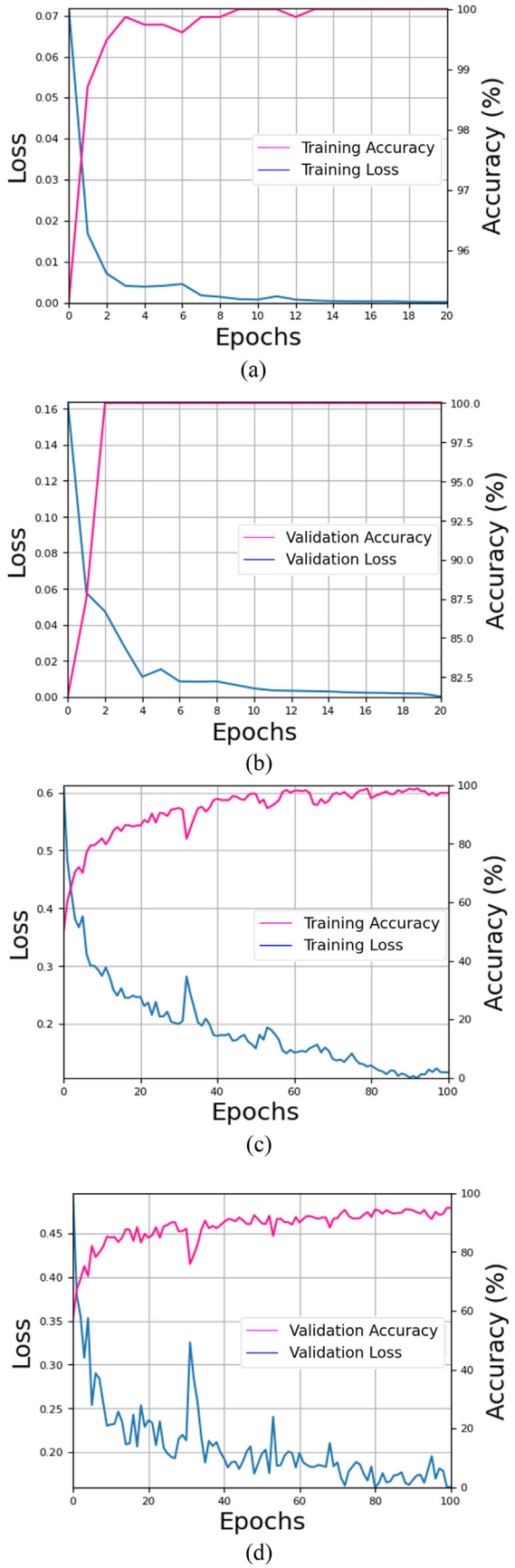

Fig. 10. Curves of loss and accuracy of proposed method. Training loss and accuracy of (a) BTEC-Net and (c) RFS-UNet. Validation loss and accuracy of (b) BTEC-Net and (d) RFS-UNet.

4.3. Testing

4.3.1. Performance evaluation metric

In this study, the classification and segmentation operations were performed on images of the BUSI dataset. The F1 score, accuracy, precision, and recall were adopted for the evaluation of our classification network in Stage 1. These evaluation metrics could be calculated using the numbers of true positives (TPo), false positives (FPo), true negatives (TNe), and false negatives (FNe). These metrics are expressed in Equations (4)–(7).

$$\text{Accuracy} = \frac{TPo + TNe}{TPo + FPo + TNe + FNe} \tag{4}$$

$$\text{Recall} = \frac{TPo}{TPo + FNe} \tag{5}$$

$$\text{Precision} = \frac{TPo}{TPo + FPo} \tag{6}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

In Stage 2, we used pixel accuracy (Pixel Acc), intersection over union (IoU), and DC to compare the segmentation performances of our RFS-UNet and conventional methods. These evaluation metrics are also commonly adopted in other conventional breast ultrasound image-segmentation applications. Notably, the number of image pixels was used as the base unit in these evaluations. Accordingly, Pixels Acc, IoU, and DC can be expressed as follows:

$$\text{Pixel Acc} = \frac{TPo + TNe}{TPo + FPo + TNe + FNe} \tag{8}$$

$$\text{IoU} = \frac{TPo}{TPo + FPo + FNe} \tag{9}$$

$$\text{DC} = \frac{2 \times TPo}{(TPo + FPo) + (TPo + FNe)} \tag{10}$$

In these equations, TPo, TNe, FPo, and FNe denote the same parameters as described in Equations (4)–(7), although they are subsequently determined in terms of the pixel count. Accordingly, Pixel Acc denotes the percentage of pixels correctly classified using the proposed network. IoU represents the intersection ratio compared with the union of outputs and ground-truth images for the tumor region. The DC offers another means for evaluating the similarity between network output and target class label. It assumes a value between 0 and 1.

Table 5
Segmentation performances of the existing and proposed methods on the BUSI without normal data. (unit: %).

Methods	Pixel Acc	IoU	DC
UNet (Ronneberger et al., 2015)	96.285	67.748	76.499
Attention UNet (Oktay et al., 2018)	96.455	69.993	78.772
FCN (Long et al., 2015)	96.046	65.445	72.456
SegNet (Badrinarayanan et al., 2017)	96.357	66.476	73.571
PSPNet (Zhao et al., 2017)	96.424	69.821	79.515
DeepLabv3+ (Chen et al., 2018)	96.672	70.345	80.124
Proposed RFS-UNet	96.975	73.904	82.005

Table 6
Segmentation performances of the existing and proposed methods on the BUSI with normal data. (unit: %).

Methods	Pixel Acc	IoU	DC
UNet (Ronneberger et al., 2015)	96.174	61.488	69.052
Attention UNet (Oktay et al., 2018)	96.201	64.419	71.707
FCN (Long et al., 2015)	95.848	59.331	68.121
SegNet (Badrinarayanan et al., 2017)	96.082	60.642	68.844
PSPNet (Zhao et al., 2017)	96.223	65.227	73.115
DeepLabv3+ (Chen et al., 2018)	96.257	64.879	72.254
Proposed RFS-UNet	96.771	70.465	77.044

4.3.2. Performances degradation of segmentation for breast tumor

When single segmentation networks are applied to normal images that do not contain any tumor cells, the number of FP errors can increase. First, an experiment was conducted to confirm the performance degradation of the RFS-UNet and the existing networks.

Furthermore, we conducted comparative experiments using the BUSI datasets, with and without normal images. Table 5 compares the performances of the single-segmentation networks. This comparison was performed using only tumor images from the BUSI dataset. In other words, the datasets used for this comparison did not include any normal images. When the test was performed without a normal image, relatively few FP errors, which were primarily distributed in the area around the breast tumor pixels, were generated.

