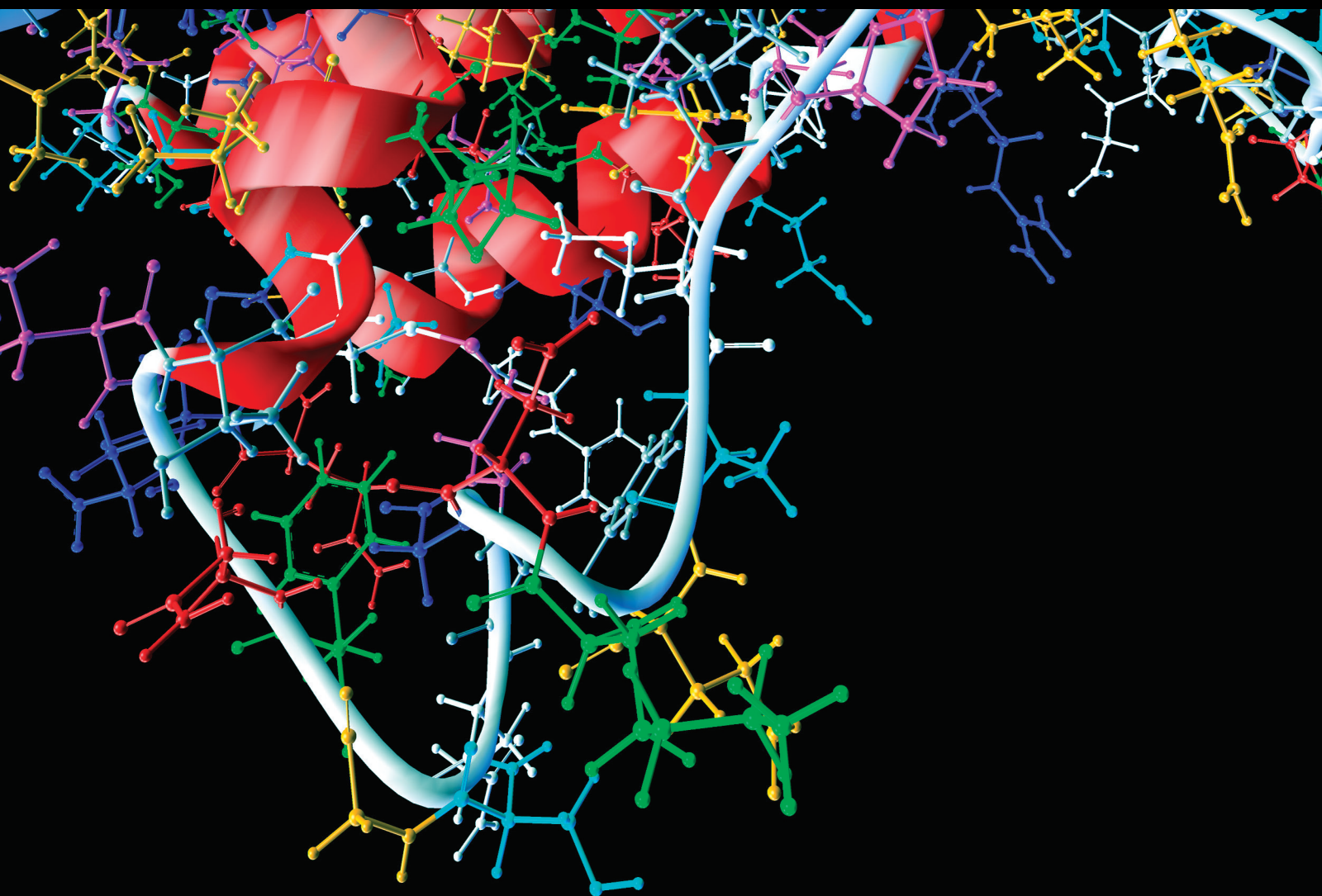


Deep Learning as a Gateway to Predictive, Preventive, Personalized, and Participatory Medicine

Lead Guest Editor: Andrea Duggento

Guest Editors: Allegra Conti, Manuel Scimeca, and Nicola Toschi





**Deep Learning as a Gateway to Predictive,
Preventive, Personalized, and Participatory
Medicine**

**Deep Learning as a Gateway to
Predictive, Preventive, Personalized,
and Participatory Medicine**

Lead Guest Editor: Andrea Duggento




Guest Editors: Allegra Conti, Manuel Scimeca, and
Nicola Toschi



Copyright © 2020 Hindawi Limited. All rights reserved.

This is a special issue published in "Computational and Mathematical Methods in Medicine." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Associate Editors

Ahmed Albahri, Iraq
Konstantin Blyuss , United Kingdom
Chuangyin Dang, Hong Kong
Farai Nyabadza , South Africa
Kathiravan Srinivasan , India

Academic Editors

Laith Abualigah , Jordan
Yaser Ahangari Nanekaran , China
Mubashir Ahmad, Pakistan
Sultan Ahmad , Saudi Arabia
Akif Akgul , Turkey
Karthick Alagar, India
Shadab Alam, Saudi Arabia
Raul Alcaraz , Spain
Emil Alexov, USA
Enrique Baca-Garcia , Spain
Sweta Bhattacharya , India
Junguo Bian, USA
Elia Biganzoli , Italy
Antonio Boccaccio, Italy
Hans A. Braun , Germany
Zhicheng Cao, China
Guy Carrault, France
Sadaruddin Chachar , Pakistan
Prem Chapagain , USA
Huiling Chen , China
Mengxin Chen , China
Haruna Chiroma, Saudi Arabia
Watcharaporn Cholamjiak , Thailand
Maria N. D.S. Cordeiro , Portugal
Cristiana Corsi , Italy
Qi Dai , China
Nagarajan Deivanayagam Pillai, India
Didier Delignières , France
Thomas Desaive , Belgium
David Diller , USA
Qamar Din, Pakistan
Irina Doytchinova, Bulgaria
Sheng Du , China
D. Easwaramoorthy , India


Esmaeil Ebrahimie , Australia
Issam El Naqa , USA
Ilias Elmouki , Morocco
Angelo Facchiano , Italy
Luca Faes , Italy
Maria E. Fantacci , Italy
Giancarlo Ferrigno , Italy
Marc Thilo Figge , Germany
Giulia Fiscon , Italy
Bapan Ghosh , India
Igor I. Goryanin, Japan
Marko Gosak , Slovenia
Damien Hall, Australia
Abdulsattar Hamad, Iraq
Khalid Hattaf , Morocco
Tingjun Hou , China
Seiya Imoto , Japan
Martti Juhola , Finland
Rajesh Kaluri , India
Karthick Kanagarathinam, India
Rafik Karaman , Palestinian Authority
Chandan Karmakar , Australia
Kwang Gi Kim , Republic of Korea
Andrzej Kloczkowski, USA
Andrei Korobeinikov , China
Sakthidasan Sankaran Krishnan, India
Rajesh Kumar, India
Kuruva Lakshmana , India
Peng Li , USA
Chung-Min Liao , Taiwan
Pinyi Lu , USA
Reinoud Maex, United Kingdom
Valeri Makarov , Spain
Juan Pablo Martínez , Spain
Richard J. Maude, Thailand
Zahid Mehmood , Pakistan
John Mitchell , United Kingdom
Fazal Ijaz Muhammad , Republic of Korea
Vishal Nayak , USA
Tongguang Ni, China
Michele Nichelatti, Italy
Kazuhisa Nishizawa , Japan
Bing Niu , China

Hyuntae Park , Japan
Jovana Paunovic , Serbia
Manuel F. G. Penedo , Spain
Riccardo Pernice , Italy
Kemal Polat , Turkey
Alberto Policriti, Italy
Giuseppe Pontrelli , Italy
Jesús Poza , Spain
Maciej Przybyłek , Poland
Bhanwar Lal Puniya , USA
Mihai V. Putz , Romania
Suresh Rasappan, Oman
Jose Joaquin Rieta , Spain
Fathalla Rihan , United Arab Emirates
Sidheswar Routray, India
Sudipta Roy , India
Jan Rychtar , USA
Mario Sansone , Italy
Murat Sari , Turkey
Shahzad Sarwar, Saudi Arabia
Kamal Shah, Saudi Arabia
Bhisham Sharma , India
Simon A. Sherman, USA
Mingsong Shi, China
Mohammed Shuaib , Malaysia
Prabhishek Singh , India
Neelakandan Subramani, India
Junwei Sun, China
Yung-Shin Sun , Taiwan
Min Tang , China
Hongxun Tao, China
Alireza Tavakkoli , USA
João M. Tavares , Portugal
Jlenia Toppi , Italy
Anna Tsantili-Kakoulidou , Greece
Markos G. Tsipouras, North Macedonia
Po-Hsiang Tsui , Taiwan
Sathishkumar V E , Republic of Korea
Durai Raj Vincent P M , India
Gajendra Kumar Vishwakarma, India
Liangjiang Wang, USA
Ruisheng Wang , USA
Zhouchao Wei, China
Gabriel Wittum, Germany
Xiang Wu, China


KI Yanover , Israel
Xiaojun Yao , China
Kaan Yetilmezsoy, Turkey
Hiro Yoshida, USA
Yuhai Zhao , China

Contents








Deep Convolutional Neural Networks-Based Automatic Breast Segmentation and Mass Detection in DCE-MRI

Han Jiao , Xinhua Jiang, Zhiyong Pang , Xiaofeng Lin, Yihua Huang , and Li Li 
Research Article (12 pages), Article ID 2413706, Volume 2020 (2020)

Interactive Echocardiography Translation Using Few-Shot GAN Transfer Learning

Long Teng , ZhongLiang Fu, Qian Ma, Yu Yao, Bing Zhang, Kai Zhu, and Ping Li
Research Article (9 pages), Article ID 1487035, Volume 2020 (2020)

Multicenter Computer-Aided Diagnosis for Lymph Nodes Using Unsupervised Domain-Adaptation Networks Based on Cross-Domain Confounding Representations

RuoXi Qin , Huike Zhang , LingYun Jiang, Kai Qiao , Jinjin Hai , Jian Chen , Junling Xu, Dapeng Shi , and Bin Yan 
Research Article (10 pages), Article ID 3709873, Volume 2020 (2020)

A Long Short-Term Memory Ensemble Approach for Improving the Outcome Prediction in Intensive Care Unit

Jing Xia , Su Pan, Min Zhu, Guolong Cai, Molei Yan, Qun Su, Jing Yan , and Gangmin Ning 
Research Article (10 pages), Article ID 8152713, Volume 2019 (2019)

Research Article

Deep Convolutional Neural Networks-Based Automatic Breast Segmentation and Mass Detection in DCE-MRI

Han Jiao ¹, Xinhua Jiang,² Zhiyong Pang ¹, Xiaofeng Lin,² Yihua Huang ¹ and Li Li ²

¹School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China

²Department of Medical Imaging, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, China

Correspondence should be addressed to Yihua Huang; huangyih@mail.sysu.edu.cn and Li Li; li2@mail.sysu.edu.cn

Received 25 November 2019; Accepted 13 February 2020; Published 5 May 2020

Guest Editor: Allegra Conti

Copyright © 2020 Han Jiao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Breast segmentation and mass detection in medical images are important for diagnosis and treatment follow-up. Automation of these challenging tasks can assist radiologists by reducing the high manual workload of breast cancer analysis. In this paper, deep convolutional neural networks (DCNN) were employed for breast segmentation and mass detection in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). First, the region of the breasts was segmented from the remaining body parts by building a fully convolutional neural network based on U-Net++. Using the method of deep learning to extract the target area can help to reduce the interference external to the breast. Second, a faster region with convolutional neural network (Faster RCNN) was used for mass detection on segmented breast images. The dataset of DCE-MRI used in this study was obtained from 75 patients, and a 5-fold cross validation method was adopted. The statistical analysis of breast region segmentation was carried out by computing the Dice similarity coefficient (DSC), Jaccard coefficient, and segmentation sensitivity. For validation of breast mass detection, the sensitivity with the number of false positives per case was computed and analyzed. The Dice and Jaccard coefficients and the segmentation sensitivity value for breast region segmentation were 0.951, 0.908, and 0.948, respectively, which were better than those of the original U-Net algorithm, and the average sensitivity for mass detection achieved 0.874 with 3.4 false positives per case.

1. Introduction

Breast cancer is one of the most common cancers amongst women worldwide [1]. Early diagnosis and treatment are proven to reduce the mortality rate [2]. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a more reliable tool for early detection of breast cancer than mammography and ultrasound [3]. The correct resolution of DCE-MRI image of the breast depends largely on the quality of visualization, operation experience, and the time needed for data analysis. Because manual analysis of MR series is time-consuming and error-prone, several specific systems have been developed to help radiologists detect and diagnose breast lesions, which greatly improves clinicians' work efficiency [4]. Although some computer-aided diagnosis (CAD) systems are currently used in the clinic, fully

automatic detection of breast lesions is still an ongoing problem [5].

Generally, a DCE-MRI image of the breast also includes other organs such as the lung, heart, liver, and pectoral muscles. Separation of the breast region from other organs is necessary for further analysis. Manual segmentation of the breast region is tedious; therefore it is impractical to segment the entire breast region from a large number of diverse datasets. Automatic segmentation of the breast region reduces result bias and can accelerate data processing. Conventional breast region segmentation methods proposed in the literature are based on threshold, morphology, and fuzziness with specific connections between pixels. Thakran et al. segmented the outer and inner breast tissue by thresholding, morphological operation, and B-spline curve fitting [6]. Wu et al. [7] combined edge enhancement and

edge linking with candidate evaluation to detect the chest-wall line in the sagittal plane. Jiang et al. [8] used dynamic programming combined with preprocessing to segment the breast region in fat-suppressed transverse DCE-MRI. While these traditional methods have shown good performance, the robustness of these algorithms is insufficient, in that they tend to fail in some specific data because they can only process the underlying information. In recent years, convolutional neural networks (CNNs) have been widely used in biomedical semantic segmentation such as FCN [9], SegNet [10], and U-Net [11]. These deep networks can automatically extract high-level features and complete pixel-wise segmentation. Xu et al. used a 2D U-Net for automatic breast region segmentation in DCE-MRI [12]. Adoui et al. built two fully CNNs based on SegNet and U-Net for breast tumor segmentation [13]. Building on the success of U-Net in medical image segmentation, this study additionally used U-Net++ [14] for comparison of DCE-MRI images for breast region segmentation. Our results indicated the effectiveness and accuracy of this method in biomedical segmentation tasks.

Because of the varying sizes, shapes, appearances, and densities of masses, CAD for breast mass detection is a challenging task [15, 16]. Conventional methods for breast mass detection mainly rely on threshold values [17] or mass templates [18] based on various kinds of filter operators. Huang et al. used multiscale Hessian-based analysis for breast mass detection [19], and Wang et al. detected breast masses based on Gestalt psychology [20]. These traditional detection algorithms are sensitive to image noise, and the hand-designed features are not adequately robust to blurred contrast and bias-field. Deep convolutional neural networks (DCNNs) have significantly outperformed traditional methods in recent years, owing to the strong feature expression ability that can further improve the detection accuracy. With the development of deep learning, DCNN has been widely used in medical image detection [21–23]. These CAD systems can be used to migrate between breast cancer and lung cancer. The application of a CAD system helps the radiologist as a second reviewer to evaluate screening medical images. A CAD system based on one of the most successful object detection frameworks—Faster RCNN [24]—could achieve high sensitivity with few false positive results on the datasets of mammograms and tomosynthesis [25, 26]. In the field of mammary MRI, the development of deep learning frameworks is limited, because only a small number of datasets are available. With this background, we incorporated Faster RCNN into our DCE-MRI datasets for detection of breast mass.

In this paper, DCNN-based frameworks were present for computer-aided segmentation of breast region and detection of breast masses. Breast region segmentation was first performed by U-Net++ to remove other organs except the breast region to eliminate interference. U-Net was also implemented as a comparison of segmentation framework. Given the limited amount of our datasets, data augmentation was used during training to reduce overfitting. Breast mass detection was implemented by training Faster RCNN with images preprocessed by segmentation network and

labels annotated by radiologists. The proposed methods were validated by a 5-fold cross validation technique to obtain more reliable results than independent tests. Performance of the automatic segmentation was evaluated using similarity coefficients. The breast mass detection model was evaluated using sensitivity at the number of false positives per case.

2. Materials and Methods

2.1. Data. The dataset consisted of 75 women (mean age, 47 years; age range, 18–68 years) with histopathologically confirmed breast masses at the Sun Yat-sen University Cancer Center (Guangzhou, China). Patients with suspicious breast masses were recruited after they provided written informed consent. The Ethics Committee of Sun Yat-sen University Cancer Center approved the study. Patients were scanned in the prone position with the bilateral breast naturally hanging into the two holes of the coil. A 3.0 T superconductive magnetic system (Discovery 750, GE Healthcare) with a bilateral 8-channel phased-array breast-specific surface coil was used for the imaging. Standard imaging was performed, including axial fast spin echo (FSE) T1WI, and axial and sagittal FSE T2WI. Subsequently, diffusion-weighted images (DWI) were acquired in the axial planes. The DCE-MRI data were acquired using the VI-BRANT-FLEX technique in the axial orientation, after one set of unenhanced baseline images were obtained using an MRI-specific automatic power injector (Medrad Inc., Pittsburgh, PA, USA) to inject 0.1 mmol/kg body weight contrast medium (gadopentetate dimeglumine; Magnevist, Bayer Schering Pharma, Berlin, Germany), with a hand venipuncture technique at a rate of 3 mL/s. Saline (10 mL at 3 mL/s) was then injected to wash the tube. Dynamic scanning was initiated by simultaneously pushing the high-pressure syringe button and the dynamic scan button. Eight postcontrast sets were acquired under the following scanning conditions: field of view, 32 cm; matrix, 320 × 320; section thickness without a gap, 1.4 mm; repetition time, 3.9 ms; echo time, 1.7 ms; and flip angle, 5°. The patients had not received any treatment before undergoing MRI. In this study, only mass-like masses showed strong contrast.

Diagnosis was confirmed following pathological analysis subsequent to core-needle biopsy or surgical excision, or lesion not changed at a minimum follow-up of 2 years defined as benign lesion.

All images were analyzed independently by two radiologists with ten years of experience in breast MRI. The images were assessed independently and any disagreements were resolved by achieving consensus. All lesions were assessed using the Breast Imaging Reporting and Data System (BI-RADS). BI-RADS category 1 (negative) and category 2 (benign) denote an essentially 0% likelihood of cancer. BI-RADS category 3 (probably benign) assessment is more intuitive and can be recommended in the case of a unique focal finding for which the likelihood of malignancy is $\geq 0\%$ but $\leq 2\%$. BI-RADS category 4 (suspicious) and category 5 (highly suggestive of malignancy) describe MRI findings that are suspicious enough to warrant tissue diagnosis. BI-RADS category 6 (known biopsy-proven

malignancy) describes MRI findings of biopsy-proven breast cancer for which surgical excision is recommended when clinically appropriate.

The objective of this study is to detect the mass. All lesions with a BI-RADS score greater than 1 were used in the dataset. According to statistics, the diameter of the mass ranges from 1.07 to 6.69 cm.

3. Methods

Our fully automatic method consists of two parts: breast region segmentation and breast mass detection. Automatic segmentation of breast region is a challenging task because of large variations in breast shapes, sizes, image artifacts, and other noise-induced errors. Inspired by the success of DCNN models on object segmentation tasks, we implemented a U-Net++ framework for breast region segmentation, along with a U-Net framework for comparison. Next, we obtained images of the breast region without interference from other organs/body parts as the input for Faster RCNN to detect breast masses. Finally, the positions and sizes of the breast masses were identified. The whole framework of this proposed method is shown in Figure 1.

U-Net [11] was used in breast region segmentation for comparison, and its architecture is shown in Figure 2. One advantage of this method is its robustness even with small training data. The architecture consists of downsampling (left side) and upsampling (right side). The left side acts as an encoder and extracts features through the network. It contains a typical convolutional structure: two 3×3 convolution operations and one 2×2 max-pooling operation. Each convolution operation is followed by batch normalization (BN) and a rectified linear unit (ReLU). The right side acts as a decoder and also contains a typical architecture: a 2×2 upsampling operation and two 3×3 convolution operations, each followed by BN and a ReLU. Contracting paths are used to combine the high resolution feature maps with the upsampling outputs to accurately classify and locate each pixel. Every pixel of input image is classified as breast region or background. To evaluate the segmentation loss, we used the binary cross entropy (BCE) loss function.

$$\text{BCE Loss} = -\frac{1}{M} \sum_{i=1}^M (Y_i \cdot \log(\hat{Y}_i) + (1 - Y_i) \cdot \log(1 - \hat{Y}_i)), \quad (1)$$

where Y is the ground truth and \hat{Y} is the predicted probability for all the M pixels.

U-Net++ [14] is an improved version of U-Net for biomedical image segmentation. For the skip connection of U-Net, U-Net++ adds modules to integrate the features of different levels through superposition. Further, to ensure gradient propagation, U-Net++ uses a deep supervision scheme that connects the middle module to the final output, finally forming a dense block structure as shown in Figure 3. The loss function is a combination of binary cross entropy and dice coefficient on each of the above four semantic levels, which is described as

$$\text{BCE\&Dice Loss} = -\frac{1}{N} \sum_{b=1}^N \left(\frac{1}{2} \cdot Y_b \cdot \log(\hat{Y}_b) + \frac{2 \cdot Y_b \cdot \hat{Y}_b}{Y_b + \hat{Y}_b} \right), \quad (2)$$

where Y_b and \hat{Y}_b denote the flattened ground truths and the flattened predicted probabilities of the b^{th} image, respectively, and N indicates the batch size.