Table 6 compares the segmentation performance on the BUSI dataset, including normal images. As shown in Table 6, the IoU and DC performances decrease significantly in all the models. This is because the segmentation network can falsely detect breast tumors that are not present in a normal image when performing the test. The number of FP errors also increases, and the numerical performance decreases significantly.

Fig. 11 shows the segmentation results of normal images obtained by testing different networks applied to the BUSI dataset. Herein, although the networks are trained with a dataset that comprises both normal and tumor-containing images, the number of FP results remains high in the case of normal images. These experimental results reveal that the performance of single segmentation networks decreases invariably when the test dataset contains normal images.

4.3.3. Ablation study of breast tumor image classification

The results discussed in 4.3.2 section prove that the performance of single-segmentation models deteriorates significantly when normal images are included in the test dataset. To address this issue, this study presents a method for classifying tumor-containing and normal images using the BTEC-Net during the first stage of the proposed network. We evaluated the performances of the standalone DenseNet121 and ResNet101 networks as well as the modified components of the proposed BTEC-Net using the BUSI dataset (Table 7). The application of the SE block and multi-scale connection to the original DenseNet121 and ResNet101 models results in a significant improvement in all the evaluation metrics. Furthermore, the proposed BTEC-Net, which combines the MSSE-DenseNet121 and MSSE-ResNet101 in an ensemble method, demonstrates the best performance, with an accuracy of 99.487 %.

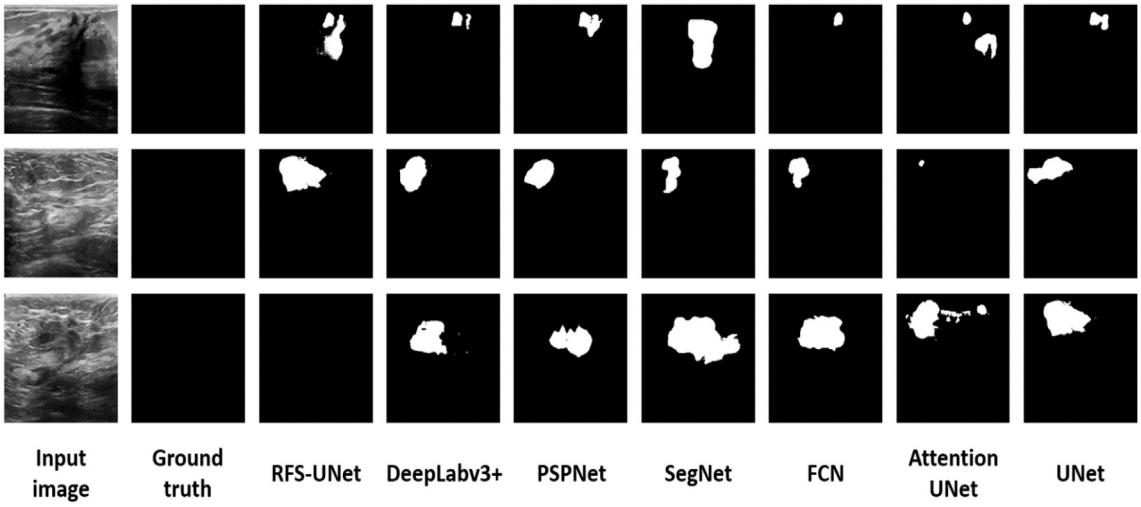

Fig. 11. Examples of false-positive results acquired by the proposed RFS-UNet and conventional single-segmentation networks. The first and second columns depict the normal-case input and ground-truth images.

Table 7
Ablation of the BTEC-Net on the BUSI dataset (unit: %) (Acc, Rec, Pre, and F1 mean Accuracy, Recall, Precision, and F1-score, respectively).

Models	Acc	Rec	Pre	F1
DenseNet121 (Huang et al., 2017)	96.282	98.301	97.248	97.771
ResNet101 (He et al., 2016)	96.538	97.682	98.137	97.909
MSSE-ResNet101 (ResNet101 (He et al., 2016) + SE block + multi-scale connection)	97.564	98.764	98.308	98.535
MSSE-DenseNet121 (DenseNet121 (Huang et al., 2017) + SE block + multi-scale connection)	98.205	98.609	99.222	98.915
BTEC-Net (proposed)	99.487	99.846	99.538	99.691

4.3.4. Comparisons of proposed BTEC-Net and State-of-the-Art models

Table 8 presents the performance of the BTEC-Net against previous classification networks when applied to the BUSI dataset. This comparison includes existing models, such as the MobileNetV2, Inception-v3, ResNet101, VGG Net-16, DenseNet121, and Xception. All the models were pretrained using the ImageNet dataset, and cross-validation was performed. The MSSE-ResNet101 and MSSE-DenseNet121 models outperformed the existing models, and the proposed BTEC-Net demonstrated the best accuracy across the four evaluation metrics. This implies that the proposed network could accurately classify nearly all the images contained within the BUSI test dataset.

For statistical analysis, we performed a *t*-test (Student's T-Test) on the proposed method and Xception (second-best), with the highest performance among the existing models (Fig. 12). The p-values of accuracy, recall, precision, and f1-score are 1.19×10^{-4},

Table 8
Ablation study of the proposed BTEC-Net using BUSI dataset (unit: %) (Acc, Rec, Pre, and F1 mean Accuracy, Recall, Precision, and F1-score, respectively).

Models	Acc	Rec	Pre	F1
MobileNetV2 (Sandler et al., 2018)	94.359	97.527	95.751	96.631
DenseNet121 (Huang et al., 2017)	96.282	98.301	97.248	97.771
Inception-v3 (Szegedy et al., 2016)	96.538	97.836	97.988	97.912
ResNet101 (He et al., 2016)	96.538	97.682	98.137	97.909
VGG Net-16 (Simonyan and Zisserman, 2015)	96.923	97.836	98.445	98.141
Xception (Chollet, 2017)	97.051	98.764	97.706	98.232
MSSE-ResNet101	97.564	98.764	98.308	98.535
MSSE-DenseNet121	98.205	98.609	99.222	98.915
BTEC-Net (Proposed)	**99.487**	**99.846**	**99.538**	**99.692**