The object detection network named Faster RCNN [24] is composed of two modules. The first module is a deep fully CNN that proposes regions and the second module is the detector that uses the proposed regions. Feature map is extracted from input image by using ResNet-101 [27]. ResNet-101 has been proven to perform well on classification tasks, which shows that it has good feature extraction ability. A region proposal network (RPN) considers the feature map as input and provides a set of rectangular object proposals as the output, each with an objectness score. The RPN structure in [24] is used in this paper without modification. The function of ROI Align is to map region of interest (ROI) areas of different sizes to feature maps of fixed sizes. ROI Align [28] can effectively solve the problem of misalignment caused by twice quantization in ROI Pooling. For the detection of large objects, the difference between the two schemes is rare. If there are more small objects in the picture to be detected, ROI Align is preferred, which is more accurate. In this paper, the masses of breast MRI image are taken as the research target, and the diameter of them is less than 48 pixels, which belongs to small targets. Therefore, ROI Align is selected for detection network. The feature is shared by RPN and fully connected (FC) layers. The structure is shown in Figure 4.

We computed four regression losses. L_{reg} was associated with predicted mass coordinates, widths, and height, and L_{cls} was the classification loss for the predicted mass probabilities. The ground truth labels are determined for each anchor as follows: If an anchor i overlaps with a mass with an intersection over union (IOU) greater than 0.7, then it is regarded as positive ($p_i^* = 1$). On the contrary, if an anchor i overlaps with a mass with an IOU less than 0.3, it is regarded as negative ($p_i^* = 0$). All other anchors do not contribute to the loss, and only positive anchors contribute to the regression loss. The final loss function for anchor i is defined as follows:

$$L(p_i, t_i) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \frac{1}{N_{\text{reg}}} \sum_i L_{\text{reg}}(t_i, t_i^*). \quad (3)$$

Here, t_i is a vector representing the four parameterized coordinates of the predicted bounding box and t_i^* is the ground truth associated with a positive anchor. We used binary cross entropy loss for L_{cls} and smooth L1 loss for L_{reg} .

3.1. Performance Evaluation. The performance of the proposed method for breast region segmentation was tested using the Dice similarity coefficient (DSC) [29], Jaccard coefficient [30], and segmentation sensitivity described by Udupa et al. [31], which are given by the following equations:

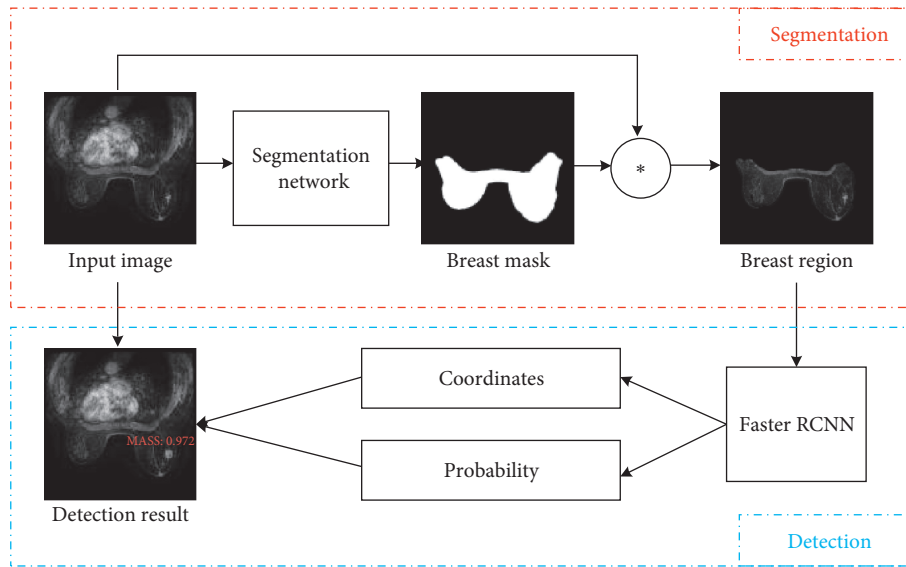


FIGURE 1: The framework of breast region segmentation and breast mass detection. The entire framework is divided into two parts. The upper part is image segmentation. Breast region mask is obtained by importing the image into segmentation network such as U-Net++ and U-Net. The breast region can be segmented by masking the input image. The following part is the target detection. The breast region image was input to Faster RCNN to obtain the location coordinates and probability of the mass.

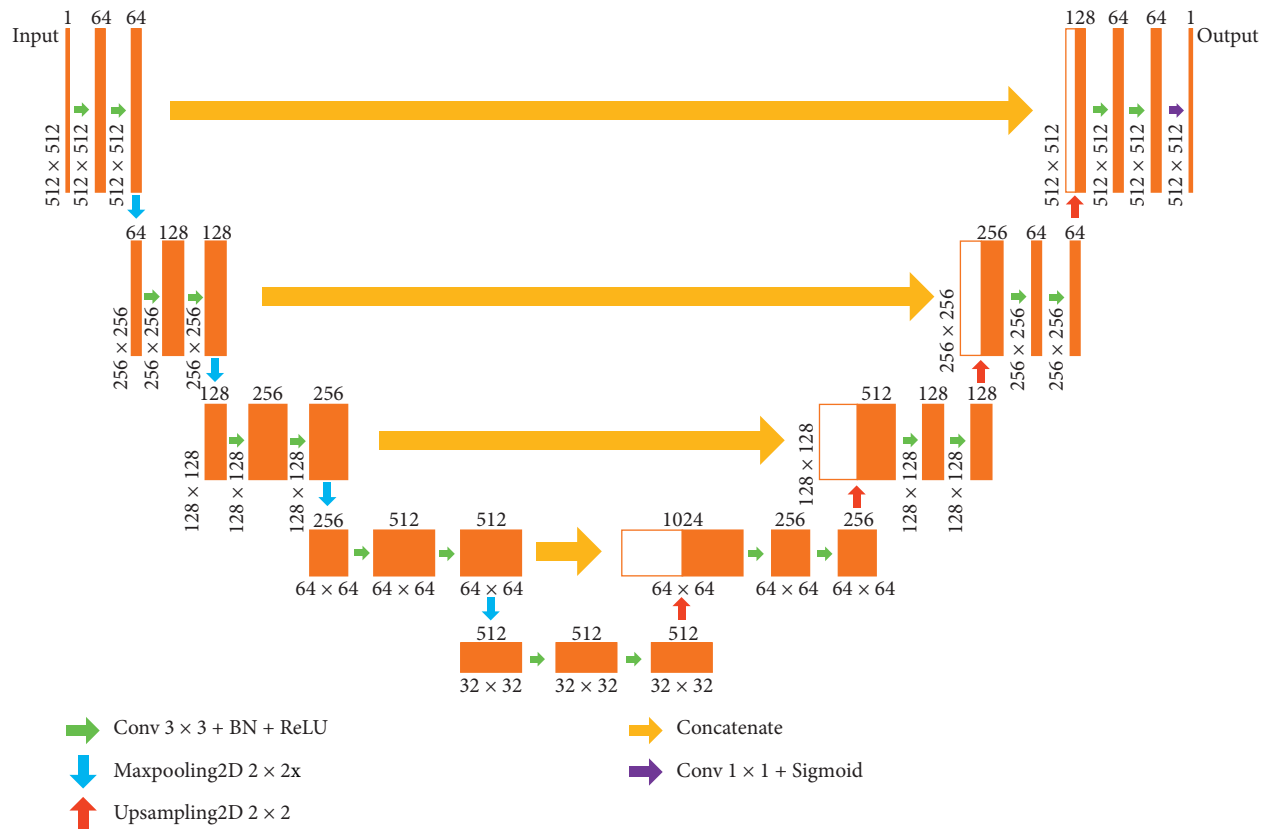


FIGURE 2: The architecture of U-Net. The network consists of two parts. The left half is feature extraction. After each pool layer, the scale of feature changes. There are five scales in total. The right half is upsampling. Every time the feature is upsampled, it will be fused with the same scale corresponding to the feature extraction part.

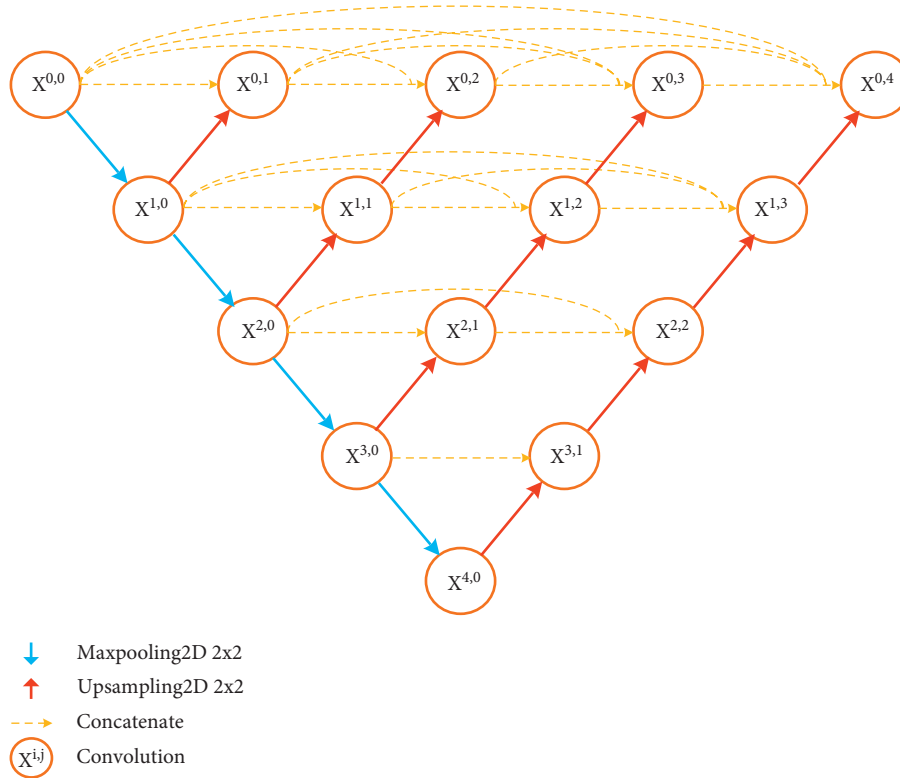


FIGURE 3: The architecture of U-Net++. Every $X^{i,j}$ includes two 3×3 convolution operations, each followed by BN and a ReLU. Channels of the same scale are connected in a dense manner for gradient propagation.

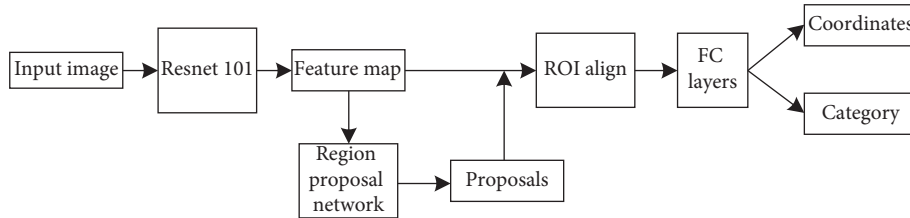


FIGURE 4: The architecture of Faster RCNN. Using ResNet-101 as the feature extraction network, the target location and probability can be obtained by processing the shared features.

$$\text{Dice Coefficient DSC} = 2 \cdot \frac{|A \cap B|}{|A| + |B|},$$

$$\text{Jaccard Coefficient Jaccard} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \quad (4)$$

$$\text{Segmentation Sensitivity SS} = \frac{|A \cap B|}{|B|},$$

where A is the automatic segmentation result and B is the ground truth.

For target detection tasks, performance is measured by sensitivity with the average number of false positives per sample. In the task of breast mass detection, the sensitivity, also known as the true positive rate (TPR), represents the proportion of the number of detected masses to the number of all masses in the dataset. This is calculated using the true positive (TP), false negative (FN), and false positive (FP). It

should be noted that true negative (TN) in target detection tasks is meaningless.

$$\text{True Positive Rate TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

The receiver operating characteristic (ROC) curve was first invented by electronics and radar engineers during World War II to detect enemy vehicles on the battlefield, i.e., signal detection theory. Soon afterwards, it was introduced into psychology to detect signal perception. Since then, it has been introduced into the field of machine learning to evaluate classification and test results and is a very important and common statistical analysis method. However, the classical ROC method cannot solve the practical problem of evaluating target detection task on an image. In the 1970s, the concept of FROC (free-response ROC) was proposed, which allows the evaluation of arbitrary anomalies on each image. The FROC curve considers the number of false

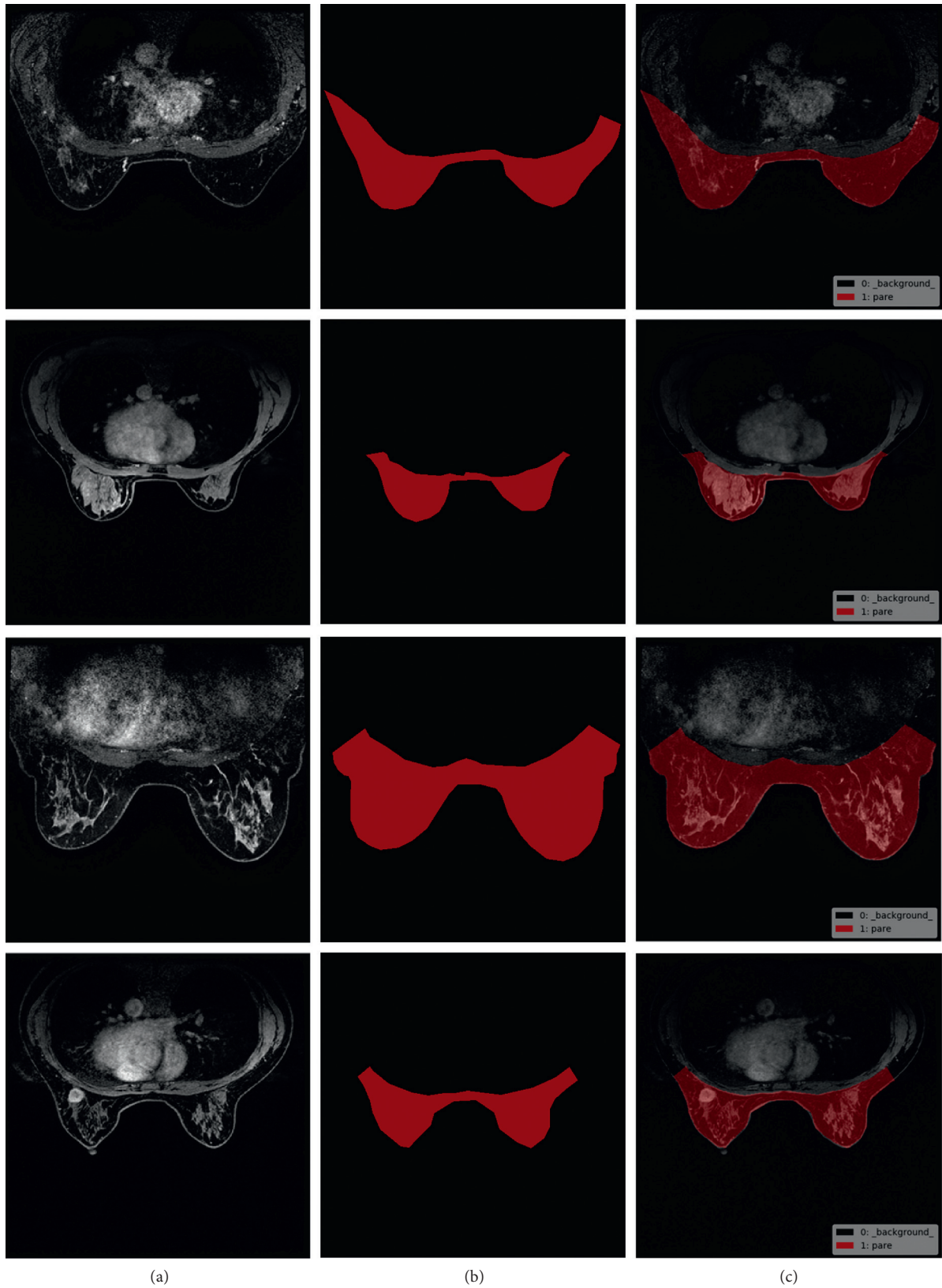


FIGURE 5: Samples of breast region segmentation dataset: (a) the original image, (b) the segmentation ground truth, and (c) (b) superimposed on (a). The pixel value of the breast region is set to one and the pixel value of the background area is set to zero during training of segmentation models.

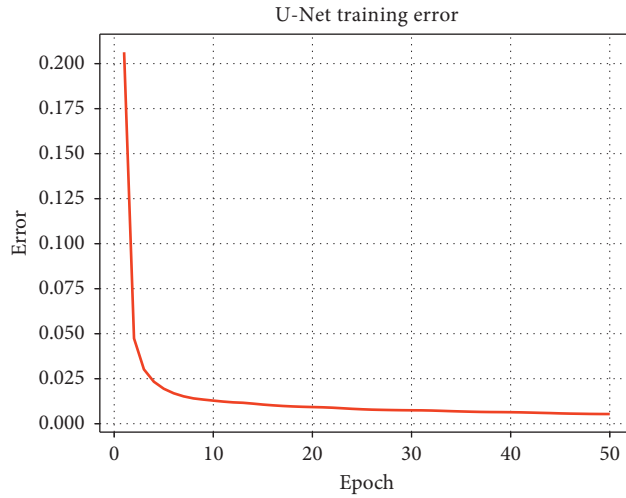


FIGURE 6: U-Net training error.

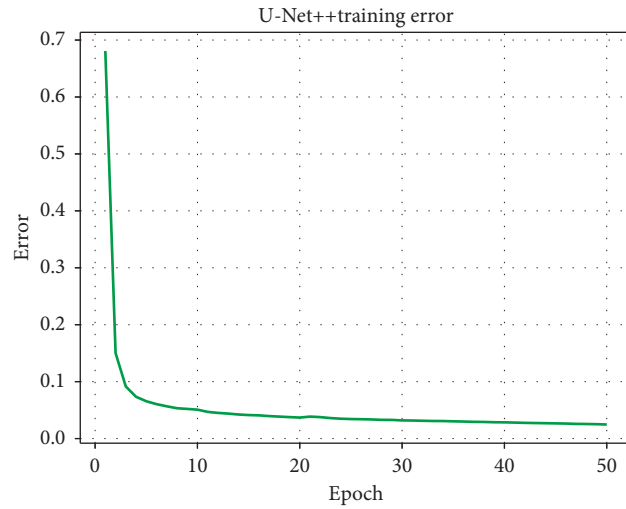


FIGURE 7: U-Net++ training error.

TABLE 1: Performance of breast region segmentation by U-Net on 5-fold cross validation.

Performance	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Average
DSC	0.949	0.952	0.926	0.938	0.938	0.941
Jaccard	0.905	0.908	0.869	0.886	0.885	0.891
Segmentation sensitivity	0.947	0.960	0.933	0.946	0.921	0.941

TABLE 2: Performance of breast region segmentation by U-Net++ on 5-fold cross validation.

Performance	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Average
DSC	0.959	0.960	0.945	0.948	0.942	0.951
Jaccard	0.921	0.924	0.897	0.904	0.894	0.908
Segmentation sensitivity	0.957	0.964	0.940	0.953	0.927	0.948

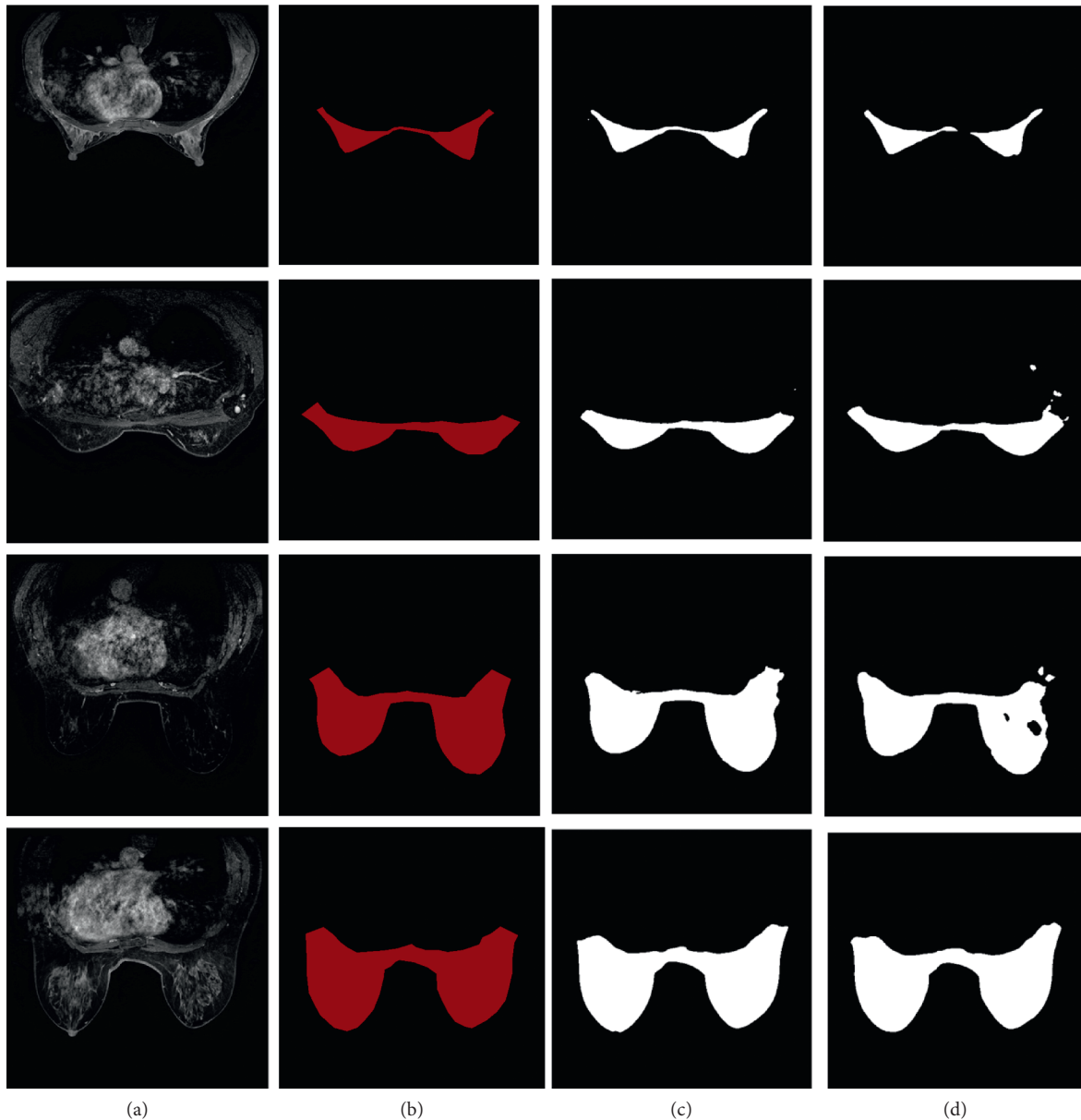


FIGURE 8: Samples of breast region segmentation result: (a), (b) source images and ground truths, respectively, (c) the breast region segmentation result of U-Net++, and (d) segmentation from U-Net.

positives in each sample as the x -axis and the recall as the y -axis. The closer the curve to the upper left corner, the better the performance of the model.

4. Results

4.1. Breast Segmentation. Owing to the uncertainty of breast MRI, traditional methods cannot at times segment breast region very well. Therefore, in this study, we adopted the deep learning method for breast region segmentation.

We proposed breast region segmentation by training a U-Net++ model, which removes interference from different DCE-MRI series external to the breast region. We used the dataset of DCE-MRIs from 75 patients. For the segmentation task, we created breast region labels for each patient

(Figure 5). Because of the similarity in consecutive images in the DCE-MRI series, we allocated one breast region label for every 10 consecutive images. We employed patient-level, 5-fold cross validation. Each subset contains 15 cases of DCE-MRI. Moreover to illustrate the robustness of our algorithm, we trained our model on four subsets and validated it on the other subset. For data augmentation, the training sets were flipped horizontally. We trained the U-Net++ model and the U-Net model as comparison for 50 epochs using stochastic gradient descent (SGD) as the optimizer and used the last epoch to predict validation datasets. The batch size was set to five given the limitation of GPU memory. The training processes of U-Net and U-Net++ are shown in Figures 6 and 7, respectively, which show the final convergence of the networks.

The segmentation performance was validated by DSC, Jaccard, and segmentation sensitivity on the validation dataset. We computed performance values of U-Net++ for each validation image, and the average of them was 0.951, 0.908, and 0.948, respectively, which were better than those of U-Net. Evaluation results are provided in Tables 1 and 2. Exp 1 to 5 are experiments from 5-fold cross validation. The performance of each experiment result was evaluated, and mean values of the Dice coefficient (DSC), Jaccard coefficient, and segmentation sensitivity between segmentation results and ground truths from 15 different cases are calculated. The average performance was assessed. The breast region segmentation results on the validation dataset are shown in Figure 8. From the overall segmentation effect, the integrity and robustness of U-Net++ segmentation were better than those of U-Net.

4.2. Mass Detection. Encouraged by the overall success of Faster RCNN and deep residual networks in natural images, we used them in our study of breast mass detection. To eliminate the impact of redundant information and external noise, we first used the well-trained U-Net++ model to preprocess images. After that, the training and validation dataset only contained information pertaining to the breast region. All the slices containing the mass were extracted, and the central coordinates and the width and height of the mass on the image were marked for training. In validation, the whole case was segmented and processed. The detection results obtained by the network were mapped to three-dimensional space, and the bounding boxes close to each other in the space would be merged into one candidate. Candidates would be evaluated as TP if the IOUs with ground truth were greater than 0.5 while FP if less than 0.5. We employed patient-level 5-fold cross validation. Each subset contains 15 cases of DCE-MRI. Further, to illustrate the robustness of our algorithm, we trained our model on 4 subsets and validated it based on the other remaining subset. In the training subset, we used 30 epochs in total with SGD optimization and a momentum of 0.9. Because of the limitation of GPU memory, the batch size parameter was set to 8. We used a weight decay of 5×10^{-4} . The initial learning rate was 0.001 and multiplied by 0.1 every 10 epochs. Before training the network, the weights of pretrained ResNet-101 were loaded for transfer learning, which effectively accelerated convergence. The training process of the network is shown in Figure 9, which shows the final convergence of the network.

We validated the Faster RCNN model based on the validation dataset and the average mass-level sensitivity achieved 0.874 at 3.4 false positives per case. The comparison of several detection results is shown in Figure 10. The network outputs the specific location and size information of different kinds of mass, as well as the probability. Regarding the 5-fold cross validation, Figure 11 shows the FROC curve of each experiment and Figure 12 shows the average FROC curve for 5-fold cross validation. The total size of the dataset is small so that it is probable that a few samples are obviously different from

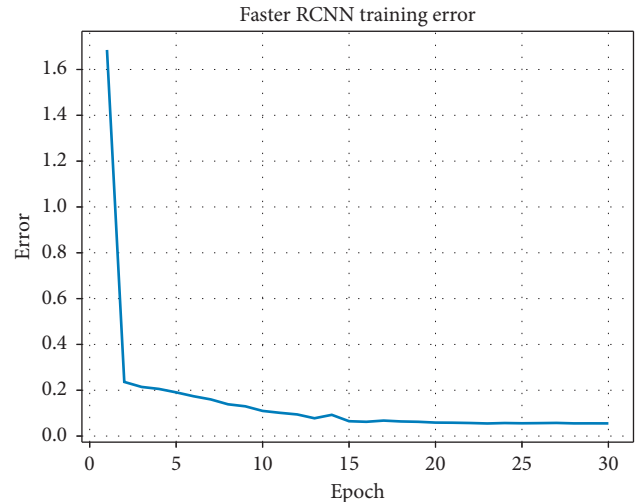


FIGURE 9: Faster RCNN training error.

other cases, which affects the accuracy of the experimental results. Therefore, the cross validation method is used to make the experimental results more convincing in small datasets. The performance of the curves supports the proposed breast mass detection method as a useful tool to assist doctors with the diagnosis.

5. Discussion

We proposed a method based on DCNN to effectively segment breast region and detect breast masses on DCE-MRI images. The success of the proposed method was mainly dependent on two aspects. First, breast region segmentation is necessary because other organs outside the breast occupy most of the image area, which greatly interferes with detection of masses. U-Net++ is a fully convolutional network applicable to various biomedical segmentation problems. The network contains convolutional layers and max-pooling layers and does not have any fully connected layers. Specifically, there are dense blocks between skip connections, which can make full use of extracted features. Owing to the lack of datasets, 5-fold cross validation was adopted. We used our own data to train U-Net++ for 50 epochs, and the DSC, Jaccard, and segmentation sensitivity values obtained were 0.951, 0.908, and 0.948, respectively, which were better than those obtained with U-Net. The experimental results showed that U-Net++ segmentation of the breast region was more precise and complete than the U-Net segmentation. As seen in Figure 8, the segmentation of the end of the breast region was not very accurate, but the key areas were well preserved, which was helpful in mass detection. Second, Faster RCNN is a flexible generic object detection framework that is easily extended to biomedical detection tasks. This study used ResNet-101 as the backbone of Faster RCNN to extract feature map from the input image. Then, there were two parallel branches to share the feature map, bounding box regression, and classification. We used the well-trained U-Net++ model to preprocess the training dataset to eliminate interference.

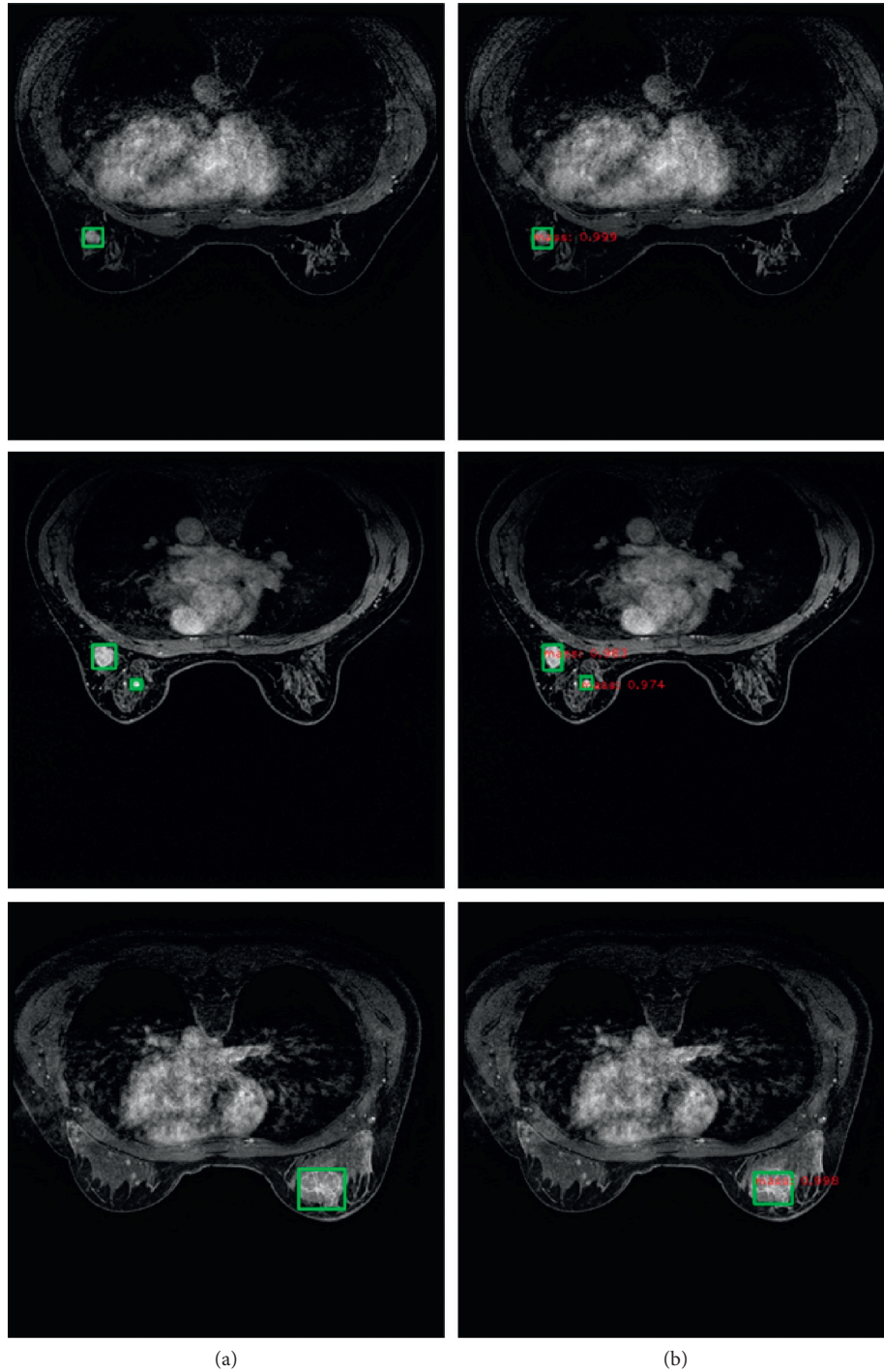


FIGURE 10: Ground truth and detection result: (a) the ground truth of breast mass detection and (b) the detection results of Faster RCNN, including target position, size, and probability.

Before training the network, the weights of pretrained ResNet-101 were loaded for transfer learning, which effectively accelerated convergence. For the performance evaluation, Faster RCNN accurately found the position of the breast mass and provided corresponding confidence information. We used 5-fold cross validation, and the average sensitivity performance achieved 0.874 with 3.4 false positives per case. This meant that our results may have a

potential value for early diagnosis and treatment of breast cancer, but more clinical cases are needed to carry out research for verification.

Our study has some limitations, but it can verify the feasibility of this method in our dataset. Future research includes detection from multiple directions as well as using 3D CNN for extracting 3D feature map. The dataset used in this paper is relatively small. We aim to establish a large-

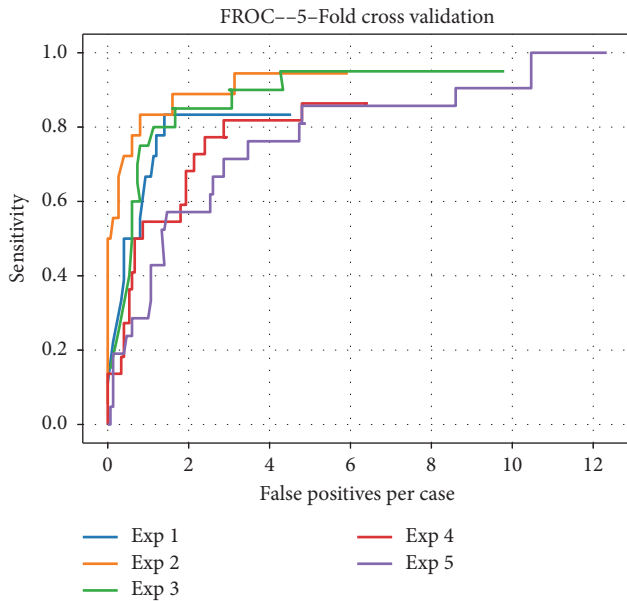


FIGURE 11: FROC of each experiment. The FROC curves of Exp 1 to Exp 5 are listed. Horizontal coordinates represent the average number of false positives per case. The vertical coordinates represent the sensitivity corresponding to the average number of false positives per case.

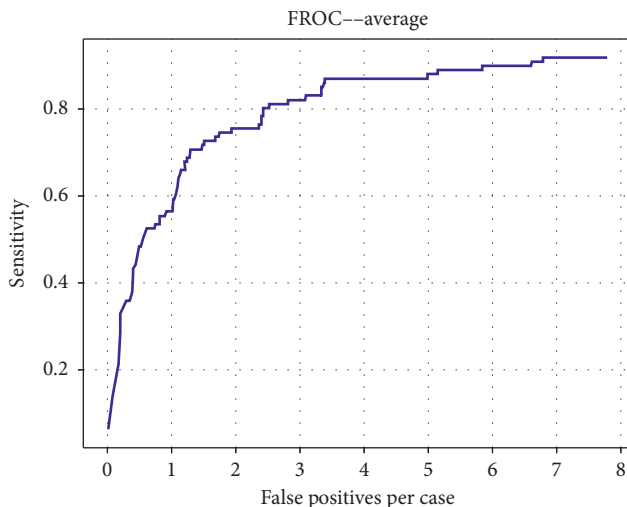


FIGURE 12: Average FROC for 5-fold cross validation.

scale open dataset for convenient performance comparison of papers with different algorithms in future.

6. Conclusion

Herein, we presented an automatic method for breast region segmentation and mass detection based on DCNN in DCE-MRIs. Our method consists of the breast region segmentation by U-Net++ and the breast mass detection by Faster RCNN. The DCNNs were trained and validated in DCE-MRIs from Sun Yat-sen University Cancer Center and showed good performance. We believe that this method provides a powerful clinical tool to help doctors accurately

and quickly diagnose breast masses. Future research will include the establishment of larger open data and the opinions of radiologists and surgeons in clinical practice.

Data Availability

The DCE-MRI data used to support the findings of this study were supplied by the Sun Yat-sen University Cancer Center (Guangzhou, China) under license and have not been made freely available because of patient privacy. If our dataset is useful to you, please contact the corresponding authors by e-mail.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Han Jiao and Xinhua Jiang contributed equally to this study.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no.61771494) and the Science and Technology Planning Project of Guangdong Province (no. 2016A010104007). This work was funded by the Science and Technology Planning Project of Guangdong Province, China (no. 2016B090918066) and the Science and Technology Program of Guangzhou, China (no. 201704020060, 201807010057).

References

- [1] N. Howlader, A. M. Noone, M. Krapcho et al., *SEER Cancer Statistics Review, 1975–2008*, National Cancer Institute, Bethesda, MD, USA, 2011.
- [2] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.
- [3] E. Warner, D. B. Plewes, R. S. Shumak et al., “Comparison of breast magnetic resonance imaging, mammography, and ultrasound for surveillance of women at high risk for hereditary breast cancer,” *Journal of Clinical Oncology*, vol. 19, no. 15, pp. 3524–3531, 2001.
- [4] Z. Pang, D. Zhu, D. Chen, L. Li, and Y. Shao, “A computer-aided diagnosis system for dynamic contrast-enhanced MR images based on level set segmentation and relief feature selection,” *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 450531, 10 pages, 2015.
- [5] A. Gubern-Mérida, R. Martí, J. Melendez et al., “Automated localization of breast cancer in DCE-MRI,” *Medical Image Analysis*, vol. 20, no. 1, pp. 265–274, 2015.
- [6] S. Thakran, S. Chatterjee, M. Singhal, R. K. Gupta, and A. Singh, “Automatic outer and inner breast tissue segmentation using multi-parametric MRI images of breast tumor patients,” *PLoS One*, vol. 13, no. 1, Article ID e0190348, 2018.
- [7] S. Wu, S. P. Weinstein, E. F. Conant, M. D. Schnall, and D. Kontos, “Automated chest wall line detection for whole-

- breast segmentation in sagittal breast MR images,” *Medical Physics*, vol. 40, no. 4, Article ID 042301, 2013.
- [8] L. Jiang, X. Hu, Q. Xiao, Y. Gu, and Q. Li, “Fully automated segmentation of whole breast using dynamic programming in dynamic contrast enhanced MR images,” *Medical Physics*, vol. 44, no. 6, pp. 2400–2414, 2017.
- [9] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2014.
- [10] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: a deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [11] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, pp. 234–241, Munich, Germany, October 2015.
- [12] X. Xu, L. Fu, Y. Chen et al., “Breast region segmentation using convolutional neural network in dynamic contrast enhanced MRI,” in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 750–753, Honolulu, HI, USA, July 2018.
- [13] M. Adoui, S. A. Mahmoudi, M. A. Larhman, and M. Benjelloun, “MRI breast tumor segmentation using different encoder and decoder CNN architectures,” *Computers*, vol. 8, no. 3, p. 52, 2019.
- [14] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: a nested U-net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2018, ML-CDS 2018: Lecture Notes in Computer Science*, et al. Vol. 11045, Springer Cham, Berlin, Germany, 2018.
- [15] C. Gallego-Ortiz and A. L. Martel, “Improving the accuracy of computer-aided diagnosis for breast MR imaging by differentiating between mass and nonmass lesions,” *Radiology*, vol. 278, no. 3, pp. 679–688, 2015.
- [16] N. Karssemeijer and G. M. te Brake, “Detection of stellate distortions in mammograms,” *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 611–619, 1996.
- [17] G. Kom, A. Tiedeu, and M. Kom, “Automated detection of masses in mammograms by local adaptive thresholding,” *Computers in Biology and Medicine*, vol. 37, no. 1, pp. 37–48, 2007.
- [18] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby, “A concentric morphology model for the detection of masses in mammography,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 6, pp. 880–889, 2007.
- [19] Y.-H. Huang, Y.-C. Chang, C.-S. Huang, J.-H. Chen, and R.-F. Chang, “Computerized breast mass detection using multi-scale hessian-based analysis for dynamic contrast-enhanced MRI,” *Journal of Digital Imaging*, vol. 27, no. 5, pp. 649–660, 2014.
- [20] W. Hongyu, F. Jun, B. Qirong et al., “Breast mass detection in digital mammogram based on gestalt psychology,” *Journal of Healthcare Engineering*, vol. 2018, Article ID 4015613, 13 pages, 2018.
- [21] W. Zhu, C. Liu, W. Fan, and X. Xie, “Deeplung: deep 3d dual path nets for automated pulmonary nodule detection and classification,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 673–681, Lake Tahoe, NV, USA, March 2018.
- [22] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, “Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.
- [23] D. Selvathi and A. A. Poornila, “Deep Learning Techniques for Breast Cancer Detection Using Medical Image Analysis,” *Biologically Rationalized Computing Techniques for Image Processing Applications*, Springer, Berlin, Germany, pp. 159–186, 2018.
- [24] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [25] D. Ribli, H. Anna, Z. Unger, P. Pollner, and I. Csabai, “Detecting and classifying lesions in mammograms with deep learning,” *Scientific Reports*, vol. 8, no. 1, 2017.
- [26] M. Fan, Y. Li, S. Zheng, W. Peng, W. Tang, and L. Li, “Computer-aided detection of mass in digital breast tomosynthesis using a faster region-based convolutional neural network,” *Methods*, vol. 166, pp. 103–111, 2019.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Las Vegas, NV, USA, June 2016.
- [28] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in *Proceedings of the 2017 IEEE international conference on computer vision (ICCV)*, IEEE, Venice, Italy, October 2017.
- [29] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, pp. 297–302, 1912.
- [30] P. Jaccard, “The distribution of the flora in the alpine Zone.1,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [31] J. K. Udupa, V. R. Leblanc, Y. Zhuge et al., “A framework for evaluating image segmentation algorithms,” *Computerized Medical Imaging and Graphics*, vol. 30, no. 2, pp. 75–87, 2006.