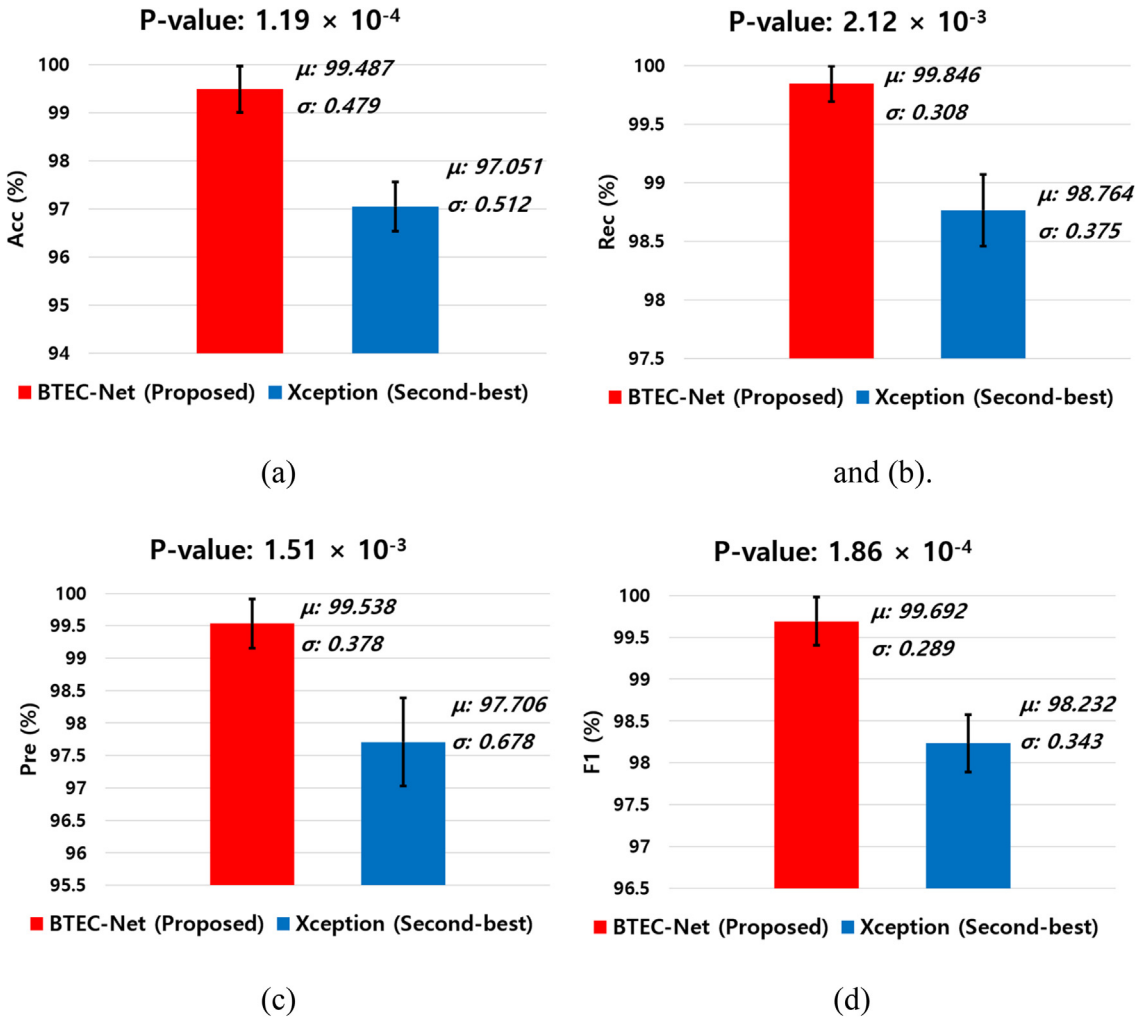

Fig. 12. Results of *t*-test for BTEC-Net (Proposed) and Xception (Second-best). (Acc, Rec, Pre, and F1 mean Accuracy, Recall, Precision, and F1-score).

Table 9
Ablation of the RFS-UNet on the BUSI dataset. (unit: %).

Methods	Pixel Acc	IoU	DC
UNet (Ronneberger et al., 2015)	96.174	61.488	69.052
UNet + AM (attention UNet (Oktay et al., 2018))	96.201	64.419	71.707
UNet (Ronneberger et al., 2015) + SAM	96.459	65.567	72.419
UNet (Ronneberger et al., 2015) + SAM + RB	96.564	67.586	74.095
UNet (Ronneberger et al., 2015) + SAM + RB + RFSM	96.652	69.273	76.102
UNet (Ronneberger et al., 2015) + SAM + RB + RFSM + CR (RFS-UNet)	96.771	70.465	77.044

Table 10
Ablation of the RFS-UNet with and without using BTEC-Net on the BUSI dataset. (unit: %).

Methods	Pixel Acc	IoU	DC
RFS-UNet	96.771	70.465	77.044
BTEC-Net + RFS-UNet (Proposed method)	97.253	77.835	84.856

2.12×10^{-3}, 1.51×10^{-3} and 1.86×10^{-4}. This means that the performance of BTEC-Net is statistically superior to that of the second-best method.

In addition, we used Cohen's d method (Cohen, 1992) to show the difference in the standardized mean between the two methods using the effective size. In the Cohen's d method, cases, where Cohen's d values were 0.2, 0.5, 0.8, or more, were classified as small, medium, and large, respectively. Cohen's d values of accuracy, recall, precision, and f1-score were 4.91, 3.15, 3.33, and 4.6, respectively. Because all Cohen's d values were greater than 0.8, the difference between the two methods represented a large effect size.

4.3.5. Ablation study of breast tumor image segmentation

Table 9 presents the ablation study of our RFS-UNet. We considered the performance of several models, starting from the UNet and attention UNet to the final proposed RFS-UNet. The UNet (Ronneberger et al., 2015) demonstrated the lowest Pixel Acc, IoU, and DC values. The original attention UNet applied an attention module (AM) to the existing UNet structure, and its performance was slightly better than that of UNet. The model that replaces the AM with the spatial attention module (SAM) also improved the accuracy compared to the attention UNet. A comparison of the UNet + SAM model before and after the replacement of parts of its convolutional layers with the residual block (RB) reveals a performance improvement in terms of all the metrics. Compared with those when RB was applied exclusively when the RB and the residual feature selection module (RFSM) were used together, the IoU and DC values improved by approximately 1.68 % and 2 %, respectively. Finally, we considered the application of the channel reduction (CR) technique to the RFS-UNet to reduce the number of filters in all the layers. This reduced the computational cost by approximately half and also slightly improved the accuracy.

In the next experiment, we analyzed the results of Table 10 to confirm the effect of BTEC-Net in our multistage segmentation method. Table 10 represents the accuracies of the proposed algorithm before and after the addition of BTEC-Net to the RFS-UNet. When the standalone RFS-UNet was used, the values of the Pixel Acc, IoU, and DC were 96.771 %, 70.465 %, and 77.044 %, respectively. The final model, which combined the RFS-UNet and BTEC-Net, significantly reduced the occurrence of these FP errors. As described in Table 10, the values of the Pixel Acc, IoU, and DC cor-

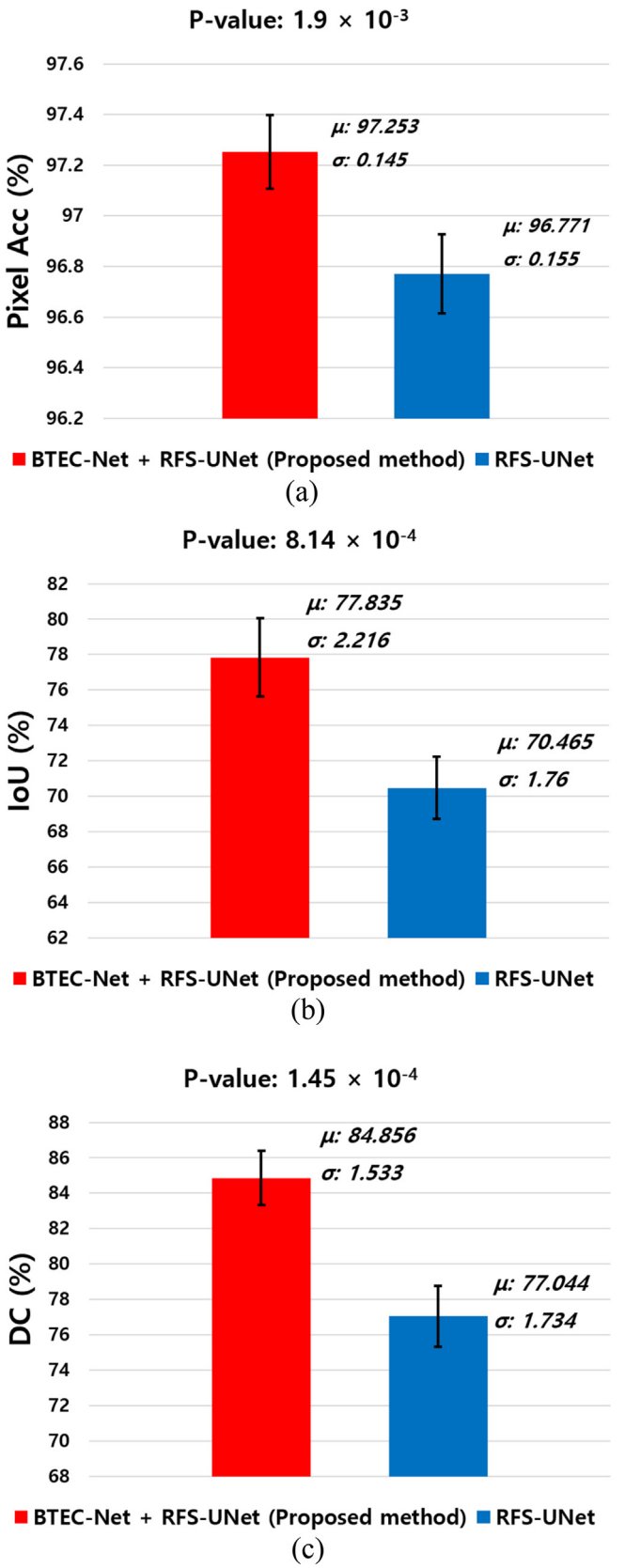

Fig. 13. Results of *t*-test for BTEC-Net + RFS-UNet (Proposed method) and RFS-UNet.

Table 11
Segmentation performances of the existing and proposed methods on the BUSI dataset. (unit: %).

Methods	Pixel Acc	IoU	DC
UNet (Ronneberger et al., 2015)	96.174	61.488	69.052
Attention UNet (Oktay et al., 2018)	96.201	64.419	71.707
FCN (Long et al., 2015)	95.848	59.331	68.121
SegNet (Badrinarayanan et al., 2017)	96.082	60.642	68.844
PSPNet (Zhao et al., 2017)	96.223	65.227	73.115
DeepLabv3+ (Chen et al., 2018)	96.257	64.879	72.254
Xue et al. (Xue et al., 2021)	96.6	56.5	64.1
Byra et al. (Byra et al., 2020)	95.6	–	70.9
Zhang et al. (Zhang et al., 2021)	95.56	–	81.42
Zhu et al. (Zhu et al., 2020)	–	76.39	84.7
Proposed method	**97.253**	**77.835**	**84.856**

responding to the final two-stage segmentation network are 97.253 %, 77.835 %, and 84.556 %, respectively, which are indicative of superior performance.

Next, we performed a *t*-test on the proposed method and RFS-UNet. The p-values of the Pixel Acc, IoU, and DC are 1.9×10^{-3}, 8.14×10^{-4}, and 1.45×10^{-4}, as shown in Fig. 13, respectively. This means that the performance difference between the two models is large.

Next, we used Cohen's d method to demonstrate the difference in the standardized means between the two methods using the effective size. Cohen's d values of pixel Acc, IoU, and DC were 3.2, 3.68, and 4.77, respectively. Because These values were greater than 0.8, the difference between the two methods represented a large effect size. These results demonstrated that our method could significantly enhance the performance of a single network.

4.3.6. Comparison of proposed and State-of-the-Art methods

In this experiment, we compared the previous and proposed segmentation networks. Table 11 lists the values of Pixel Acc, IoU, and DC for all the methods considered in this experiment. Moreover, the performances of these methods were compared with those proposed in previous studies. Compared with conventional algorithms, our proposed method demonstrated the highest