Research Article

Interactive Echocardiography Translation Using Few-Shot GAN Transfer Learning

Long Teng ¹, ZhongLiang Fu,¹ Qian Ma,² Yu Yao,¹ Bing Zhang,² Kai Zhu,¹ and Ping Li³

¹Chengdu Institute of Computer Application, University of Chinese Academy of Sciences, Beijing, China

²Sichuan University of Media and Communications, Chengdu, China

³West China Hospital, Sichuan University, Chengdu, China

Correspondence should be addressed to Long Teng; 52867567@qq.com

Received 9 July 2019; Revised 19 January 2020; Accepted 17 February 2020; Published 19 March 2020

Guest Editor: Andrea Duggento

Copyright © 2020 Long Teng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Interactive echocardiography translation is an efficient educational function to master cardiac anatomy. It strengthens the student's understanding by pixel-level translation between echocardiography and theoretically sketch images. Previous research studies split it into two aspects of image segmentation and synthesis. This split makes it hard to achieve pixel-level corresponding translation. Besides, it is also challenging to leverage deep-learning-based methods in each phase where a handful of annotations are available. **Methods.** To address interactive translation with limited annotations, we present a two-step transfer learning approach. Firstly, we train two independent parent networks, the ultrasound to sketch (U2S) parent network and the sketch to ultrasound (S2U) parent network. U2S translation is similar to a segmentation task with sector boundary inference. Therefore, the U2S parent network is trained with the U-Net network on the public segmentation dataset of VOC2012. S2U aims at recovering ultrasound texture. So, the S2U parent network is decoder networks that generate ultrasound data from random input. After pretraining the parent networks, an encoder network is attached to the S2U parent network to translate ultrasound images into sketch images. We jointly transfer learning U2S and S2U within the CGAN framework. Results and conclusion. Quantitative and qualitative contrast from 1-shot, 5-shot, and 10-shot transfer learning show the effectiveness of the proposed algorithm. The interactive translation is achieved with few-shot transfer learning. Thus, the development of new applications from scratch is accelerated. Our few-shot transfer learning has great potential in the biomedical computer-aided image translation field, where annotation data are extremely precious.

1. Background

Echocardiography education has dramatically helped students to master cardiac structure assessment by combining cardiac ultrasound images with simulators. However, a more efficient method of interactive translation between ultrasound images and theoretically sketch images is still lacking. This causes the image processing difficulties in our case: echocardiography is characterized by the deformable appearance and poor spatial resolution, while limited annotations are available, building obstacles to achieve good performance as well as leverage state-of-the-art deep learning methods.

U2S and S2U are often investigated in different approaches. U2S is often specified in the segmentation task. It is addressed with the following methods: Level set (LS) [1] segmentation,

Deformable templates [2, 3], Active shape models (ASM) [4, 5], Active contour methods, Active appearance models (AAM), Bottom-up approaches, and Database-guided (DB-guided) segmentation. LS and deformable templates present some drawbacks regarding the prior knowledge included in the optimization function. Active contour methods inspire the development of level set (LS) methods. ASM- and DB-guided approaches require a large number of annotated training images [6]. Bottom-up approaches are sensitive to initial conditions and lack of robustness. Additionally, none of those approaches are used to infer the sector boundary, which is essential for comprehension during education.

S2U typically models the tissue response as a collection of point scattering centers [7]. Different amplitudes are assigned to scatter from the blood pool or muscle. However,

due to ignoring surrounding conditions like papillary muscles, clutter noise, and local intensity variations, the genuineness of the synthetic ultrasound images is still unsatisfactory. Some improvements in combining ultrasound recording as a template to synthetic realistic speckle textures are proposed to address the above issue [8, 9]. However, those approaches unavoidably introduced unrealistic warping in simulated speckle texture.

GAN-based translation approach recently shows its potential in generative applications [10]. Structure [11] and texture [12, 13] generation are explored in different applications. While giving an outstanding performance, the GAN approach requires sufficient annotation, which is time-consuming and expensive for biomedical applications.

In this paper, we design a GAN-based transfer learning framework to interactively translate ultrasound images into sketch images (U2S translation) and sketch images into ultrasound images (S2U translation) with a handful of annotations. Figure 1 shows the example results of final U2S translation and S2U translation.

2. Methods

Our approach of interactive translation consists of two steps: pretrain U2S parent network and S2U parent network and train the two networks together with end-to-end transfer learning.

Transfer learning is used for fast adaption and avoiding overfitting since we got only a handful of annotations. In our case, parent networks are carefully designed and pretrained with supervised and unsupervised learning. GAN-based few-shot transfer learning is then designed to fine-tuning the final result.

The proposed U2S network (Figure 2) contains a parent network that follows the U-net [14] architecture. In this paper, the U-net structure contains 10 block layers. The first five blocks are convolutional downsampling networks. Kernel size here is 3, the stride is 2, and padding is 1. Each layer is followed by a batch norm layer and a relu layer. Correspondingly, the last five layers are deconvolutional upsampling networks. Its kernel size is 4, the stride is 2, and padding is 1. The batch norm and relu layer are also adopted. Skip-connection is realized by a concatenate layer between the symmetrical layers. U2S parent network is pretrained on VOC2012 dataset [15]. During the pretraining process, the loss function is class-balanced cross-entropy.

When U2S parent network is ready, we would then transfer the U2S Parent Network into sketch translation. The Conditional Generative Adversarial Network (CGAN) [16] framework is chosen here during transfer learning to infer sector boundary. Now, the U2S Parent Network is regarded as the generation network part of CGAN. It translates ultrasound images into sketch images. The CGAN framework could intuitively generate sketch images with sector boundaries. Also, we add L1 loss as an optional criterion.

$$L_S = -E[\log[D_S(S, U)]] - E[\log[1 - D_S(G_S(U), U)]] + L_{l1}. \quad (1)$$

In equation (1), D_S is the discriminator. It contains 5 block layers. Block layers contain convolution, batch normalization, and relu layers. D_S determines whether the input image is translated data or ground truth. S represents the ground truth sketch image. U represents ground truth ultrasound image. G_S is the generator (initialized with U2S Parent Network). It translates the ultrasound image into a sketch image.

S2U recovers the ultrasound texture from the sketch. Sketch image contains only the structure and no texture information at all. We first extract and maintain texture within the parent network and then synthesis texture on the specific sketch.

As shown in Figure 3, the S2U Parent Network is the decoder network. Our approach trains GAN to generate an ultrasound image on the condition of random input. In this way, as the generator part of GAN, the S2U Parent Network learns the ultrasound texture from training dataset. The S2U Parent Network consists of 4 block layers. The first 3 blocks contain a deconvolution layer, a batch normalization layer, and a relu layer. The last block contains a deconvolution layer and a tanh layer.

The S2U Parent Network training phase is shown in Figure 4. The generator and discriminator loss graphs are listed in the second row. The result of S2U Parent Network is illustrated in the first row. The generator and discriminator play against each other. As a result, the generator learns a growing quality of ultrasound textures.

When S2U Parent Network is ready, we could move forward to S2U transfer learning. Till now, our S2U Parent Network still has two flaws. Firstly, it cannot generate an ultrasound image on the condition of sketch input, not even pixel-level translation. Secondly, unexpected twist and image blur occur in Ultrasound Parent Network.

Aiming at making up for those two flaws, we further reform the network into S2U architecture that is shown in Figure 5. Pretrained S2U Parent Network is the dark blue part. An encoder network marked in light blue is connected to S2U Parent Network. This connection enables generation from a sketch to ultrasound image, other than from random initialization. In fact, the encoder network turns sketch image into the subset of random input. Thus, transfer learning learns the pixel-wise corresponding translation between sketch and ultrasound images. Besides, perceptual loss [17] and total variation loss are attached to the loss function. We try to maximize the fidelity of spatial resolution by minimizing GAN loss and perceptual loss. The loss function is shown in

$$L_U = -E[\log[D_U(U, S)]] - E[\log[1 - D_U(G_U(S), S)]] + \lambda_1 L_{\text{pct}} + \lambda_2 L_{\text{TV}} + \lambda_3 L_{l1}. \quad (2)$$

Intuitively, the loss function of S2U is similar to equation (1). D_U is discriminator. It determines whether the input image is synthesized by the network, or comes from the ground truth. D_U has 5 block layers and is shown in Figure 5. U represents ultrasound ground truth. G_U is the generator

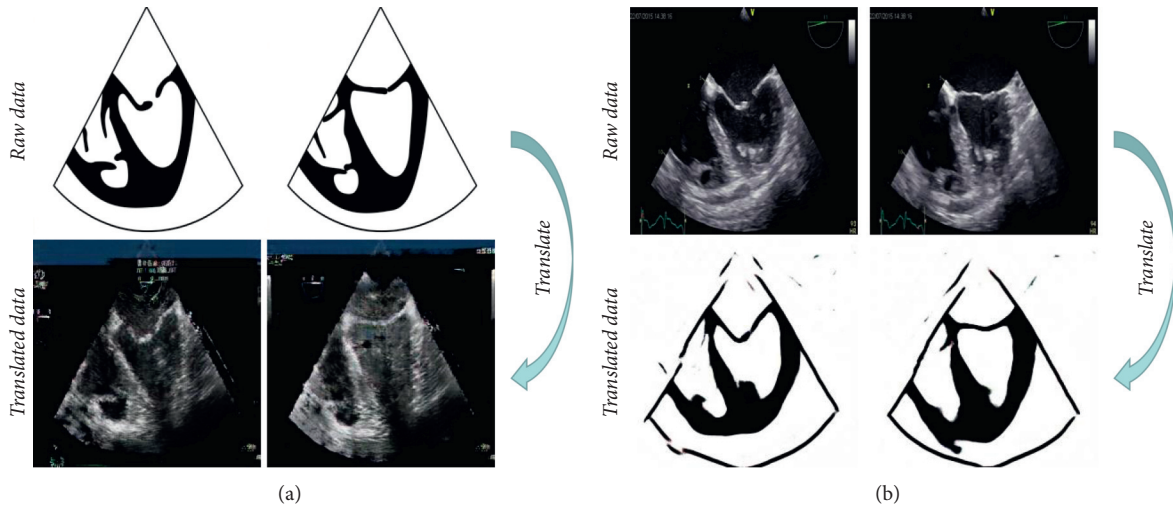


FIGURE 1: Example of interactive translation. The first two columns illustrate (a) S2U translation and the last two columns illustrate (b) U2S translation.

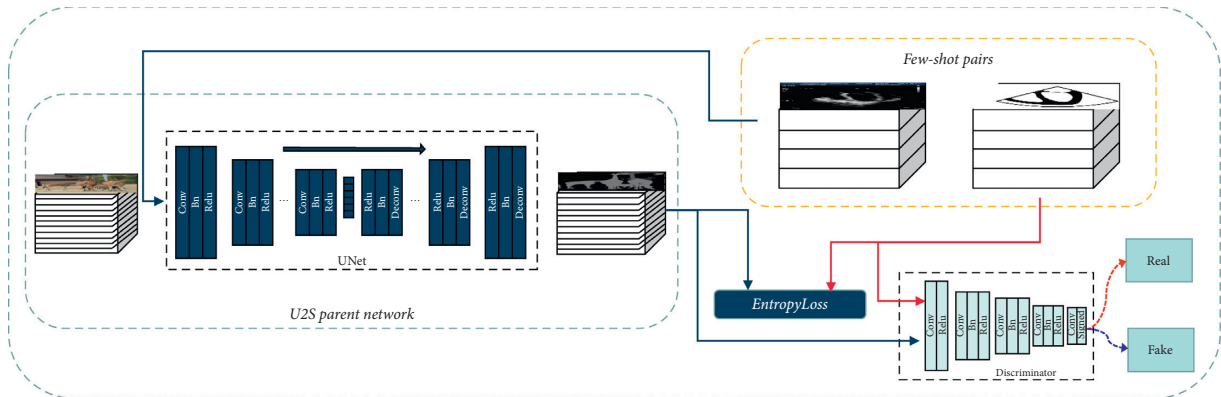


FIGURE 2: U2S network.

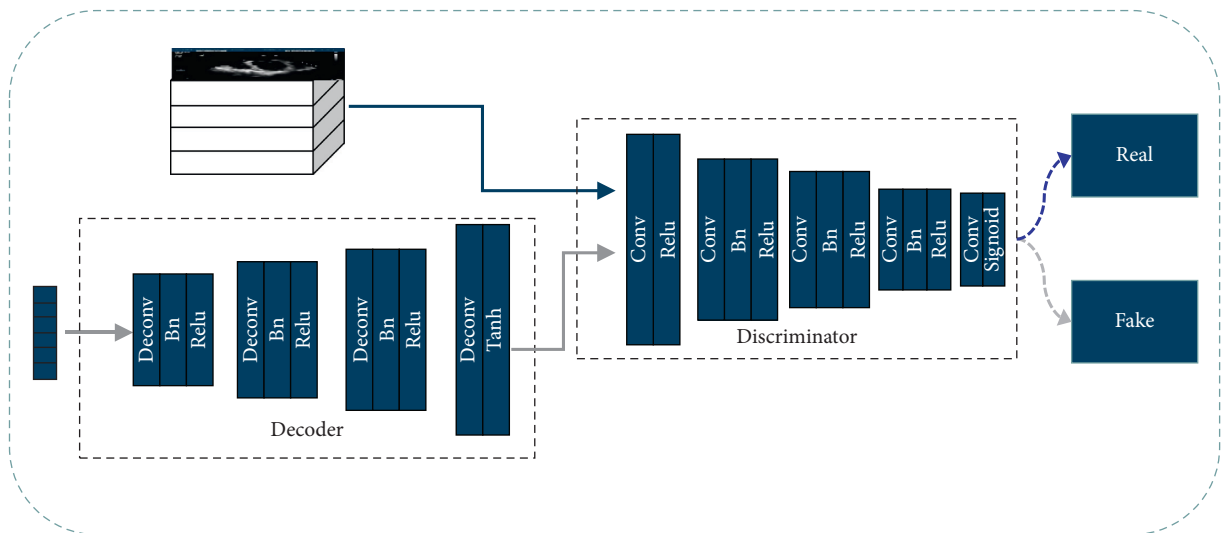


FIGURE 3: S2U Parent Network.

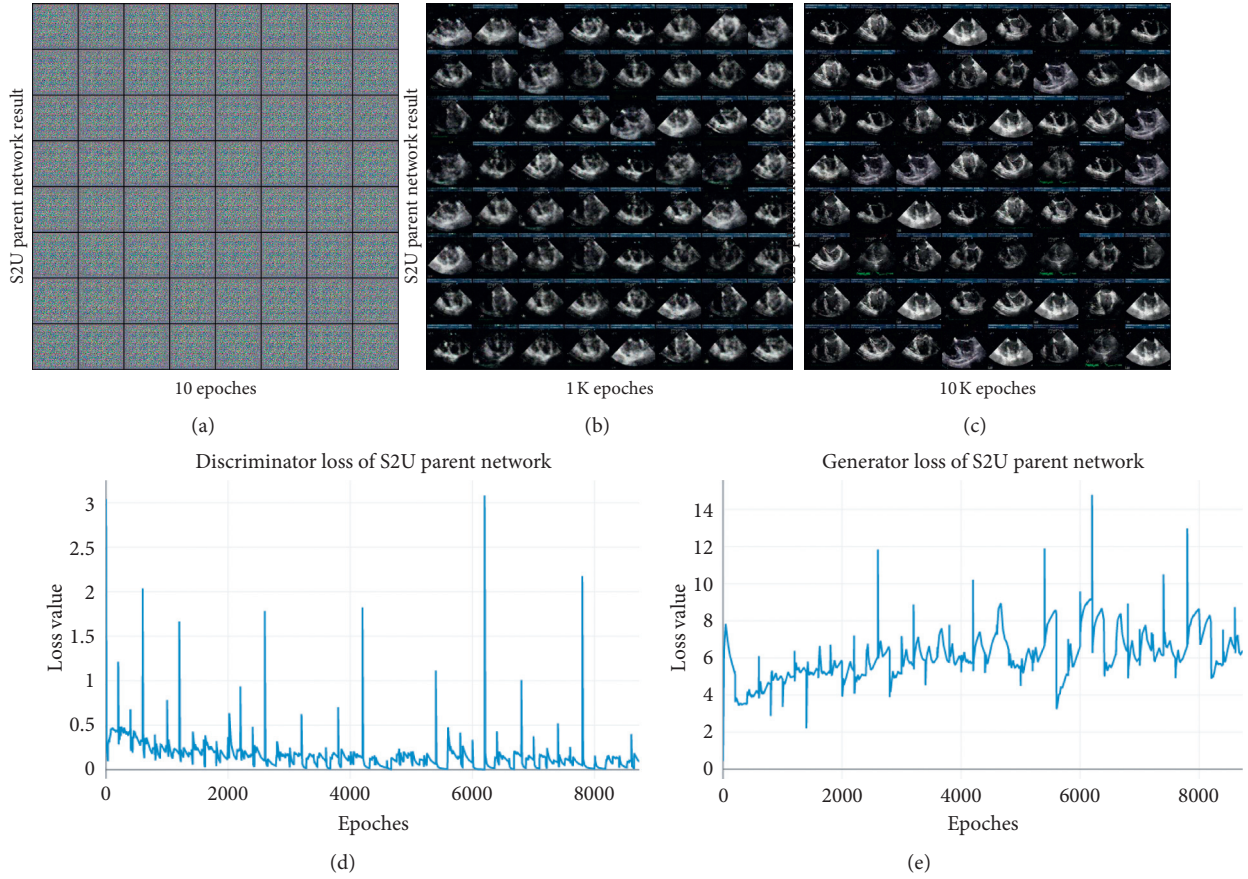


FIGURE 4: S2U training phase.

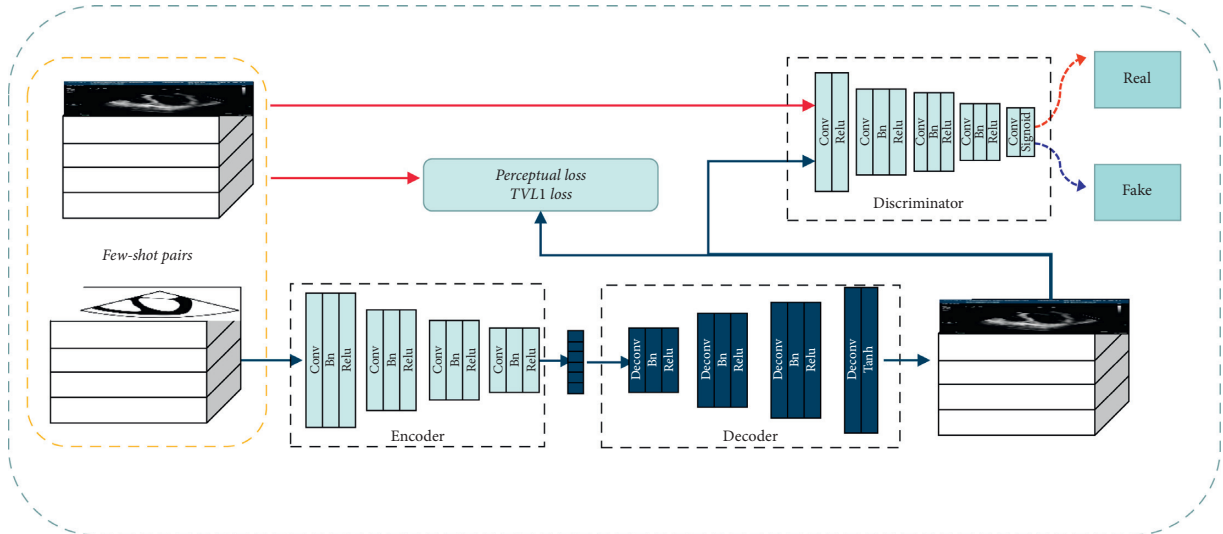


FIGURE 5: S2U network.

(the combination of navy encoder and decoder). It translates a sketch image into an ultrasound image.

L_{Pcpt} is the perceptual loss between ground truth ultrasound image and generated ultrasound image. The perceptual loss here is calculated with the feature maps of VGG16 network, which are more invariant to changes in

pixel space [18]. L_{TV} is the L1 smoothness of generated image. $\lambda_1, \lambda_2, \lambda_3$ in this paper are $6e-3, 2e-8,$ and $1,$ which could be further optimized.

As is mentioned above, loss function in U2S and that in S2U are similar to each other. Both of them are trained under CGAN framework. Furthermore, they share the same input

pairs. In Figures 2 and 4, we emphasized this similarity by marking the yellow dash blocks.

Therefore, we integrate U2S and S2U for interactive translation.

$$L_{\text{total}} = L_S + L_U. \quad (3)$$

During transfer learning, the S2U network is trained with TVL1 loss, perceptual loss, L1 loss, and CGAN loss to maintain ultrasound texture. After transfer learning for both two networks, each network splits into the following interactive application (Figure 6 shows our applications).