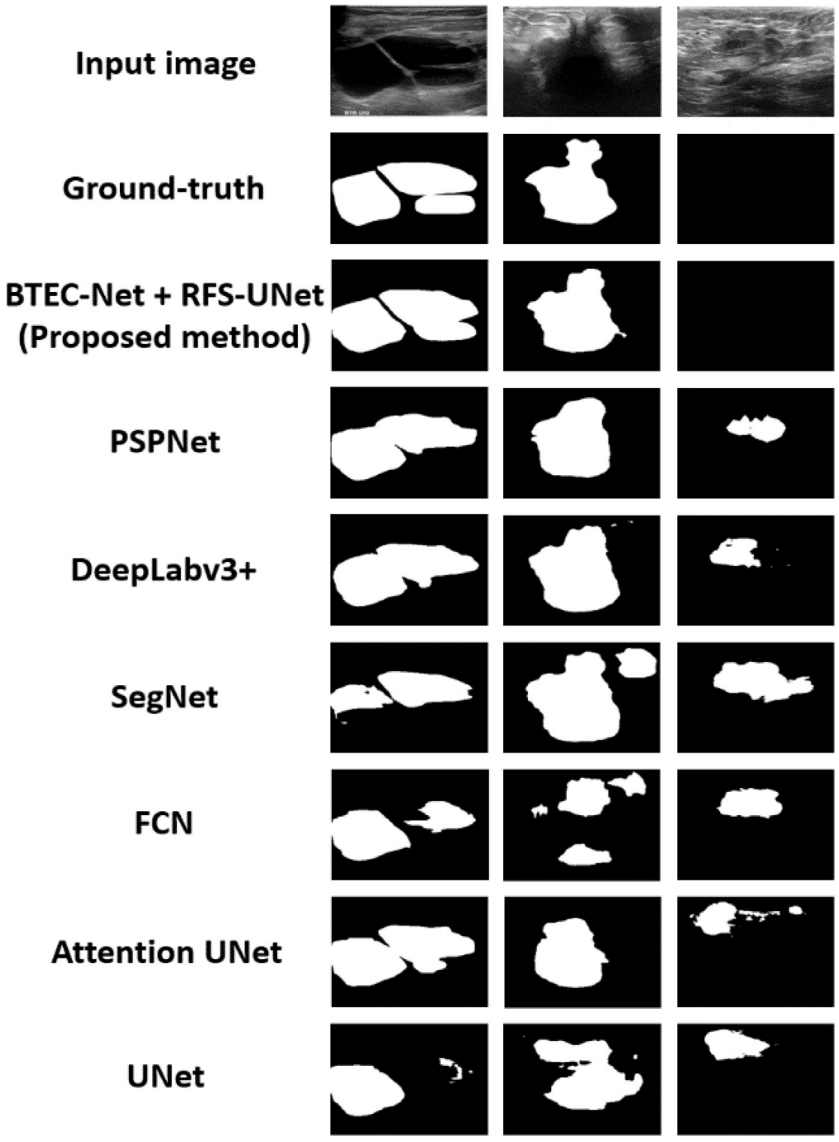

Fig. 14. Segmentation results on BUSI dataset by the previous and proposed methods. The first and second columns depict benign and malignant cases, respectively, and the last column depicts a normal case.

Table 12
Segmentation performances of the existing and the proposed methods on UDIAT. (unit: %).

Methods	Pixel Acc	IoU	DC
UNet (Ronneberger et al., 2015)	98.288	70.129	79.155
Attention UNet (Oktay et al., 2018)	98.293	71.533	80.581
FCN (Long et al., 2015)	97.676	66.667	75.924
SegNet (Badrinarayanan et al., 2017)	97.787	67.655	76.719
PSPNet (Zhao et al., 2017)	98.295	72.461	81.217
DeepLabv3+ (Chen et al., 2018)	98.342	73.122	81.968
Lee et al. (Lee et al., 2020)	97.794	62.26	76.58
Shareef et al. (Shareef et al., 2020)	–	69.5	78.2
Byra et al. (Byra et al., 2020)	98.5	–	79.1
Abraham and Khan (Abraham and Khan, 2019)	–	–	80.4
Proposed RFS-UNet	**98.601**	**77.094**	**85.366**

performance across all the metrics. Fig. 14 depicts examples of the resulting images obtained using the segmentation networks. Notably, the proposed two-stage model yields result wherein the tumor regions are accurately detected with significant resemblance to the ground-truth image. Furthermore, no FP errors occurred when the proposed model was adopted with the BUSI dataset that contained normal images. The results presented in Fig. 14 prove that segmentation result images of our method are visually the best.

Finally, we used the existing methods and the proposed RFS-UNet to compare their performances when applied to the UDIAT. The performances were compared without applying the BTEC-Net because the UDIAT dataset did not contain normal images. The results of Table 12 confirmed that the RFS-UNet is superior to the existing methods. Compared with the other networks in Fig. 15, the output images of our proposed network are characterized by noticeable tumor region boundaries.

Fig. 16 is the segmentation results images for the BUSI and UDIAT datasets. Segmentation quality of malignant tumors is

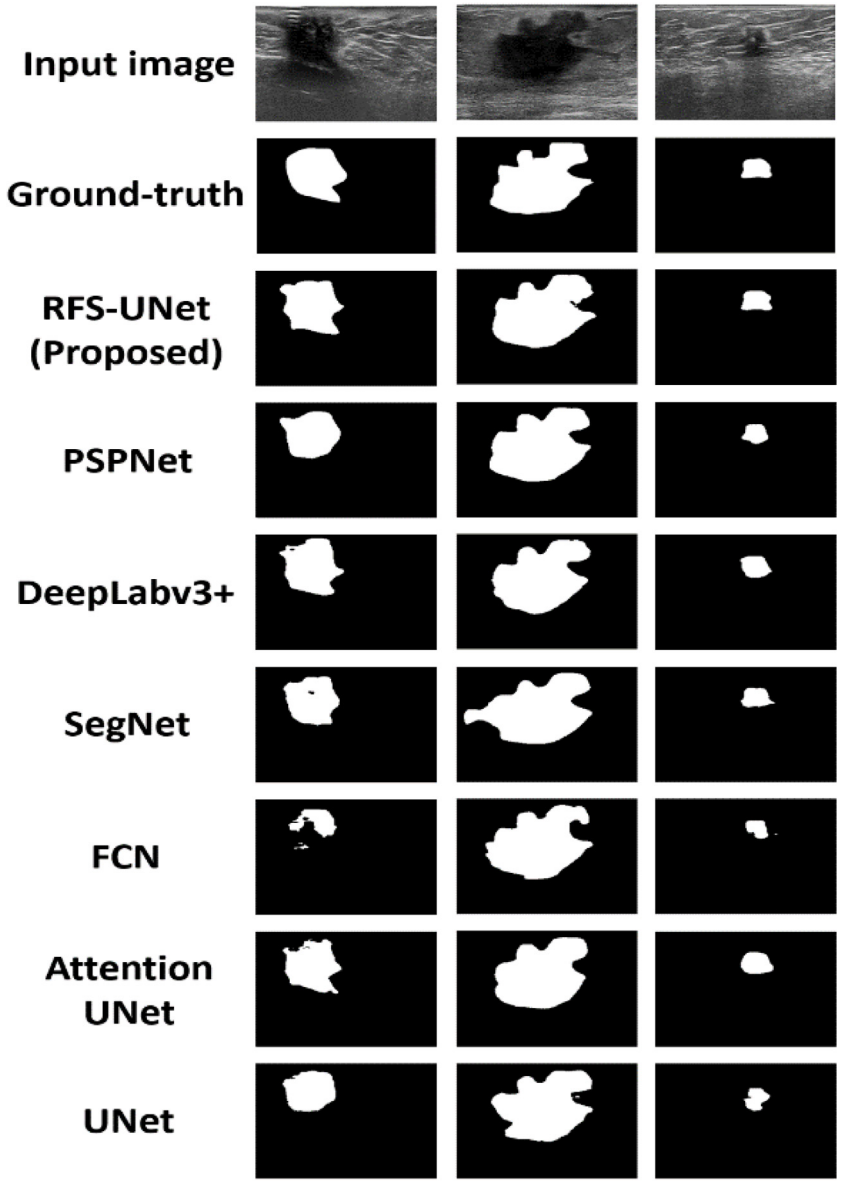

Fig. 15. Segmentation results on UDIAT dataset using proposed and existing methods.