2.1. Interactive U2S Translation. In some scenarios, the student would carefully study the static picture that captured in dynamic echo video. During this interaction, the local area should be amplified and translated into a sketch at a breakneck speed. Otherwise, the interaction would get stuck and result in a terrible experience.

In this paper, we complete sketch translation at the start of the interaction. Region of interest (ROI) is then selected and amplified to the size of the original image. Notice that, sketch image is the black-and-white image, cubic interpolation is chosen for amplification. The cubic interpolation is efficient and enough for identification.

2.2. Video U2S Translation. During training, automatic U2S translation would greatly help students to comprehend. Here, we split the U2S Network part from the whole networks. U2S network inputs ultrasound images and outputs sketch images. So every frame is translated into sketch images. We process frame-by-frame, converting all frames into a video. This translated sketch video is dynamically contrast to echocardiography to illustrate structural information.

2.3. Interactive S2U Translation. If the student draws a sketch, which outlines the cardiac structure, how the sketch corresponds to the clinical ultrasound image? This interaction could be thought-provoking and, in turn, help for comprehension.

We extract the decoder network in the S2U Parent Network and turn it into an S2U network with an encoder network. S2U inputs sketch and outputs an ultrasound image. It strictly generates output with an appropriate ultrasound texture. So, after students complete their sketch in the drawing board, the sketch image could interactively be translated into an ultrasound image.

3. Results

In this section, we compare the method of U2S translation and S2U translation with 1-shot, 5-shot, and 10-shot transfer learning. Firstly, the performance is analyzed through the visual comparison and the visualization of transfer learning process. Then, the performance is investigated through numerical comparison. In numerical comparison, each experiment is summarized through 45 pairs of annotations.

Besides, we supplement S2U translation performance with and without perceptual loss and TVL1 loss during numerical comparison.

3.1. Dataset. Two datasets are used in this paper, VOC2012 and echocardiography dataset. VOC2012 is an open access segmentation dataset used for the pretraining of the U2S parent network. The echocardiography dataset is collected in the hospital under the guidance of doctors. It contains 5152 four-chamber view echocardiographs with no annotation, and 55 pairs of annotated four-chamber view echocardiographs (in this paper, we use 10 pairs for training and left 45 pairs of the annotated images for validation). Those annotations are made by the teamwork of doctors and art teachers. Images are fully annotated with the chamber (atrial and ventricular), sector boundary, and myocardial. Sensitive patient information is manually removed.

3.2. Visual Comparison. A pair of validation images is chosen to analyze the performance of our proposed network. As shown in Figure 7, the left column is a pair of ground truth. The first row shows S2U results from 1-shot, 5-shot, and 10-shot. The contrast between myocardium and chamber is getting obvious while inputting more transfer learning data. Also, the image resolution is getting better, which makes the myocardium more realistic.

Compared with the real ultrasound images, the S2U results' texture is more similar to the training data. The blue bar and some comments from training data are synthesized on S2U results. In the second row, 1-shot, 5-shot, and 10-shot results of U2S are shown in order. The shape of the U2S result is getting similar to the ground truth. The sector boundary of U2S is also getting reasonable with more training data.

3.3. Transfer Learning Process. The performance of transfer learning process is investigated in two aspects, the loss function value and the corresponding performance during training. The loss function value of S2U and U2S is a representative, shown with 5-shot in Figure 8.

As is shown in Figure 8, the first row is the first three terms of L_U , and the second row is the terms of L_S . The discriminator and generator loss of S2U and U2S are the first two images in the first and second rows. In both S2U and U2S, the generator and discriminator contest against each other, while the perceptual loss of S2U and the L1 loss of U2S keep decreasing. The adversarial loss function and extra loss function work together to fine-tune the final result. Figure 9 shows the performance on testing data.

In Figure 9, the Intersection over Union (IOU) and peak signal to noise ratio (PSNR) result are representatively illustrated in 1-shot, 5-shot, and 10-shot. As a result of the proposed loss function, S2U and U2S achieve improving performance during training. Specifically, the more the training samples, the better the performance achieved. 10-shot transfer learning achieves better performance than 5-shot, while 5-shot achieves better performance than 1-shot.

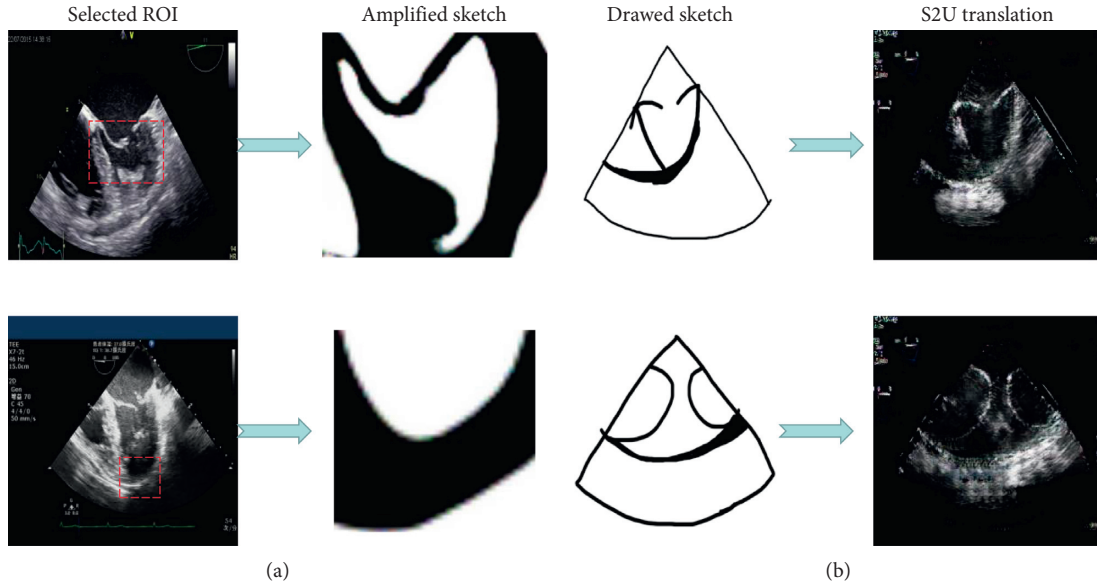


FIGURE 6: Interactive translation applications. First two columns illustrate interactive (a) U2S translation and last two columns illustrate interactive (b) S2U translation.

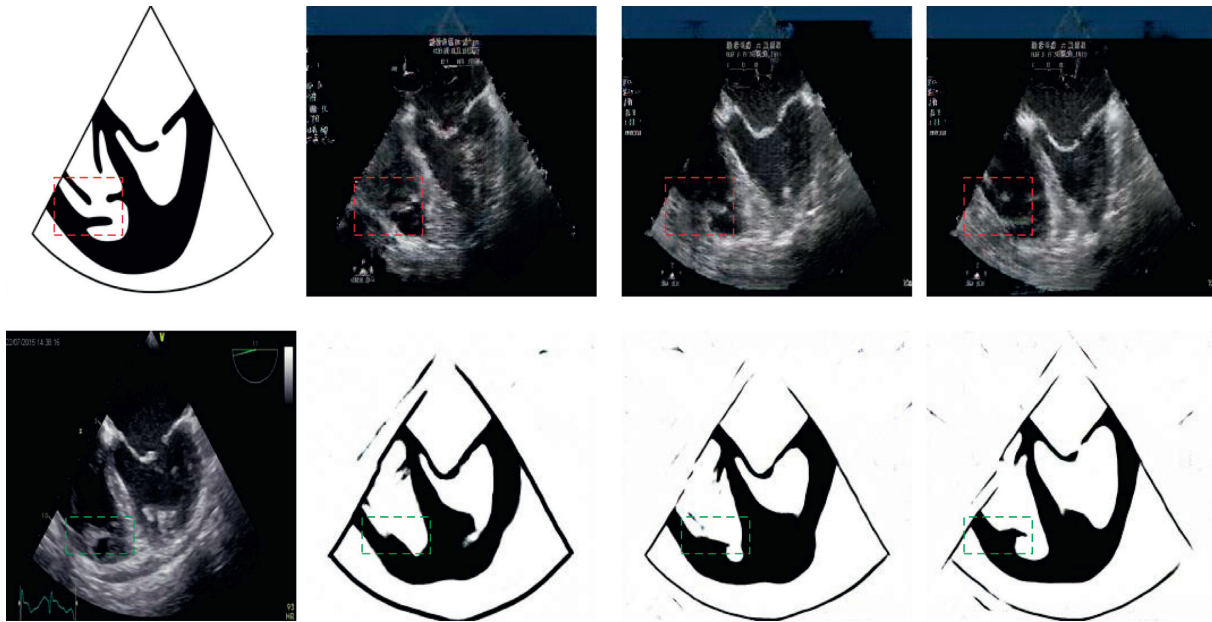


FIGURE 7: Example performance of few-shot transfer learning. The first row shows S2U of 1-shot, 5-shot, and 10-shot; the second row shows U2S of 1-shot, 5-shot, and 10-shot. The left column is a pair of ground truth.

3.4. Numerical Comparison. In U2S translation, we adopt the medical image segmentation index of dice loss, volumetric overlap error (VOE), and intersection over union (IOU). In S2U translation, we use peak signal to noise ratio (PSNR) and structural similarity index (SSIM) to evaluate our performance.

The convincing result below (Tables 1 and 2) shows the effectiveness of proposed few-shot transfer learning with

1-shot, 5-shot, and 10-shot. In Table 1, the gradual increase of training samples leads to better performance of the index. In Table 2, the indexes of PSNR and SSIM are compared with and without extra loss function.

As is shown in Table 1, few-shot learning has led to acceptable results in all of the indexes. It enables us to present the initial version of the U2S function while lacking annotations.

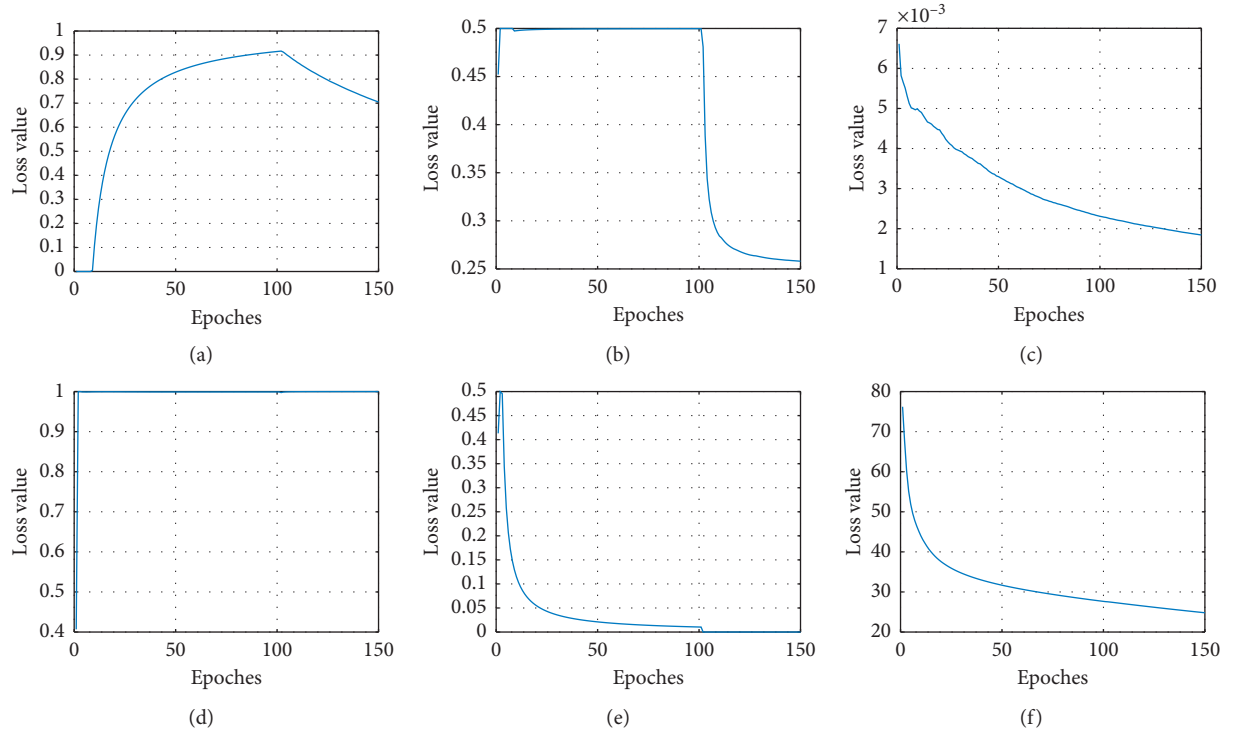


FIGURE 8: Loss function value during transfer learning. The S2U and U2S loss value during transfer learning. (a) S2U generator loss. (b) S2U discriminator loss. (c) S2U generator perceptual loss. (d) U2S generator loss. (e) U2S discriminator loss. (f) S2U generator L1 loss.

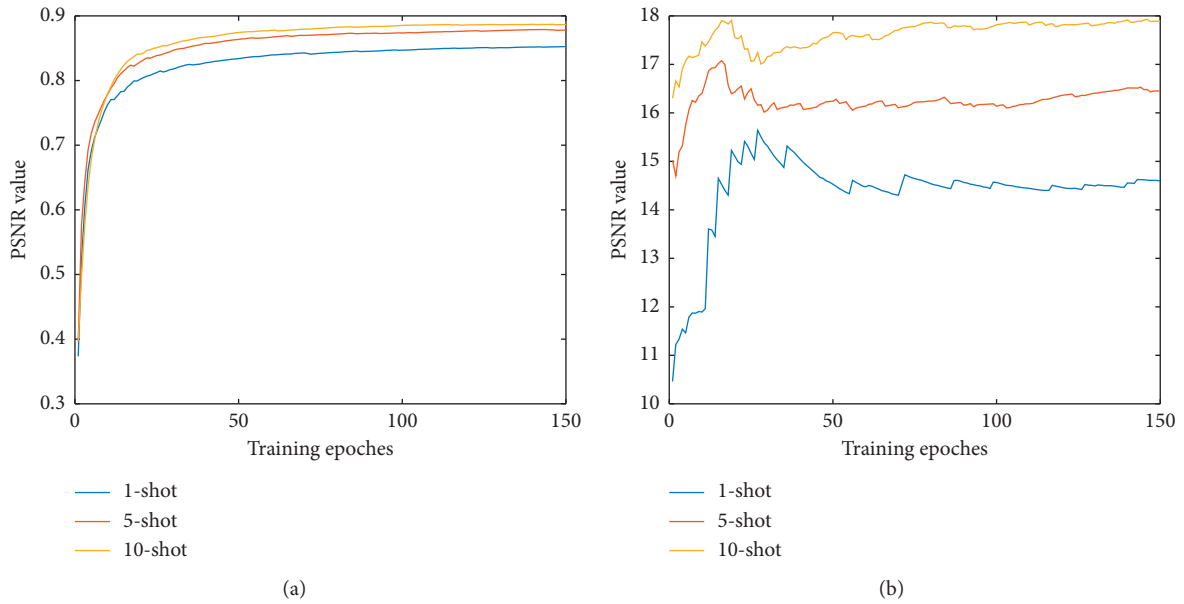


FIGURE 9: U2S and S2U performance during transfer learning. The IOU index of U2S and the PSNR index of S2U during transfer learning. (a) IOU—few-shot transfer learning. (b) PSNR—few-shot transfer learning.

TABLE 1: U2S translation accuracies on our dataset.

U2S	DICE	VOE	IOU
1-shot	0.902	0.001	0.852
5-shot	0.913	0.018	0.872
10-shot	0.921	-0.009	0.887

TABLE 2: S2U translation accuracies on our dataset.

S2U	PSNR	PSNR (No. perceptual and TVL1 loss)	SSIM	SSIM (No. perceptual and TVL1 loss)
1-shot	14.614	14.898	0.417	0.406
5-shot	16.180	16.074	0.499	0.433
10-shot	17.905	17.575	0.544	0.553

According to the result of Table 2, S2U that trained with the perceptual and TVL1 loss is generally better without those loss functions.

4. Conclusion

This paper proposed a few-shot GAN Transfer Learning for Interactive Echocardiography Translation. U2S Parent Network and S2U Parent Network are individually designed and pretrained beforehand. Then, they are assembled together for transfer learning. This joint transfer learning transfers prior knowledge into target networks. Qualitative analysis of visual comparison and visualization of the transfer learning process, quantitative analysis of numerical index shows the effectiveness of the proposed method.

The proposed method has two advantages over previous researches. Firstly, it simultaneously achieves interactive translation between ultrasound and sketch images with few-shot annotations, enabling a new educational interactive function before getting enough annotation. Secondly, it is also promising in further improvement with more training data and is promising in other related biomedical applications.

Data Availability

Part of our dataset used in the current study is available from the corresponding author on a reasonable request. Our code is open source at: <https://github.com/tlok666/Interactive-Echocardiography-Translation-with-Few-Shot-GAN>.

Ethical Approval

This study was approved by the Medical Ethics Committee of the West China Hospital, Sichuan University, and written informed consent was obtained from each participant.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Long Teng contributed equally to this work. Long Teng, ZhongLiang Fu, and Kai Zhu designed the research. Long Teng completes all the code and paper material. Qian Ma, Bing Zhang, and Ping Li prepared the dataset. Yu Yao is responsible for the application of the proposed algorithm.

Acknowledgments

The authors would like to thank to the doctors in the Department of Anesthesiology, West China Hospital, Sichuan

University, for their helpful contribution with collecting and validating the data. This study was supported by Sichuan Province's New Generation of Artificial Intelligence Major Special Project (Grant no. 2018GZDZX0036).

References

- [1] S. Mazaheri, P. S. B. Sulaiman, R. Wirza et al., "Echocardiography image segmentation: a survey," in *Proceedings of the 2013 International Conference on Advanced Computer Science Applications and Technologies*, pp. 327–332, Kuching, Malaysia, December 2013.
- [2] K. Chauhan and R. Chauhan, "Boundary detection of echocardiographic images during mitral regurgitation," in *Recent Advances in Computer Vision*, pp. 281–303, Springer, Berlin, Germany, 2019.
- [3] G. Veni, M. Moradi, H. Bulu, G. Narayan, and T. Syeda-Mahmood, "Echocardiography segmentation based on a shape-guided deformable model driven by a fully convolutional network prior," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 898–902, Washington, DC, USA, April 2018.
- [4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
- [5] S. C. Mitchell, B. P. F. Lelieveldt, R. J. Van Der Geest, H. G. Bosch, J. H. C. Reiver, and M. Sonka, "Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac mr images," *IEEE Transactions on Medical Imaging*, vol. 20, no. 5, pp. 415–423, 2001.
- [6] S. Leclerc, E. Smistad, J. Pedrosa et al., "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE Transactions on Medical Imaging*, vol. 20, 2019.
- [7] M. Alessandrini, M. De Craene, O. Bernard et al., "A pipeline for the generation of realistic 3d synthetic echocardiographic sequences: methodology and open-access database," *IEEE Transactions on Medical Imaging*, vol. 34, no. 7, pp. 1436–1451, 2015.
- [8] A. Prakosa, M. Sermesant, H. Delingette et al., "Generation of synthetic but visually realistic time series of cardiac images combining a biophysical model and clinical images," *IEEE Transactions on Medical Imaging*, vol. 32, no. 1, pp. 99–109, 2012.
- [9] S. Marchesseau, H. Delingette, M. Sermesant et al., "Preliminary specificity study of the Bestel-Clément-Sorine electromechanical model of the heart using parameter calibration from medical images," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 20, pp. 259–271, 2013.
- [10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, Honolulu, HI, USA, July 2017.

- [11] B. Chang, Q. Zhang, S. Pan, and L. Meng, "Generating handwritten Chinese characters using cyclegan," in *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 199–207, Lake Tahoe, NV, USA, March 2018.
- [12] C. Chu, A. Zhmoginov, and M. Sandler, "Cyclegan, a master of steganography," 2017.
- [13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, Venice, Italy, October 2017.
- [14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Munich, Germany, October 2015.
- [15] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," in *Pattern Analysis, Statistical Modelling and Computational Learning*, Technical Report, Springer, Berlin, Germany, 2011.
- [16] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.
- [17] C. Ledig, L. Theis, F. Huszár et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, Honolulu, HI, USA, July 2017.
- [18] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2479–2486, Las Vegas, NV, USA, June 2016.

Research Article

Multicenter Computer-Aided Diagnosis for Lymph Nodes Using Unsupervised Domain-Adaptation Networks Based on Cross-Domain Confounding Representations

RuoXi Qin ¹, Huike Zhang ², LingYun Jiang¹, Kai Qiao ¹, Jinjin Hai ¹, Jian Chen ¹, Junling Xu², Dapeng Shi ² and Bin Yan ¹

¹PLA Strategy Support Force Information Engineering University, Zhengzhou 450001, China

²Department of Radiology, Henan Provincial People's Hospital, Zhengzhou 450002, China

Correspondence should be addressed to Bin Yan; ybspace@hotmail.com

Received 9 September 2019; Revised 14 December 2019; Accepted 26 December 2019; Published 24 January 2020

Guest Editor: Andrea Duggento

Copyright © 2020 RuoXi Qin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To achieve the robust high-performance computer-aided diagnosis systems for lymph nodes, CT images may be typically collected from multicenter data, which cause the isolated performance of the model based on different data source centers. The variability adaptation problem of lymph node data which is related to the problem of domain adaptation in deep learning differs from the general domain adaptation problem because of the typically larger CT image size and more complex data distributions. Therefore, domain adaptation for this problem needs to consider the shared feature representation and even the conditioning information of each domain so that the adaptation network can capture significant discriminative representations in a domain-invariant space. This paper extracts domain-invariant features based on a cross-domain confounding representation and proposes a cycle-consistency learning framework to encourage the network to preserve class-conditioning information through cross-domain image translations. Compared with the performance of different domain adaptation methods, the accurate rate of our method achieves at least 4.4% points higher under multicenter lymph node data. The pixel-level cross-domain image mapping and the semantic-level cycle consistency provided a stable confounding representation with class-conditioning information to achieve effective domain adaptation under complex feature distribution.

1. Introduction

Novel techniques for medical image analysis, computer-aided diagnosis algorithms, and related topics have been recently emerging. The research in these directions depends not only on the development of artificial intelligence methods but also on the promotion of big data techniques. In fact, deep learning algorithms for computer-aided diagnosis typically use realistic big medical datasets that are manually labeled. Problems pertaining to medical image analysis are complicated by the fact that clinical images of the same modality may differ if these images are generated by devices of different models, manufacturers, or imaging parameters. For big-data-driven medical diagnosis algorithms, a model trained only by one centrally acquired

dataset cannot typically generalize well, a shortcoming which limits the application scope of such model [1, 2]. However, collecting data with multiple models, parameters, or locations increase the cost of manual labeling, and balancing the numbers of data samples collected from each of these factors for one dataset is complicated leading to inflexibility in practical applications. Classification of medical imaging data under such variations is usually referred to as a cross-domain classification problem. Indeed, if there is no comprehensive dataset with a balanced quantity of samples for each factor, it is difficult to achieve optimal classification performance for these factors. For example, computerized tomography (CT) imaging shows clarity variations in some tissues due to differences in equipment or parameters. If we have enough labeled data samples for one of the CT imaging

devices, we can train a stable deep learning network. The CT images collected from another device may not be suitable by themselves for training a stable network if the high labeled data requirement is not met [3, 4]. In view of this, we can define the problem of unsupervised domain adaptation in deep learning as one dealing with how to use the labeled data from one domain to achieve the common recognition of two-domain images, especially when the images are unlabeled. This problem is properly handled by a multidomain robust classification algorithm, which integrates multicenter data with less manual labeling.

There are two types of methods in domain adaptation: data-centric methods and subspace-centric methods [5]. On the one hand, a data-centric method finds a unified transformation that maps data from source and target domains into a domain-invariant space so as to reduce the distributional disparities of data from the source and target domains, while retaining the data attributes of the original space. The features from the domain-invariant space are used to achieve the final classification on a different domain. On the other hand, subspace-centric methods reduce domain shifts by manipulating subspaces of two domains. For example, this manipulation can be done by establishing a linear map or by using Grassmann kernels [6] so that the subspace of each domain contributes to the formation of the final map. By comparing the two types of methods, we find that the data-centric methods require extracting features into a common implicit space for classification, while the subspace-centric method transforms the target subspace to the source subspace. Utilizing the multimodal characteristics of medical data while paying attention to the flexibility considerations in practical applications, we constructed a unified network for the classification of multicenter data.

In domain adaptation problems, the most commonly used methods are the adversarial training ones such as the gradient reversal layer (GRL) method or the maximum mean discrepancy (MMD) method. Ganin et al. [7] proposed computing adversarial losses in the embedded space using the gradient reversal layer (GRL). The goal of the adversarial training is to find the parameters of a domain classifier or discriminator that maximizes its classification accuracy while also setting the parameters of a generator to maximize the confusion of the domain classifier. The biggest problem with this domain classification loss is that a good domain-invariant feature should not let the discriminator know which domain it came from. Such a domain discriminator based on binary classification tends to promote the network to extract domain-specific features, which might be irrelevant to both domains. The maximum mean discrepancy (MMD) [8] is used in many studies to depict the distributional disparity between feature spaces from different domains. Tzeng et al. [9] and Rozantsev et al. [10] use MMD to measure the distance between two feature spaces and to map the source and the target spaces to a unified feature space by minimizing this distance. Furthermore, Sun and Saenko [11] use the CORAL loss to align the mean value and the variance between two domains to achieve correlation alignment. These methods, striving to

characterize feature loss only, consider the distance in the feature space but do not take into account other correlative representations between the feature spaces of the source and the target domains. When the prior class-conditioning information of the source and target domains differs greatly, the similarity of the feature spaces means the loss of the class-conditioning information of the target domain, and hence higher classification errors occur in the target domain. Furthermore, domain separation networks (DSNs) [12] use four extraction feature networks to disentangle the common and domain-specific features. Through the GRL for domain similarity constraints with soft subspace orthogonality loss for difference constraints, the obtained domain-invariant features still suffer from the problems caused by the above specific losses. The domain-specific and shared features of medical images are very complicated, indeed, because of the high resolution of these images.

The unsupervised image translation problem is similar to the domain adaptation problem. For a data-centric problem, whether it is an unsupervised domain-adaptation problem or a style translation problem, the main purpose is to find a domain-invariant representation across the two domains. In comparison with the methods that use specific constraint losses such as the MMD or CORAL methods, a more flexible implicit-loss scheme named cross-domain image mapping is proposed for unsupervised image style translation [13, 14], by which pixel-level domain-invariant features are obtained. The unsupervised image-to-image translation (UNIT) [15] is based on a shared-latent-space hypothesis. The UNIT cross-domain reconstruction mapping, based on generative adversarial networks (GANs) such as cycle GAN [16] and coupled GAN (CoGAN) [17], forces the network to extract common implicit spatial features. Combined with a pixel-space variational autoencoder (VAE) reconstruction loss [18], the UNIT loss function uses implicit loss constraints through image reconstruction and adversarial training to achieve more flexible feature extraction [19]. Following the approach of the domain separation networks (DSNs) [12], methods that utilize diverse image-to-image translation (DRIT) [20] and multimodal unsupervised image-to-image translation (MUNIT) [21] disentangle the common and domain-specific features and then perform cross-domain reconstruction. These methods add more image reconstruction loss to the domain-invariant constraints to obtain a more realistic reconstructed image. However, such a loss function limits feature extraction and mapping. Some studies have also applied the domain-adaptation method to image translation. The XGAN method [22] uses the GRL idea instead of the pixel-level VAE loss to achieve semantic cycle consistency of the shared feature representation allowing the network to realize flexible structural transformations.

The biggest difference between the domain-adaptation problem and the image translation problem is that the former problem is characterized by a weakening of image reconstruction quality and an enhancement of the class-conditioning information in the domain-invariant space. As mentioned above, the domain-invariant space in

the domain-adaptation problem should have the characteristics of being classifiable and domain invariant. The features extracted from the two domains can be reconstructed across domains, and a generated image cannot be distinguished as to which domain it originates from. The GTA [23] uses the joint generative-discriminative method to add a classification loss to the GAN loss [24] to constrain the class-conditioning information of the generated images. The pixel-level loss corresponding to two images is not used. So, the domain adaptation of the network does not depend on the fine image reconstruction. Moreover, the importance of the gradient of the classification loss for domain adaptation is verified in [23]. Because the target-domain image is unlabeled, GTA only uses the source label to back propagate the gradients. Such a method might not be well generalized for the classification in the target domain. After all, a classifier trained only with the source domain may have certain domain-specific features. To solve this problem, we use a pair of decoders to establish a cross-domain image mapping to map the labeled source domain images to the target domain. In order to realize this cross-domain mapping and make the network robust to shifts of the class-conditioning information between the source and target domains, we propose a new concept called the classification cycle consistency in the image generation process. The classification cycle consistency, which is similar to the cycle semantic loss in XGAN, can more effectively perceive changes in the class-conditioning information and provide a more flexible structural transformation while regularizing the ill-posed unpaired cross-domain mapping problem.

In another main contribution of this paper, a domain-adaptation network is built based on the cross-domain confounding representation to solve the multicenter problem of computer-aided diagnosis data. Inspired by the unsupervised image translation problem, a cross-domain confounding representation is used to implicitly force the network to extract domain-invariant features through cross-domain mapping. The network architecture consists of a pair of encoders and decoders, as well as a discriminator and a classifier. The pair of encoders has common parameters for feature extraction across the two domains, while the classifier achieves the final classification in the domain-invariant space. Because of the complex feature distribution of medical images and the requirement of class-conditioning information in the domain-invariant space, the classification cycle consistency is added to the loss function after the cross-domain mapping to help the network update the gradient with class-conditioning information. Figure 1 shows a conceptual diagram of the method employed in this paper. Compared to the previous methods on domain adaptation such as those of GRL, MMD, and GTA, our method can achieve a better and more stable performance for high-resolution medical images with complex feature distributions.

2. Materials and Methods

We introduce here our approach in three stages. In the first stage, we introduce the network architecture and the main

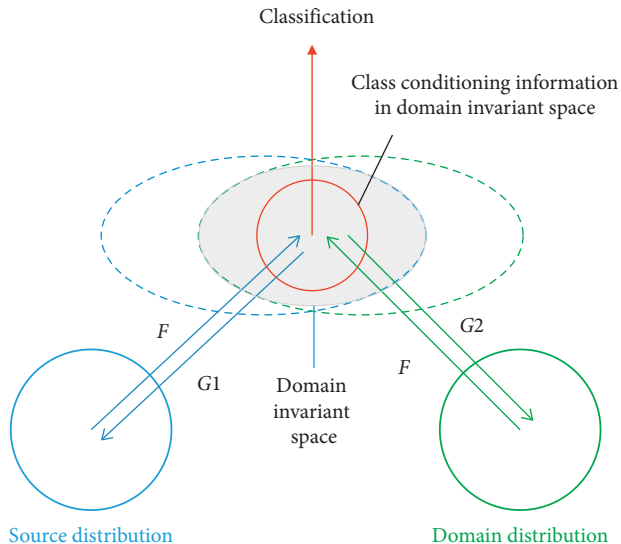


FIGURE 1: A graphical illustration of the proposed method: the cross-domain confounding representation is generated by constraining the cross-domain mapping reconstruction. The classification cycle consistency enables the network to perceive the significant discriminative representation in a domain-invariant space for final classification.

functions of each component. In the second stage, we combine the functions of all parts to describe the construction of the loss function. In the third stage, we introduce how the whole network uses loss function to achieve the main functions of each component and the verification of network.

2.1. Architecture of the Proposed Network. Our network architecture consists of pairs of decoders and discriminators as shown in Figure 2(a). The labeled data represent the source image, and it is shown in blue while the unlabeled data represents the target image, which is shown in green. We assume that pairs of images, $X1$ and $X2$, from different domains can be mapped into a domain-invariant space. The encoder network F is a fully-weighted shared network with six convolution layers whose inputs are images from two domains. After the encoder, we apply the classification network which has two fully connected layers to classify the domain-invariant features. The decoder G has six deconvolution layers to reconstruct an image through the domain-invariant features. The network function structure is mainly composed of two parts. One part extracts the cross-domain confounding representation through the loss of GAN, and the other part guarantees the consistency of the class condition information in the codec process through the classification cycle consistency. Figure 2(a) indicates the process of the extraction of crossing domain confounding representation under whole network. Figure 2(b) indicates the process of classification cycle consistency after image reconstruction through the source image phase. Features from the source domain with label information are decoded to generate an image of the corresponding

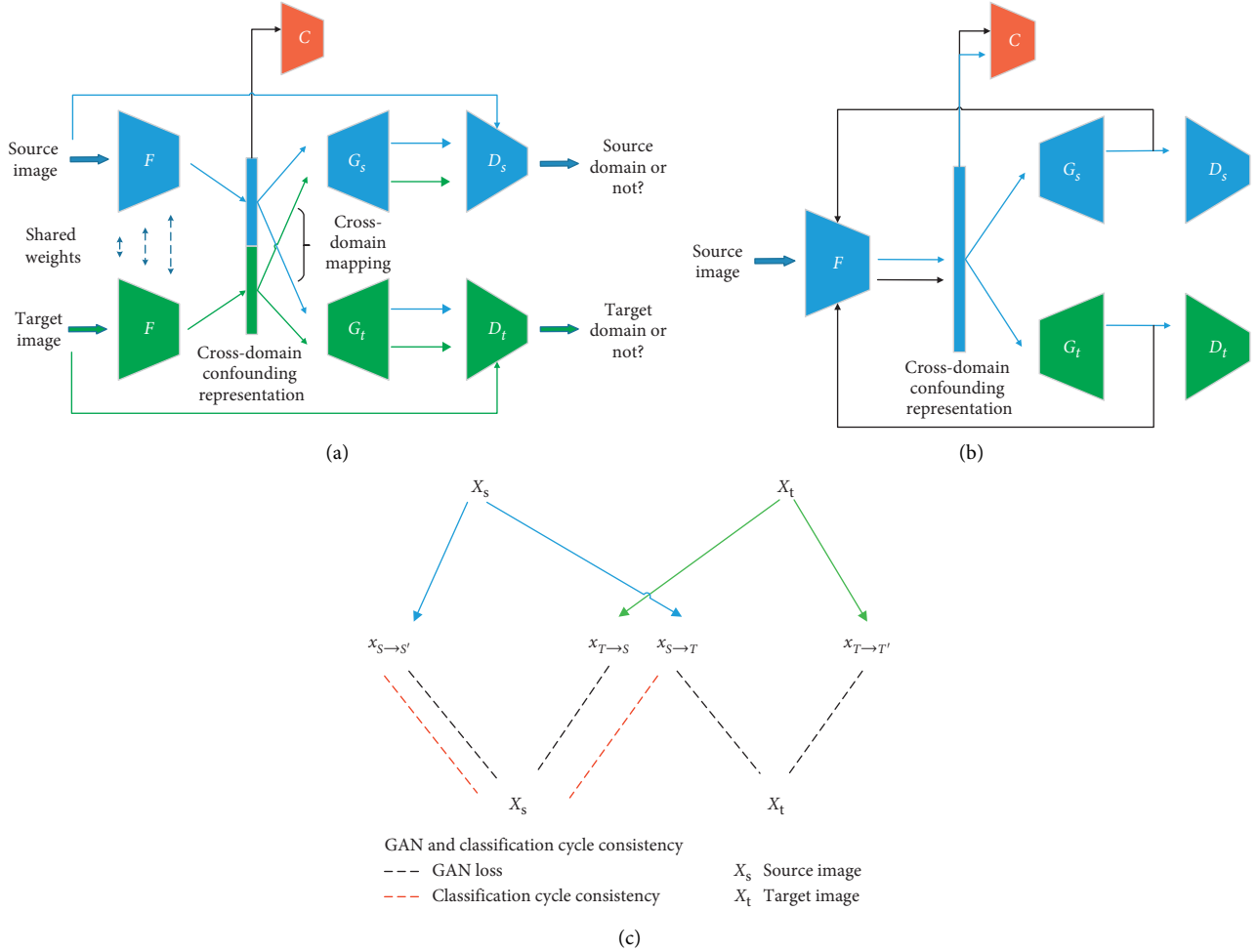


FIGURE 2: Illustration of the proposed network architecture. (a) Domain confounding representation through cross-domain mapping: the encoder F and the decoder G constitute the VAE architecture for unsupervised representation learning. The D module constitutes the GAN discriminator, while the C module constitutes a classifier. The encoder F uniformly encodes images from two domains. Paired decoders process different domain features, enabling cross-domain pixel-level image reconstruction and adversarial discrimination. (b) Classification cycle consistency: the reconstructed image based on source-domain features, as shown by the black line, will be constrained by classification cycle consistency through F and C . (c) Illustration of the loss overview.

domain. The generated image is then reencoded and reclassified. This update step introduces the gradient of the class-conditioning information into the decoder and encoder of the each domain and implements semantic-level consistency constraints. Figure 2(c) shows the GAN loss and classification cycle consistency overview. Compared with the joint generative discriminative method [23], the proposed method ensures the consistency of the classification information and the domain-invariant information in the encoding and decoding processes.

2.2. Loss Function Construction. Let the source and target distributions be $S(x, y)$ and $T(x, y)$, respectively, and S' and T' be the generate image, respectively. Also, let the source and target images be x_s and x_t . In order to achieve cross-domain confounding representation and classification cycle

consistency, we optimize the domain discriminator D , decoder G , classifier C , and encoder F as follows.

2.2.1. Domain Discriminator D . The purpose of the domain discriminator is to determine whether an input image belongs to the considered domain. The design goal is to minimize the domain discriminator loss, forcing the codec to generate images that are more in line with the domain features in the adversarial training. The application of the pixel-level adversarial losses follows the approaches of the DC-GAN [19] and the GTA [23]. To solve the problem that domain-invariant feature should not let the discriminator know which domain it came from, we apply two domain discriminators D in each domain so that the discriminator discriminates the domain other than the two domains. The loss function of D is shown in equations (1) and (2). The S

and T subscripts represent the discriminators of the source and target domains, respectively:

$$\begin{aligned}\mathcal{L}_D^S &= \mathbb{E}_{x_s \sim S} \max_{D_S} \log(D_S(x_s)) \\ &\quad + \log(1 - D_S(x_{S \rightarrow S'})) + \log(1 - D_S(x_{T \rightarrow S})), \\ \mathcal{L}_D^T &= \mathbb{E}_{x_t \sim \Gamma} \max_{D_T} \log(D_T(x_t)) \\ &\quad + \log(1 - D_T(x_{T \rightarrow T'})) + \log(1 - D_T(x_{S \rightarrow T})),\end{aligned}\tag{1}$$

$$\mathcal{L}_D = \mathcal{L}_D^S + \mathcal{L}_D^T.\tag{2}$$

2.2.2. Decoder G . The purpose of each decoder or generator G is to maximize the domain discriminator confusion by generating images that fool this domain discriminator as much as possible. This confusion constitutes a part of the adversarial loss. As the training progresses, the decoder G generates images with more domain characteristics. The classification cycle consistency ($C(F(x_{S \rightarrow S'}))$ and $C(F(x_{S \rightarrow T}))$) is introduced to update G . The generated images are reclassified through the encoder F and the discriminator D . By minimizing the classification loss, the decoding network captures more flexible class structure changes, retaining the class-conditioning information in the generated image. The loss function of G is given by equations (3) and (4). The S and T subscripts represent the discriminators of the source and target domains, respectively:

$$\begin{aligned}\mathcal{L}_G^S &= \mathbb{E}_{x_s \sim S} \min_{G_S} -\log(C(F(x_{S \rightarrow S'}))) \\ &\quad + \log(1 - D_S(x_{S \rightarrow S'})),\end{aligned}\tag{3}$$

$$\begin{aligned}\mathcal{L}_G^T &= \mathbb{E}_{x_t \sim \Gamma} \min_{G_T} -\log(C(F(x_{S \rightarrow T}))) \\ &\quad + \log(1 - D_T(x_{T \rightarrow T'})),\end{aligned}$$

$$\mathcal{L}_G = \mathcal{L}_G^S + \mathcal{L}_G^T.\tag{4}$$

2.2.3. Classifier C . The classifier C is applied to the domain-invariant space. The goal of the classifier is to correctly

classify the source and target images, x_s and x_t . For image classification, we directly optimize the classifier C by minimizing a binary cross-entropy loss. For the generated image, the classifier also minimizes the classification loss of the cross-domain reconstructed image in the target domain. So, the label information can be introduced in the target domain to improve the classification performance of the target-domain-specific features. It should be noted that when we use cross-domain-generated images to update C , we do not use the gradient information of F and G . The loss function of C is shown in the following equation:

$$\mathcal{L}_C = \mathbb{E}_{\substack{x_s \sim S \\ x_t \sim \Gamma}} \min_C -\log(C(F(x_s)_y)) - \log(C(F(x_{S \rightarrow T})_y)).\tag{5}$$

2.2.4. Encoder F . A part of the loss function of the encoder F is the same as the decoder G . By minimizing the loss of the classifier C , classification cycle consistency ($C(F(x_{S \rightarrow S'}))$ and $C(F(x_{S \rightarrow T}))$) enhances the ability to extract certain domain-specific features. The GAN adversarial loss combines this encoder loss with the domain discriminator loss. In addition, the cross-domain discriminator loss is introduced to ensure the extraction of domain-invariant features. The purpose of the generator G is only to generate an image of each domain to fool the corresponding domain discriminator D without cross-domain mapping, while the purpose of the encoder F is mainly to map the image to the domain-invariant space that maximizes the domain discriminator confusion during cross-domain generation. We adjust the proportion of the same-domain reconstruction loss and cross-domain reconstruction loss in F by the parameters α (0.1) and β (0.1). The functions of the module F and module G are different from each other in terms of the optimization process, implicitly enforcing domain invariance of the extracted features. At the same time, the goal of the encoder network F is to minimize the label prediction loss with L2 regularization in the cross-domain confounding representation. Equation (5) shows the loss function of the encoder F :

$$\begin{aligned}\mathcal{L}_F &= \mathbb{E}_{\substack{x_s \sim S \\ x_t \sim \Gamma}} \min_F -\log(C(F(x_s)_y)) - \log(C(F(x_{S \rightarrow S'})_y)) - \log(C(F(x_{S \rightarrow T})_y)) \\ &\quad + \log(1 - D_T(x_{S \rightarrow T})) + \log(1 - D_S(x_{T \rightarrow S})) + \alpha \log(1 - D_S(x_{S \rightarrow S'})) \\ &\quad + \beta \log(1 - D_T(x_{T \rightarrow T'})) + L_2(x_{gS}) + L_2(x_{gT}).\end{aligned}\tag{6}$$

2.3. Model Training. Because each module has a different loss function, the training of our network model follows a different loss update rule for each module to achieve the isolation of parameter updates. With a training input of

unpaired source and target images, we alternately update the D , G , C , and F modules as mentioned in GTA [23]. For updating the discriminator D , we fix the parameters of all other modules and minimize the discriminator loss by

moving in the direction of the gradient information of the discriminator. For the update of the decoder G , the parameters of the other modules are fixed, while the G parameters are updated by minimizing the classification cycle consistency loss and maximizing the discriminant loss. For the update of the classifier C , we fix all parameters except for those of the classifier and use only the source-domain images and the cross-domain generated images with the label information to minimize the classification loss under the gradient information of the encoder F and the classifier C . The objective for the encoder F is to maximize the domain discriminant loss, minimize the classifier loss, minimize the classification cycle consistency loss, and maximize the discriminant loss of the cross-domain generation. From the composition of the loss function, we can see that G , C , and F have the same part in the loss function but also have their own unique parts, so we need to update the gradient independently. On the contrary, the loss of F comes from C , D , and G at the same time. On the basis of updating G and C , it can better provide effective gradient information for F , thereby better optimizing the parameters of F . In the update order of F , G , and C , our principle is to use the gradient of the previous network to better train the subsequent network based on the order of the network architecture and the composition of the loss function.