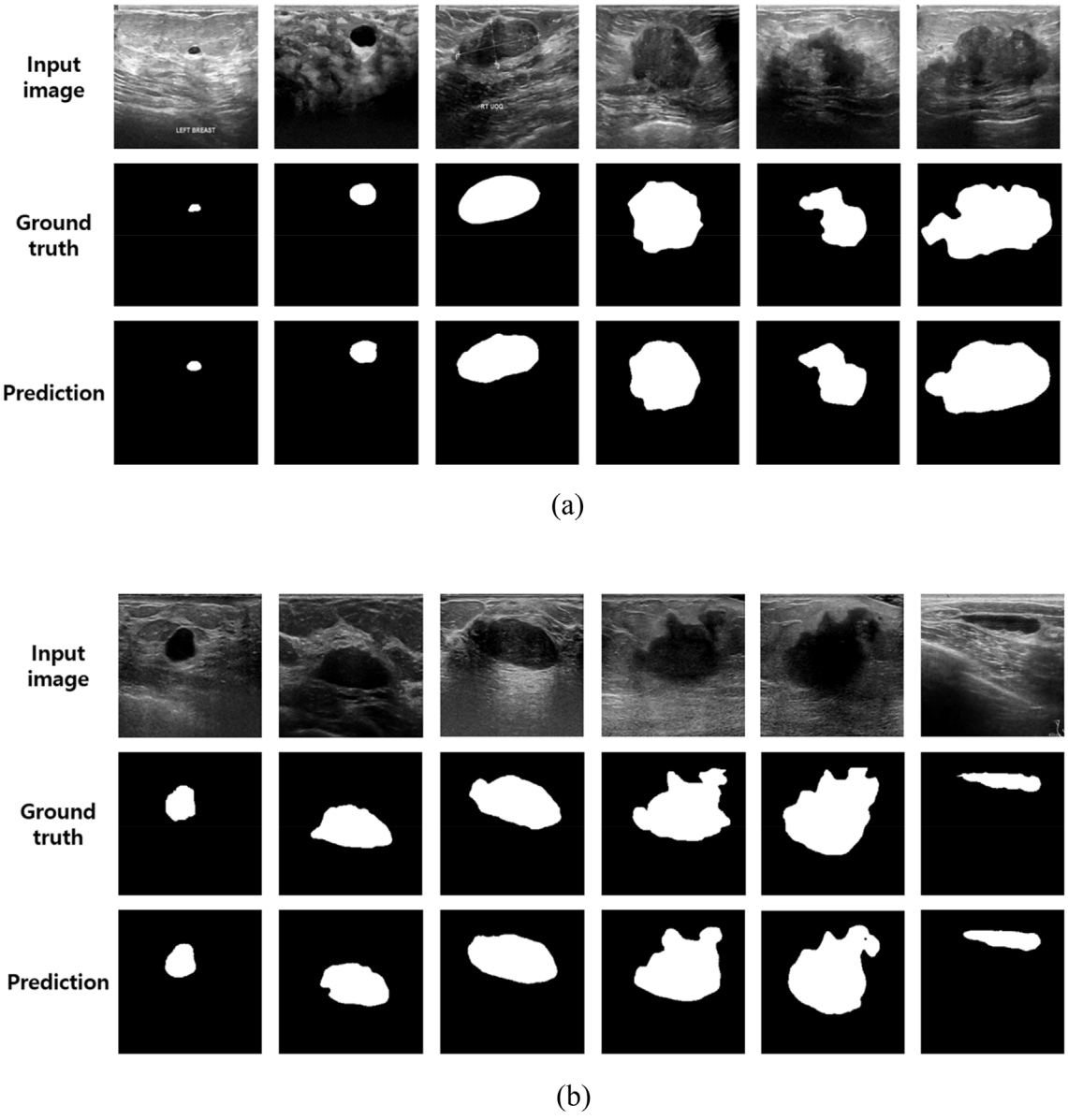

Fig. 16. Visualization of segmentation results using proposed RFS-UNet. (a) BUSI dataset and (b) UDIAT dataset. In both (a) and (b), the first to third columns are benign cases, and the fourth to last columns are malignant cases.

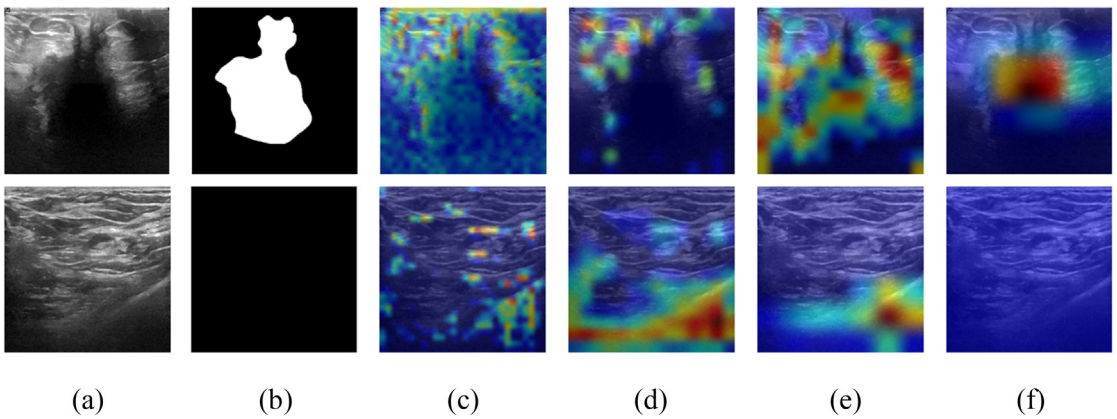

Fig. 17. Grad-CAM outcomes extracted from middle layers of MSSE-ResNet101. The first and second rows depict abnormal and normal images: (a) input images, (b) ground-truth image, and (c)–(f) Grad-CAM outcomes extracted from the last convolutional layer of Residual Blocks 1–4.

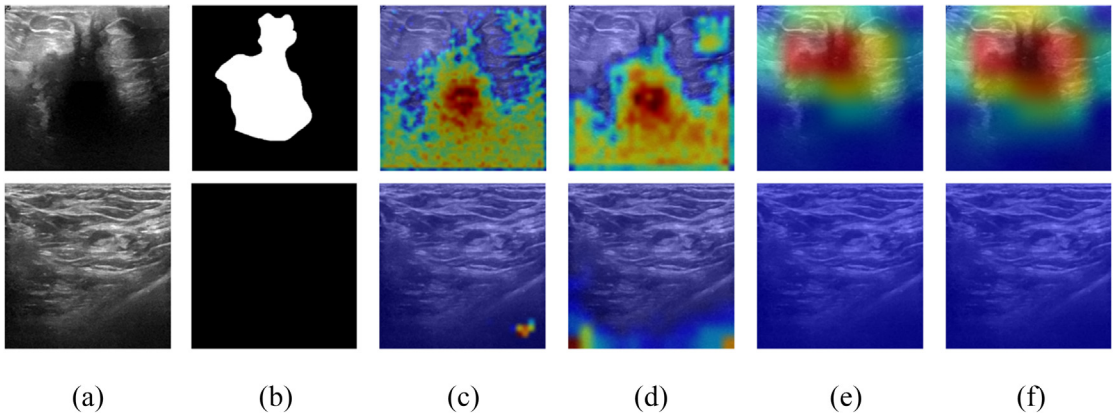

Fig. 18. Grad-CAM outcomes extracted from middle layers of MSSE-DenseNet121. The first and second rows depict abnormal and normal images: (a) input images, (b) ground-truth images, and (c)–(f) Grad-CAM outcomes extracted from the last convolutional layer of Dense Blocks 1–4.

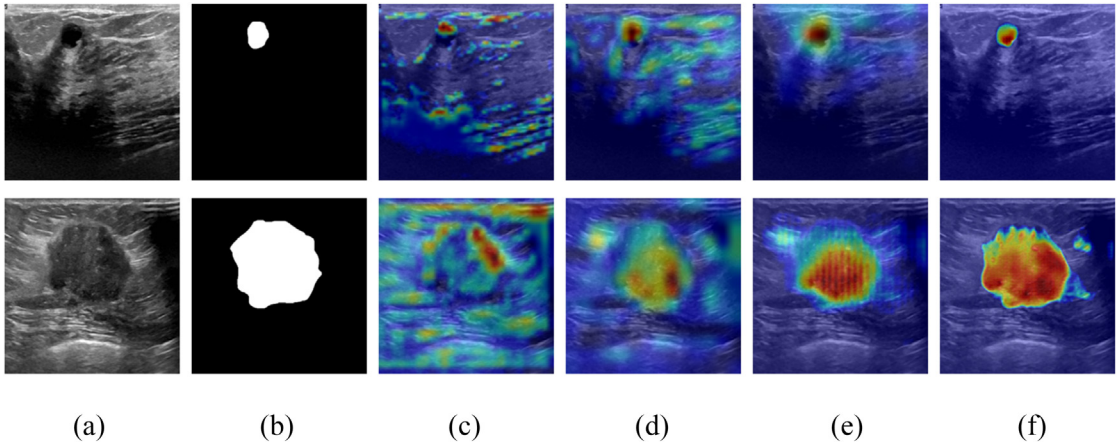

Fig. 19. Grad-CAM outcomes extracted from middle layers of proposed RFS-UNet. The first and second rows depict results obtained for benign and malignant tumor images: (a) input images, (b) ground-truth images, and Grad-CAM outcomes extracted from (c) RFS modules1 of EB1, (d) RFS module5 of EB5, (e) RFS module6 of DB1, and (f) RFS module9 of DB4.

slightly lower because they have an uneven shape. The shape of benign tumors is usually round, and a difference in brightness or contrast between the tumor and normal areas is observed. Therefore, the segmentation results are relatively better. Our network effectively detects tumor pixels for breast tumors of various sizes. Fig. 16(b) depicts the segmentation results on the UDIAT dataset. Despite the occurrence of various breast tumors, most tumor pixels are well-classified.

5. Discussion

Finally, we applied the gradient-weighted class-activation map (Grad-CAM) method. That is, we prove that our networks can learn tumor features segmented from breast ultrasound images.

Because the layers are stacked deep in the CNN model, determining whether the important features have been appropriately extracted from the input is challenging. Therefore, the reliability of an outcome can be improved by identifying the exact position in an image, where the CNN focuses on extracting and determining features of a certain class. By using Grad-CAM, the part of an image considered by the CNN during training can be identified. In the Grad-CAM image, the minimum and maximum activation values are represented in blue and red.

Fig. 17 depicts the Grad-CAM images acquired from the middle layers of the MSSE-ResNet101. In the input image of the abnormal case, the tumor boundary is unclear. Therefore, tumors cannot easily be identified with the naked eye. However, as the network deepens, a region characterized by high activations is formed in the central part of the tumor. By contrast, in the normal image, the activation value is high in the initial layer of the network. However, in the Grad-CAM image depicted in Fig. 17f, the activation value decreases significantly. Fig. 18 depicts the Grad-CAM images acquired from the middle layers of the MSSE-DenseNet121. In the case of an abnormal image, a high activation value corresponding to the presence of a tumor is observed in the deep layers. In the normal image, the activation value of the tumor remained low in the initial layers, as illustrated in Fig. 18c.

Fig. 19 shows the Grad-CAM images acquired from the middle layers of the proposed RFS-UNet. In Fig. 19c and 19d, it is seen that the activation values in the parts excluding the tumor region decrease as the information passes through the encoder blocks. As shown in Fig. 19e and 19f, a tumor region characterized by high activation values is formed as the information passes through successive decoder blocks. The Grad-Cam results reveal that both the proposed BTEC-Net and RFS-UNet can accurately identify breast tumor regions. Therefore, both the models were appropriately trained to extract good features and enhance segmentation accura-

cies. As shown by the experimental results, our method is suitable for classifying and detecting breast tumors from ultrasound images.

Figs. 17 and 18 depict the Grad-CAM of the MSSE-DenseNet121 and MSSE-ResNet101, and the high activation values formed in the pixel area of the breast tumor are attributed to the final classification results. In the case of the Grad-CAM results of the segmentation network in Fig. 19, the high activation values are distributed similarly to the shape of the breast tumor as the layer deepens. This is similar to the results of the final segmentation output. By comparing with the results depicted in Fig. 19 with those depicted in Figs. 17 and 18, it is seen that the activation is well distributed in the central location of the breast tumor, but it does not accurately represent the shape of the breast tumor. This is because the output of the segmentation network contains and maintains more spatial information than the classification network.

6. Conclusion

This study presented a multistage breast tumor segmentation network that identified tumor regions in breast ultrasound images. The proposed network comprised classification and segmentation stages. The first classified whether an input image was normal or abnormal using the BTEC-Net model, which performed feature-level fusion by combining two subnetworks, thereby forming an ensemble network. If the image was classified as normal, the output was a result in which all the image pixel values were set to zero. Based on this approach, the occurrence of FP errors could be significantly reduced. In the next stage, the RFS-UNet model was used to perform segmentation only on images classified as abnormal. The experimental results obtained in this study using the BUSI and UDIAT datasets confirmed that the accuracy of our method was higher than those of conventional methods. Through Grad-CAM analysis, our model showed that the final output was derived by focusing on the breast tumor area and extracting the features of that area.