2.4. Model Verification. In order to assess the domain adaptation of the proposed network, we evaluate the classification performance on a source-domain validation set and a target-domain dataset. Figure 3 shows the verification phase, in which only the encoder and the classifier are reserved. For domain verification, the dataset is specifically divided into verification sets that are independent of the training set. Finally, the target-domain and source-domain verification accuracies are used to determine the model stability and to assess the performance of domain adaptation.

2.5. Dataset Description. CT images are widely used in medical examinations, and scans of different resolution are also performed in clinical diagnosis depending on the patient's condition. In particular, CT imaging has been widely used for computer-aided diagnosis of benign and malignant lymph nodes [25, 26]. During the initial diagnosis of lymph nodes, CT scans are divided into two types: plain scans and enhanced scans, where the enhanced scans show sharper details of soft tissues. A doctor will subjectively decide which scanning method to use based on the patient's initial situation. To achieve the domain adaptation for the two scan types, the trained network should eliminate well the background interference and pay attention to the class-conditioning features. The dataset used in our experiments is provided by the Department of Radiology of the Henan Provincial People's Hospital, a governmental public medical institution in China. The Department of Radiology of the Henan Provincial People's Hospital approved this study and waived the need to obtain informed consent from the patients. The enhanced CT images included 1409 malignant cases and 1099 benign cases, while the plain-scan CT images

included 1310 malignant cases and 1358 benign cases. Based on the limitation of processing 2D data by our network, when processing 3D CT scan data, we extract the center slice of the lesion area as a sample from the vertical axis direction of each 3D data. The prediction of 3D images based on 2D slices can be used as the direction of later research. Figure 4 shows benign and malignant images of the two scan types which are the two domains we address in this work. The original size of each image is 512×512 . In order to reduce the memory consumption, we uniformly scale the images to a size of 256×256 before network training.

In order to verify the general applicability of our model, we also performed experiments on datasets with simple data distributions but high domain shifts. In particular, we selected the street view house number (SVHN) dataset [27] and MNIST [28] as experimental datasets. The SVHN dataset is a nonhandwritten digit dataset of color images with few structural changes, but with certain background interference. The MNIST dataset is a handwritten digit dataset of binarized images, with clear digital structure variations. To achieve domain adaptation for the two datasets, the network needs to eliminate well the background interference and accurately capture the common feature representation of the handwritten and nonhandwritten digits. We also experimented with digital datasets on a classification network of an encoder F and a classifier C for comparison with later experiments.

2.6. Implementation Details. Because the target-domain labels are not available in practical domain-adaptation problems, the final domain-adaptation model is generally obtained as the end-of-training model. Indeed, the model stability during training is crucial for reaching a highly optimal model in practical applications. We mainly evaluate the model performance in terms of stability and accuracy. For experimental setup, batch normalization was employed prior to the rectified linear-unit (ReLU) activation function in network training. Moreover, the Adam optimizer with a momentum of 0.99 was adopted. This framework was executed using PyTorch under the NVIDIA TITAN V 12-GB GPU.

3. Results and Discussion

3.1. Source-Only Verification Experiment. We demonstrate experimentally the classification performance based on our data and assess the subsequent domain adaptation effects on the associated source domain. In particular, we use the CT data of each single center to learn the parameters of the network architecture. The results are derived for the validation set and are shown in Table 1. Indeed, our data can achieve a good classification performance under the same deep learning model. The classification performance is similar for the cases of the two single-center CT datasets. Nevertheless, the features of the enhanced CT scans and the MNIST dataset are easier to extract and lead to a higher verification accuracy.

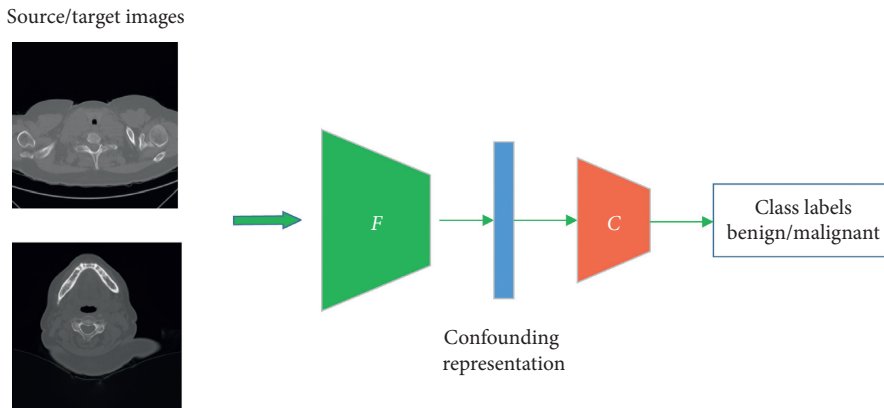


FIGURE 3: Illustration of the verification phase. F refers to the encoder which encodes the image to domain invariant space and C refers to the classifier. All those parameters are fixed during the verification.

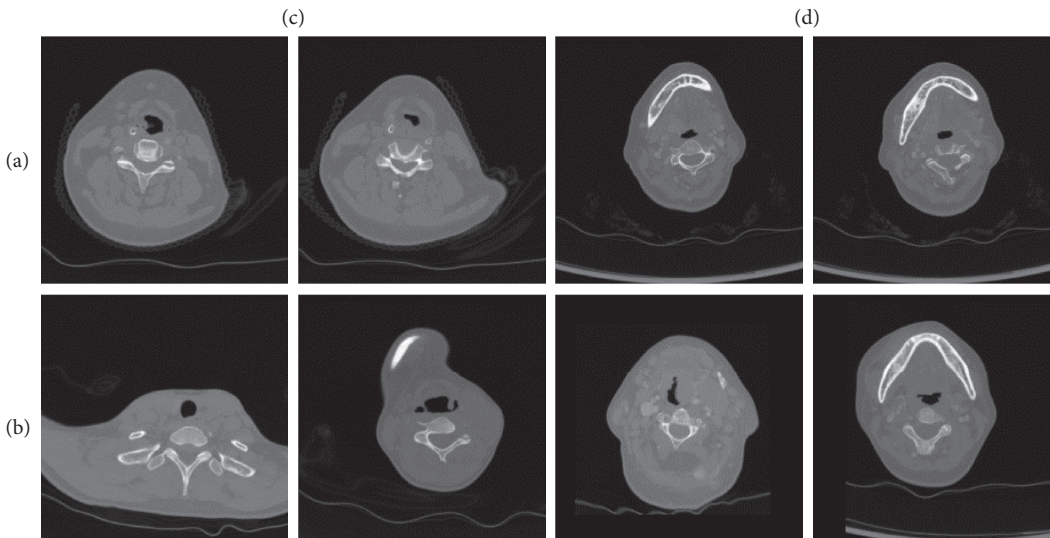


FIGURE 4: Samples of the multicenter CT images used for network training: (a) benign cases; (b) malignant cases; (c) plain CT scans; (d) enhanced CT scans.

TABLE 1: Accuracy values on different datasets using the verification network model.

	Plain CT scan	Enhanced CT scan	SVHN	MNIST
Verification accuracy (%)	84.6	88.4	98.4	99.2

3.2. *Domain Adaptation Experiments.* We report here the experimental validation results of our approach. We use the commonly used domain-adaptation methods for comparison. All the mentioned procedures are reproduced in the same deep learning framework according to the reference papers. In addition to the medical dataset, we also validate our method on the simple data distributions of the digit datasets to demonstrate the universality of our method.

We test the three common domain adaptation settings. The first setting, SVHN MNIST, refers to domain adaption with SVHN as the source domain and MNIST as the target domain, while the setting MNIST SVHN refers to opposite domain-adaptation process. In the same way, the two settings, plain enhanced and enhanced plain, represent the domain adaptation on our medical dataset of the plain and

enhanced CT scans. Table 2 shows the performance under different methods, datasets, and domain settings. The performance measure is the average accuracy variance of a ten-fold validation scheme.

The source-only method means that the model was trained with the source-domain dataset and directly tested on the target-domain dataset. The network architecture is the combination of the encoder F and discriminator C networks. From Table 2, we can observe that our method significantly closes the gap between the two domain classification spaces. For the simple data distribution (MNIST and SVHN) experiment, our method almost attains the best performance obtained by GTA [23] but with lower variance value, which demonstrates the robustness of model. GRL also performed well under simple distribution data (85.4 ± 1.7 ,

TABLE 2: Accuracy values (mean \pm std%) with different models, datasets, and domain settings.

	MN \rightarrow SV	SV \rightarrow MN	Enhanced \rightarrow plain	Plain \rightarrow enhanced
Source only	73.1 \pm 1.4	68.3 \pm 1.5	61.5 \pm 2.3	61.6 \pm 3.5
GRL [7]	85.4 \pm 1.7	87.2 \pm 2.1	65.4 \pm 3.9	60.2 \pm 2.3
MMD [8]	62.6 \pm 0.7	66.1 \pm 0.8	63.5 \pm 2.7	65.7 \pm 3.2
DSN [12]	81.3 \pm 1.4	86.4 \pm 0.5	58.5 \pm 4.1	55.2 \pm 3.4
GTA [23]	92.5 \pm 1.2	92.4 \pm 1.3	69.4 \pm 1.1	67.4 \pm 1.8
Ours	91.6 \pm 0.3	91.8 \pm 0.4	73.8 \pm 0.9	72.5 \pm 1.3

87.2 \pm 2.1), while MMD performed poorly (62.6 \pm 0.7, 66.1 \pm 0.8). This difference in performance reflects that the MMD, as a loss description of feature space distance, may focus network to extract feature that is not related to class information when the prior class-conditioning information of the source and target domains differs greatly. GRL improves discrimination of domain-invariant features with more flexible adversarial losses. DSN (81.3 \pm 1.4, 86.4 \pm 0.5) uses GRL to perform similarity constraints, and by explicitly modeling the individual features of each domain, it improves the constraint of domain invariant features. The suboptimal performance of these methods reflects the rationality extraction of domain invariant feature through cross-domain confounding representations based on adversarial loss.

For the complex data distribution of the CT images, our method achieves the best performance (73.8 \pm 0.9, 72.5 \pm 1.3). The worse performance of DSN (58.5 \pm 4.1, 55.2 \pm 3.4) shows the complexity between domain-invariant features and domain-specific features under complex data distributions. GRL (65.4 \pm 3.9, 60.2 \pm 2.3) and MMD (63.5 \pm 2.7, 65.7 \pm 3.2) do not perform well on complex data distributions compared with the source-only performance, which demonstrate the effectiveness of class-conditioning information on domain-invariant spaces. The other methods show large fluctuations in network performance stability when the features become complex. Meanwhile, the performance of our method is the most stable and is also robust to the shift of the class-conditioning information in the complex feature distribution. This stability makes our method more applicable in practice for domain adaptation. Compared to the results of Table 1 and the source-only method, the results of all domain-adaptation methods are close to the results of training the same classification network with images of a specific domain under supervised conditions. We can make some observations about the verification accuracy of the source and target domains. For the CT image dataset, the accuracy of the target-domain verification will rise slowly. The source-domain verification accuracy will first rise rapidly, decline, and then rise slowly. Compared to the highest source-domain accuracy, the final stable source-domain accuracy is lower than the highest initial value, but corresponds to a higher target-domain accuracy.

We run additional experiments to assess the effect of the classification cycle consistency on domain adaptation and thus verify the importance of the class-conditioning information in domain-invariant spaces. Specifically, we test the proposed framework with and without the classification cycle consistency loss. The results are shown in Table 3. Compared with models without category information

TABLE 3: Effect of the classification cycle consistency on the classification accuracy.

	Enhanced \rightarrow plain	Plain \rightarrow enhanced
Without classification cycle consistency	67.6 \pm 1.4	65.4 \pm 1.1
With classification cycle consistency	73.8 \pm 0.9	72.5 \pm 1.3

constraints, the introduction of category information is critical and useful. The network with a classification cycle consistency loss has better accuracy and stability (73.8 \pm 0.9, 72.5 \pm 1.3) under complex data distribution than a network without this loss (67.6 \pm 1.4, 65.4 \pm 1.1). The results show that the classification cycle consistency, as a method of adding class-conditioning gradients in the F and G , constrains the class-conditioning feature in domain invariant space. Compared with the result from GTA (Table 2), if it was not the classification cycle consistency for the proposed method, the performance is worse. The reason for analysis is that the GTA model introduces class information to the encoder through the classification loss of discriminator. These two results show that, no matter which domain-invariant feature extraction method, the introduction of class-conditioning information is more critical and useful. We can see that our classification cycle consistency concept not only imposes consistency constraints on the cross-domain mapping but also ensures flexible perception of the class-conditioning information and leads to a higher accuracy and a more stable performance.

4. Discussion

To solve the computer-aided diagnosis for multicenter lymph node data, a domain adaptation network used for complex data distribution is proposed. Unlike other domain adaptation methods, the proposed unsupervised domain adaptation method extracts domain-invariant feature through cross-domain confounding representations so that it can get a domain-invariant space for target and source data. Moreover, the classification cycle consistency captures the class-conditioning information in domain-invariant space in order to solve the difficult-to-perceive-detail class-conditioning information. The effectiveness of the proposed method is proved through the experimental results. Meanwhile, the proposed method achieves the more robust performance under both simple data distribution and complex data distribution, which may be because of the

available constraint of the class-conditioning feature in domain invariant space. Moreover, it is found that the cross-domain confounding representations based on adversarial training are also robust for the change of update sequence. Although the accuracy rates are higher than 70% through domain adaptation, it is much lower than what it should be. Maybe it is because the weight for each loss constraint needs to be well selected to balance the effect of various constraints on different parts. Furthermore, the domain adaptation performance is also needed to be demonstrated under 3D image application.

5. Conclusions

This paper proposes a domain adaptation method to solve the multicenter problem of the computer-aided diagnosis for lymph nodes. Aiming at the high-resolution and complex feature distribution of the lymph CT images, this paper constructs a confounding representation of domain features based on a cross-domain mapping to achieve domain-invariant feature extraction. For the complex data distribution of lymph CT images, the classification cycle consistency guides the proposed model to perceive significant classifiable representations in a domain-invariant space. Through the above method, our model achieves a stable domain adaptation of high-resolution images in complex medical field. The experimental results for simple data distributions also show the versatility of the proposed method. The training stability also allows us to simply get the optimal model under the target domain, which can further achieve the multidomain medical data integration.

Data Availability

The code used in the research can be obtained from https://pan.baidu.com/s/1-m6PckJgM9v_JI-hCUg50g.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Nation Key R&D Program of China under grant 2018YFC0114500.

References

- [1] J. Wang, Q. Wang, and J. Peng, "Multi-task diagnosis for autism spectrum disorders using multi-modality features: a multi-center study," *Human Brain Mapping*, vol. 38, no. 6, pp. 3081–3097, 2017.
- [2] L. Yuan, X. Wei, H. Shen, L.-L. Zeng, and D. Hu, "Multi-center brain imaging classification using a novel 3D CNN approach," *IEEE Access*, vol. 6, pp. 49925–49934, 2018.
- [3] E. A. Van, A. C. van Dijk, and M. T. B. Truijman, "Multi-center MRI carotid plaque component segmentation using feature normalization and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 34, no. 6, pp. 1294–1305, 2015.
- [4] T. Zuo, H. Zeng, and H. Li, "The influence of stage at diagnosis and molecular subtype on breast cancer patient survival: a hospital-based multi-center study," *Chinese Journal of Cancer*, vol. 36, no. 1, p. 84, 2017.
- [5] Z. Jing, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1859–1867, Honolulu, HI, USA, July 2017.
- [6] J. Hong, H. Chen, and L. Feng, "Disturbance Grassmann kernels for subspace-based learning," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1521–1530, ACM, London, UK, August 2018.
- [7] Y. Ganin, E. Ustinova, and H. Ajakan, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, pp. 2096–2030, 2017.
- [8] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Scholkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.
- [9] E. Tzeng, J. Hoffman, Z. Ning, K. Saenko, and T. Darrell, "Deep domain confusion: maximizing for domain invariance," *Computer Science*, 2014, <http://arxiv.org/abs/1412.3474>.
- [10] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 4, pp. 801–814, 2018.
- [11] B. Sun and K. Saenko, "Deep CORAL: correlation alignment for deep domain adaptation," in *Proceedings of the European Conference on Computer Vision*, pp. 443–450, Amsterdam, Netherlands, October 2016.
- [12] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 343–351, Barcelona, Spain, December 2016.
- [13] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, *Learning to Discover Cross-Domain Relations with Generative Adversarial Networks*, Korea Intelligent Information Systems Society, Seoul, Republic of Korea, 2017.
- [14] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, <http://arxiv.org/abs/1611.02200>.
- [15] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 700–708, Long Beach, CA, USA, December 2017.
- [16] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, Venice, Italy, October 2017.
- [17] M. Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 469–477, Barcelona, Spain, December 2016.
- [18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, <http://arxiv.org/abs/1312.6114>.
- [19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, <http://arxiv.org/abs/1511.06434>.
- [20] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–51, Springer, Berlin, Germany, 2018.

- [21] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, Springer, Berlin, Germany, 2018.
- [22] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, and K. Murphy, "XGAN: unsupervised image-to-image translation for many-to-many mappings," 2017, <http://arxiv.org/abs/1711.05139>.
- [23] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: aligning domains using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, Salt Lake City, UT, USA, June 2017.
- [24] I. J. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2672–2680, Montreal, Canada, December 2014.
- [25] M. Giovannini, T. Botelberge, and E. Bories, "Endoscopic ultrasound elastography for evaluation of lymph nodes and pancreatic masses: a multicenter study," *World Journal of Gastroenterology*, vol. 15, no. 13, pp. 1587–1593, 2009.
- [26] S. M. Dudea, B. J. Carolina, D. Dana, V. Dan, M. Simona, and L. Manuela Lavinia, "Differentiating benign from malignant superficial lymph nodes with sonoelastography," *Medical Ultrasonography*, vol. 15, no. 2, pp. 132–139, 2013.
- [27] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, vol. 5, Daejeon, South Korea, 2011.
- [28] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

Research Article

A Long Short-Term Memory Ensemble Approach for Improving the Outcome Prediction in Intensive Care Unit

Jing Xia ¹, Su Pan,¹ Min Zhu,¹ Guolong Cai,² Molei Yan,² Qun Su,³ Jing Yan ²,
and Gangmin Ning ¹

¹Department of Biomedical Engineering, Zhejiang University, 38 Zheda Road, Hangzhou 310027, China

²Department of ICU, Zhejiang Hospital, 12 Lingyin Road, Hangzhou 310013, China

³Department of ICU, The First Affiliated Hospital, Zhejiang University, 79 Qingchun Road, Hangzhou 310003, China

Correspondence should be addressed to Jing Yan; yanjing201801@163.com and Gangmin Ning; gmning@zju.edu.cn

Received 29 November 2018; Revised 23 September 2019; Accepted 8 October 2019; Published 3 November 2019

Guest Editor: Andrea Duggento

Copyright © 2019 Jing Xia et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In intensive care unit (ICU), it is essential to predict the mortality of patients and mathematical models aid in improving the prognosis accuracy. Recently, recurrent neural network (RNN), especially long short-term memory (LSTM) network, showed advantages in sequential modeling and was promising for clinical prediction. However, ICU data are highly complex due to the diverse patterns of diseases; therefore, instead of single LSTM model, an ensemble algorithm of LSTM (eLSTM) is proposed, utilizing the superiority of the ensemble framework to handle the diversity of clinical data. The eLSTM algorithm was evaluated by the acknowledged database of ICU admissions Medical Information Mart for Intensive Care III (MIMIC-III). The investigation in total of 18415 cases shows that compared with clinical scoring systems SAPS II, SOFA, and APACHE II, random forests classification algorithm, and the single LSTM classifier, the eLSTM model achieved the superior performance with the largest value of area under the receiver operating characteristic curve (AUROC) of 0.8451 and the largest area under the precision-recall curve (AUPRC) of 0.4862. Furthermore, it offered an early prognosis of ICU patients. The results demonstrate that the eLSTM is capable of dynamically predicting the mortality of patients in complex clinical situations.