In future studies, we will develop a method capable of simultaneously performing classification and segmentation within a single network. Furthermore, we intend to employ deep learning-based augmentation techniques to secure the data required for network training.

Declaration of Competing Interest

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

Acknowledgments

This research was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) through the Basic Science Research Program (NRF-2021R1F1A1045587), in part by the NRF funded by the MSIT through the Basic Science Research Program (NRF-2022R1F1A1064291), and in part by the MSIT, Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Abraham, N., Khan, N.M., 2019. A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. In: In: IEEE International Symposium on Biomedical Imaging (ISBI), pp. 683–687. https://doi.org/10.1109/ISBI.2019.8759329.

Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. Data Brief 28, 1–5. https://doi.org/10.1016/j.dib.2019.104863.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., Andre, M., 2020. Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomed. Signal Process. Control 61, 1–10. https://doi.org/10.1016/j.bspc.2020.102027.

Chen, G., Dai, Y., Zhang, J., 2022. C-Net: cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation. Comput. Methods Programs Biomed. 255, 1–9. https://doi.org/10.1016/j.cmpb.2022.107086.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: In: European Conference on Computer Vision (ECCV), pp. 801–818.

Chollet, F., 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258.

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155.

Gómez, W., Leija, L., Alvarenga, A.V., Infantosi, A.F.C., Pereira, W.C.A., 2010. Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation. Med. Phys. 37, 82–95. https://doi.org/10.1118/1.3265959.

Gómez-Flores, W., Ruiz-Ortega, B.A., 2016. New fully automated method for segmentation of breast lesions on ultrasound based on texture analysis. Ultrasound Med. Biol. 42, 1637–1650. https://doi.org/10.1016/j.ultrasmedbio.2016.02.016.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2018. Squeeze-and-excitation Networks. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141.

Huang, Q., Bai, X., Li, Y., Jin, L., Li, X., 2014. Optimized graph-based segmentation for ultrasound images. Neurocomputing 129, 216–224. https://doi.org/10.1016/j.neucom.2013.09.038.

Huang, Q., Huang, Y., Luo, Y., Yuan, F., Li, X., 2020. Segmentation of breast ultrasound image with semantic classification of superpixels. Med. Image Anal. 61, 1–13. https://doi.org/10.1016/j.media.2020.101657.

Huang, Q.-H., Lee, S.-Y., Liu, L.-Z., Lu, M.-H., Jin, L.-W., Li, A.-H., 2012. A robust graph-based segmentation method for breast tumors in ultrasound images. Ultrasonics 52, 266–275. https://doi.org/10.1016/j.ultras.2011.08.011.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708.

Kekre, H., Shrinath, P., 2013. Tumour delineation using statistical properties of the breast US images and vector quantization based clustering algorithms. Int. J. Image Graph. Signal Process. 5, 1–12. https://doi.org/10.5815/ijigsp.2013.11.01.

Lee, H., Park, J., Hwang, J.Y., 2020. Channel attention module with multi-scale grid average pooling for breast cancer segmentation in an ultrasound image. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67, 1344–1353. https://doi.org/10.1109/TUFFC.2020.2972573.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

Lou, M., Meng, J., Qi, Y., Li, X., Ma, Y., 2022. MCRNet: multi-level context refinement network for semantic segmentation in breast ultrasound imaging. Neurocomputing 470, 154–169. https://doi.org/10.1016/j.neucom.2021.10.102.

Mishra, A.K., Roy, P., Bandyopadhyay, S., Das, S.K., 2022. CR-SSL: a closely related self-supervised learning based approach for improving breast ultrasound tumor segmentation. Inter. J. Imaging Syst. Technol. 32, 1209–1220. https://doi.org/10.1002/ima.22693.

NVIDIA GeForce RTX 3080. https://www.nvidia.com/ko-kr/geforce/graphics-cards/30-series/rtx-3080-3080ti/ (accessed 11 July 2021).

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas. In: In: Conference on Medical Imaging With Deep Learning (MIDL), pp. 1–10.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520.

Shan, J., Cheng, H.D., Wang, Y., 2012a. A novel segmentation method for breast ultrasound images based on neutrosophic L-means clustering. Med. Phys. 39, 5669–5682. https://doi.org/10.1118/1.4747271.

Shan, J., Cheng, H.D., Wang, Y., 2012b. Completely automated segmentation approach for breast ultrasound images using multiple-domain features. Ultrasound Med. Biol. 38, 262–275. https://doi.org/10.1016/j.ultrasmedbio.2011.10.022.

Shareef, B., Xian, M., Vakanski, A., 2020. Stan: Small Tumor-aware Network for Breast Ultrasound Image Segmentation. In: In: IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. https://doi.org/10.1109/ISBI45749.2020.9098691.

Shen, X., Wang, L., Zhao, Y., Liu, R., Qian, W., Ma, H., 2022. Dilated transformer: residual axial attention for breast ultrasound image segmentation. Quant. Imaging Med. Surg. 12, 4512–4528. https://doi.org/10.21037/qims-22-33.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-scale Image Recognition. In: In: International Conference on Learning Representations (ICLR), pp. 1–14.

Student's T-Test. https://en.wikipedia.org/wiki/Student%27s_t-test (accessed 29 August 2022).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826.

TensorFlow. https://www.tensorflow.org/api_docs/python/tf?version=nightly (accessed 12 July 2021).

Wang, Y., Yao, Y., 2022. Breast lesion detection using an anchor-free network from ultrasound images with segmentation-based enhancement. Sci. Rep. 12, 1–11. https://doi.org/10.1038/s41598-022-18747-y.

Xue, C., Zhu, L., Fu, H., Hu, X., Li, X., Zhang, H., Heng, P.-A., 2021. Global guidance network for breast lesion segmentation in ultrasound images. Med. Image Anal. 70, 1–16. https://doi.org/10.1016/j.media.2021.101989.

Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R., 2018. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Health Inform. 22, 1218–1226. https://doi.org/10.1109/JBHI.2017.2731873.

Zhang, G., Zhao, K., Hong, Y., Qiu, X., Zhang, K., Wei, B., 2021. SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. Int. J. Comput. Assist. Radiol. Surg. 16, 1–7. https://doi.org/10.1007/s11548-021-02445-7.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890.

Zhou, Q., Wang, Q., Bao, Y., Kong, L., Jin, X., Ou, W., 2022. LAEDNet: a lightweight attention encoder–decoder network for ultrasound medical image segmentation. Comput. Electr. Eng. 99, 1–12. https://doi.org/10.1016/j.compeleceng.2022.107777.

Zhu, L., Chen, R., Fu, H., Xie, C., Wang, L., Wan, L., Heng, P.-A., 2020. A Second-Order Subregion Pooling Network for Breast Lesion Segmentation in Ultrasound. In: In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 160–170. https://doi.org/10.1007/978-3-030-59725-2_16.