1. Introduction

Mortality prediction is essential for the clinical administration and treatment, especially in the intensive care unit (ICU) [1, 2]. Various scoring systems have been developed and widely used for assessing the clinical outcome, and the most common ones are simplified acute physiology score (SAPS) II [3], sequential organ failure assessment (SOFA) [4], and acute physiology and chronic health evaluation (APACHE) II [5]. Scoring systems assess the patients' mortality by logistic regression model assuming a linear and additive relationship between the severity of the disease and the collected relevant physiological parameters, which are practicable but unrealistic [6]. In the recent years, machine learning was introduced in the medical application and showed its remarkable efficiency in clinical diagnosis and decision support. For admitted ICU patients, lots of physiological measurements are collected, containing symptoms, laboratory tests, and vital signs (such as

heart rate, blood pressure, and respiratory rate) [7, 8]. The clinical measurements are continuously monitored in ICU with the values fluctuating as time progresses and the temporal trends are predictive of mortality [9]. Hence, sequence of clinical records offers rich information of patients' physical condition [10, 11] and enables the utilization of machine learning in developing prognosis model from these multivariate time series data. As a decision task, mortality prediction can be solved by classification algorithms such as logistic regression, support vector machine, and random forests (RF) [12]. However, most of the methods currently used are not sensitive to the temporal link among the sequent data and thus are not able to receive full benefits of the ICU data, which limits their performances in the mortality prediction [10, 13].

Presently, recurrent neural network (RNN) was well employed in solving time series prediction problems and achieved prominent results in many fields [14–19]. Several

variants of RNN have been developed, and among them, long short-term memory (LSTM) network is one of the most popular variants [20]. LSTM learns long-term dependencies by incorporating a memory cell that is able to preserve state over time. Three gates are equipped in LSTM for deciding which information to summarize or forget before moving on to the next subsequence [21–23]. LSTM is well suited to capture sequential information from temporal data and has shown advantages in machine translation [24, 25], speech recognition [19], and image captioning [26], etc. In the medical domain, many efforts have been made to apply LSTM for clinical prediction based on electronic health records [6, 17, 27–30]. Lipton et al. employed LSTM on a collection of 10, 401 episodes to establish a model for phenotype classification [28]. Given 13 frequently sampled clinical measurements (diastolic and systolic blood pressure, peripheral capillary refill rate, end-tidal CO₂, fraction of inspired O₂, Glasgow coma scale, blood glucose, heart rate, pH, respiratory rate, blood oxygen saturation, body temperature, and urine output), the LSTM model was able to predict whether the patient suffered from 128 most common conditions, such as acute respiratory distress, congestive heart failure, and renal failure. Jo et al. used LSTM and latent topic model to extract information from textual clinical notes for assessing the severity of diseases [29]. Pham et al. conducted experiments on a diabetes cohort of 7191 patients with 53208 admissions collected in 2002–2013 from a large regional Australian hospital, and the results showed improved performances of utilizing LSTM in disease progression modeling and readmission prediction [31].

For ICU mortality prediction, the current prognosis models mostly employed single LSTM classifier [6, 29, 30]. However, in most cases, a single model is not efficient enough to handle the complex situation in ICU. Patients in ICU are heterogeneous suffering from different diseases with multiple concurrent problems, and the clinical data in ICU are highly complex [9, 32, 33]. For patients with various diseases, the underlying pathophysiologic evolutions of the patients (e.g., kidney failure) are usually manifested through different sets of physiologic variables (e.g., abnormalities in glomerular filtration rate and creatinine) [9]. Even for the patients having the same disease, they might have different comorbidities experiencing heterogeneous health conditions [33]. Thereby, hybrid learners are required for the prediction model in ICU.

An ensemble learner principally has a stronger generalization ability than a single learner [34–37]. Ensemble learning is a procedure that integrates a set of models for a given problem to obtain one composite prediction [38–43]. Diverse classifiers are constructed to learn multiple hypotheses, and the multiple resulting predictions are aggregated to solve the same problem. In contrast to the stand-alone model which builds one hypothesis space, a combination of several models can expand the space and may provide a more exact approximation to the true hypothesis [34]. It has been shown that ensemble systems outperformed single classifier systems in solving complex problems [34, 38, 39].

Therefore, we proposed an ensemble algorithm of multiple long short-term memory networks (eLSTMs) to deal with the complex situation in ICU. In eLSTM, the

diversity of LSTM models owes to the multifariousness of subsets for building the models. Two strategies are employed to produce different subsets from the entire training data, namely, bootstrapped samples and random feature subspace. Bootstrapped samples strategy generates various subsets of subjects, while random feature subspace provides different combined sets of clinical indicators. That is, the subsets are distinguished from each other at both instance and feature level. A variety of LSTM classifiers are trained accordingly, and the final score is computed as the average of predicted values from all base learners. Generally, the eLSTM algorithm selects a number of training subsets using bootstrapped instances with randomly chosen feature set, constructs multiple LSTM learners on the multiple subsets, and averages all individuals' predicted scores as final output. The main contributions of this work are as follows: (1) proposing an LSTM ensemble framework to develop hybrid sequential classification model which is able to handle complex clinical situations such as ICU and (2) applying bootstrapped samples and random feature subspace to individual LSTM classifiers for creating diversity in the ensemble. The present model will promote the application of machine learning in complex clinical situations.

The rest of this paper is organized as follows. Section 2 describes the ICU dataset, the implementation of the proposed eLSTM algorithm, and the experimental design. The empirical results yielded by various systems for mortality prediction are presented in Section 3. The advantages of eLSTM are discussed in Section 4. Finally, Section 5 concludes this paper and indicates the future work.

2. Methods

2.1. Dataset. The ICU data for this work were extracted from the Medical Information Mart for Intensive Care III (MIMIC-III) database [44]. MIMIC-III is a large and publicly available database of ICU admissions at the Beth Israel Deaconess Medical Center, USA, from 2001 to 2012. It comprises rich clinical data of patients, including the laboratory tests and vital signs. A total of 18415 patients were extracted from MIMIC-III database with age >15 years and length of stay ≥ 10 days. The prediction task of clinical outcome is 28-day postadmission mortality. The study population consists of 2162 subjects in positive group that died within 28 days after ICU admission and the other 16253 subjects in negative group that survived 28 days after ICU admission. From the tables LABEVENTS.csv and CHARTEVENTS.csv, 50 variables of continuous 10 days (denoted as D_1, D_2, \dots, D_{10}) are recorded for mortality prediction. The variables are sampled every 24 hours. These variables are commonly used clinical measurements, and the details are listed in Table 1.

2.2. LSTM Ensemble Algorithm. Ensemble methods generate multiple learners and aggregate them to provide a composite prediction. Among them, the Bagging and Boosting method are most popular. The diversity of individual learner is an important issue for ensemble model, which can be achieved

TABLE 1: Variables for mortality prediction.

Variable no.	Source table name	Variable name
1	LABEVENTS	BUN
2	LABEVENTS	WBC
3	LABEVENTS	HCO ₃ ⁻
4	LABEVENTS	Na ⁺
5	LABEVENTS	K ⁺
6	LABEVENTS	TBil
7	LABEVENTS	Plt
8	LABEVENTS	Cr
9	LABEVENTS	PH
10	LABEVENTS	HCT
11	LABEVENTS	Lactate
12	LABEVENTS	Hemoglobin
13	LABEVENTS	MCHC
14	LABEVENTS	MCH
15	LABEVENTS	MCV
16	LABEVENTS	Red Blood Cells
17	LABEVENTS	RDW
18	LABEVENTS	Chloride
19	LABEVENTS	Anion Gap
20	LABEVENTS	Glucose
21	LABEVENTS	Magnesium
22	LABEVENTS	Calcium, Total
23	LABEVENTS	Phosphate
24	LABEVENTS	INR
25	LABEVENTS	PT
26	LABEVENTS	PTT
27	LABEVENTS	Lymphocytes
28	LABEVENTS	Monocytes
29	LABEVENTS	Neutrophils
30	LABEVENTS	Basophils
31	LABEVENTS	Eosinophils
32	LABEVENTS	Base Excess
33	LABEVENTS	Calculated Total CO ₂
34	LABEVENTS	PCO ₂
35	LABEVENTS	Specific Gravity
36	LABEVENTS	ALT
37	LABEVENTS	AST
38	LABEVENTS	Alkaline Phosphatase
39	LABEVENTS	Albumin
40	LABEVENTS	PEEP
41	LABEVENTS	PaO ₂
42	CHARTEVENTS	GCS
43	CHARTEVENTS	SBP
44	CHARTEVENTS	HR
45	CHARTEVENTS	T
46	CHARTEVENTS	MAP
47	CHARTEVENTS	RR
48	CHARTEVENTS	A-aDO ₂
49	CHARTEVENTS	FiO ₂
50	LABEVENTS, CHARTEVENTS	PaO ₂ /FiO ₂

by selecting and combining the training examples or the input features, injecting randomness into the learning algorithm [34, 36].

The proposed eLSTM algorithm is an ensemble method utilizing LSTM as base learner. Two random strategies are employed to produce different training subsets, hence constructing a number of base LSTM classifiers. All predictions are integrated to give a comprehensive estimate of the outcome.

Given a training set with N training instances, each instance can be represented as (V, Y) . V is a matrix containing values of D variables and T sequences. It can be written as $[X_1, X_2, X_3, \dots, X_t, \dots, X_T]$, as expressed in equation (1). X_t is a vector given in equation (2). x_t^d represents the value of the d -th variable at t -th time step. And Y is the target label for the instance taking 0 (negative) for survival and 1 (positive) for death. The ratio of negative and positive group size is denoted as γ :

$$V = [X_1, X_2, X_3, \dots, X_t, \dots, X_T], \quad (1)$$

$$X_t = [x_t^1, x_t^2, x_t^3, \dots, x_t^d, \dots, x_t^D]. \quad (2)$$

LSTM has the advantage of capturing temporal information and is popular to be adopted in time series modeling. Detailed structure of the LSTM block is illustrated in Figure 1.

The input of LSTM block is X_t . Then, the output of hidden layer, namely, the current hidden state h_t , is computed as follows:

$$\begin{aligned} f_t &= \sigma(w_f[h_{t-1}, X_t] + b_f), \\ i_t &= \sigma(w_i[h_{t-1}, X_t] + b_i), \\ o_t &= \sigma(w_o[h_{t-1}, X_t] + b_o), \\ C_t &= f_t * C_{t-1} + i_t * \tanh(w_c[h_{t-1}, X_t] + b_c), \\ h_t &= o_t * \tanh(C_t), \end{aligned} \quad (3)$$

where f_t , i_t , and o_t are the forget, input, and output gates, respectively. h_{t-1} is the previous hidden state. C_{t-1} and C_t are previous and current cell memories. The weight matrices w_f , w_i , w_o , and w_c and the bias vectors b_f , b_i , b_o , and b_c are model parameters. The symbol σ is the sigmoid function and \tanh hyperbolic tangent function. The symbol \cdot denotes matrix multiplication and $*$ elementwise product.

A sigmoid layer is applied on the output of the LSTM block at final step for binary classification. The predicted score \tilde{y} is computed as equation (4). The loss function is the weighted cross entropy of real label and predicted score \tilde{y} with positive instances weighted γ and negative ones weighted 1. The parameters within the net are updated over several iterations to reach the minimum loss value:

$$\tilde{y} = \sigma(w_{ho} \cdot h_T + b_{ho}). \quad (4)$$

The eLSTM model is composed of multiple LSTM classifiers, and its architecture is illustrated in Figure 2.

The procedure of eLSTM consists of two stages: base learner generation and integration.

In the stage of base learner generation, the bootstrap sampling strategy [37] and random subspace method (RSM) [35] are both employed to generate different training subsets for constructing diverse base learners. As a training set sampling method, bootstrap sampling randomly draws instances with replacement from the whole training set and RSM is to randomly choose a subset of variables. The subsets resulted from different bootstrapped instances with randomly selected variables are denoted as $\{\text{Subset}_1, \text{Subset}_2, \dots, \text{Subset}_p, \dots, \text{Subset}_p\}$.

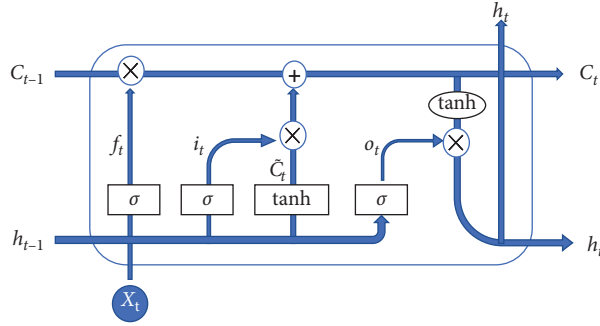


FIGURE 1: Illustration of the LSTM block's structure.

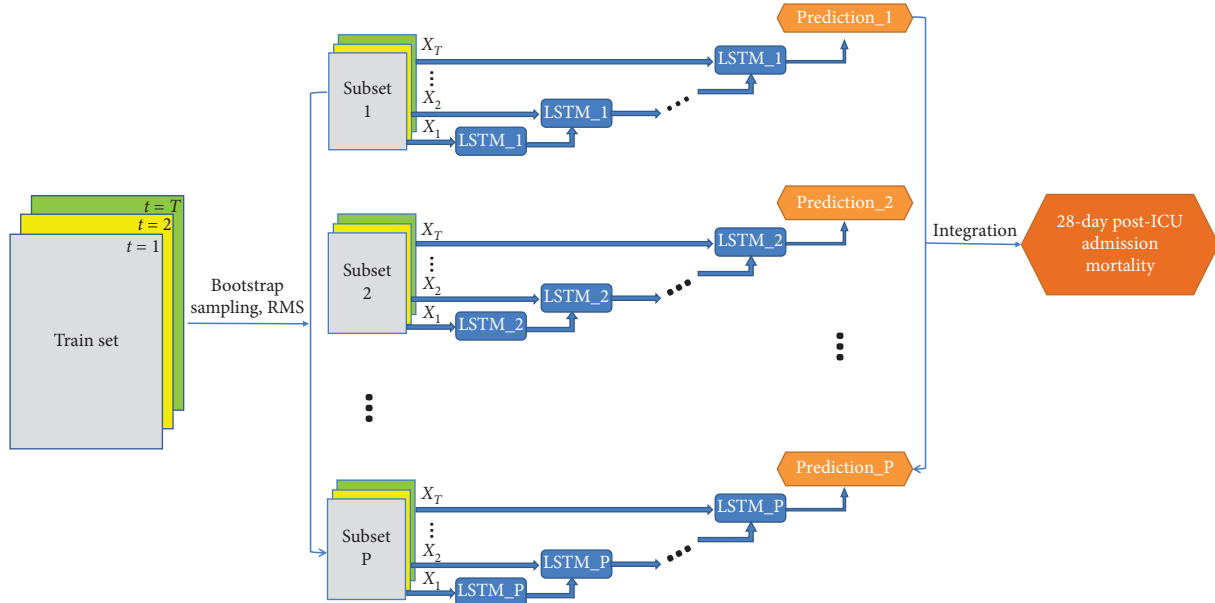


FIGURE 2: The architecture of the eLSTM algorithm.

In ensemble model rather than error control strategy, bias control is generally adopted to train multiple base classifiers benefiting the diversity of the model. Thus, appropriate number of training epochs for the classifiers is selected by experiments under a satisfied level of bias. The variance of the model due to the diversity of individual classifiers is controlled by the following ensemble operation [45, 46]. For eLSTM, the number of training epochs was set as 100, which was validated by pre-experiments.

Then, multiple LSTM classifiers learn from the subsets. Let $\{F^1, F^2, \dots, F^P\}$ denote the set of P trained base classifiers. For the input V , the p -th LSTM classifier gives an individual predicted score $\tilde{y}(p)$, as expressed in equation (5).

Finally, in the integration stage, the scores of all LSTM classifiers are averaged as the overall output and calculated as follows:

$$\tilde{y}(p) = F^p(V), \quad (5)$$

$$\tilde{Y} = \frac{1}{P} \sum_{p=1}^P \tilde{y}(p). \quad (6)$$

The procedure of the eLSTM algorithm is provided in Figure 3.

Once the eLSTM model is accomplished, it is applied in this way: for an instance, each LSTM classifier uses partial values of the corresponding variable subset and makes a prediction; different LSTM classifiers utilize different sets of variables, producing multiple prediction scores; the final prediction is obtained by averaging all scores.

2.3. Dynamic Prediction. For LSTM and eLSTM models, the full sequence of data is needed to predict the outcome. However, in practice, the patients' physiological parameters are collected day by day. To develop a dynamic procedure providing daily prediction, in this work, the values for coming days are padded by the latest available data to acquire complete sequences. Then, the LSTM algorithm and the eLSTM algorithm are employed on the complete dataset for predicting the outcome. Thus, the mortality assessment is updated daily with the replenished data approaching closer to the reliability. The process is illustrated in Figure 4.

Algorithm: eLSTM

For $p = 1$ to P // P is the number of base classifiers

- (i) Generating the Subset _{p}
 1. Generate N bootstrapped instances from the whole training set
 2. Randomly choose half of the variables
- (ii) Training individual LSTM classifier
Train the p -th LSTM classifier F^p
- (iii) Making a prediction
For the given input, predict the outcome with the score $\tilde{y}(p)$

End

Compute the final prediction as the average of all scores

FIGURE 3: Procedure of eLSTM algorithm.

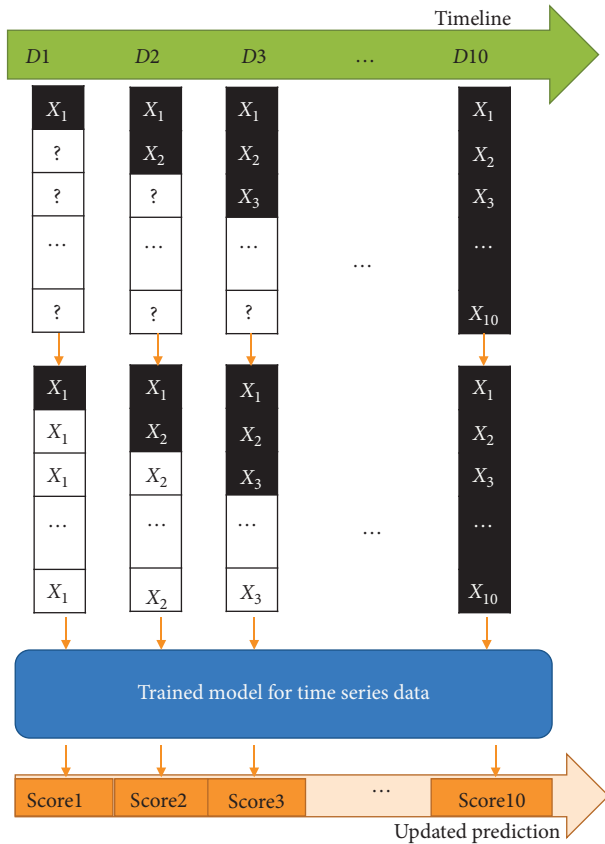


FIGURE 4: Flow diagram of dynamic prediction with data updating.

2.4. Experiment Design. The proposed eLSTM algorithm is compared with three scoring systems (SAPS II, SOFA, and APACHE II), RF algorithm, and LSTM classifier. In the LSTM classifier, a sigmoid layer is applied on top of the LSTM block for binary classification. The LSTM block has one hidden layer with 64 hidden units, and a dropout of rate 0.5 is applied to the input layer. The weight parameters are initialized randomly using Glorot uniform initialization [47]. The LSTM model is trained with the Adam optimizer of learning rate of 0.01 for a maximum of 100 epochs. 10% of

the training data are used as a validation set to find the best epoch. In eLSTM algorithm, there are two important hyperparameters: the number of base LSTM classifiers and the size of variable subset. Considering the running time, the number of base LSTM classifiers in the current work is set as 200. And, half of the variables are randomly chosen to construct individual classifier as recommended in the literature [35]. Eventually, 200 individual LSTM-based classifiers are trained on resampled instances with 25 randomly selected variables. In addition, dynamic prediction by RF algorithm is realized by training 10 models on data of the first 1, 2, . . . , 10 days, respectively.

The experiment is repeated 50 times. For each experiment, 90% of the dataset is chosen as training data and the left 10% as test data. Before the training procedure, data are preprocessed by imputation and normalization. The missing values are filled by linear interpolation imputation method, assuming a linear development in time of the variable with missing data [48]. Then, all the variables are normalized by subtracting the means and dividing the standard deviations computed across the training data.

To compare the performances of these models, several metrics are computed on predicted scores and true labels. The receiving operating characteristics (ROC) curve and the precision-recall curve are plotted to evaluate the performance of the classifiers. The ROC curve uses $1 - \text{specificity}$ as the x -axis and sensitivity as the y -axis for all potential thresholds, while the precision-recall plot applies recall and precision as the x -axis and y -axis. The area under ROC (AUROC) and the area under precision-recall curve (AUPRC) are calculated for comparison. Moreover, the bias between the predicted class labels and the true labels is comprehensively measured by sensitivity/recall, specificity, accuracy, precision, and F1 score. Sensitivity/recall calculates how many true-positive cases are correctly classified as positive, while precision counts the proportion of true-positive cases in the cases classified as positive. F1 score is the harmonic mean of recall and precision.

3. Results

3.1. Mortality Prediction Performance. The ROC curves and precision-recall curves of all models are shown in Figures 5 and 6. The eLSTM model harvests the largest AUROC of 0.8505 and the largest AUPRC of 0.45.

Detailed statistical results of repeated experiments are given in Table 2. ANOVA test shows significant differences in AUROC, AUPRC, sensitivity/recall, specificity, accuracy, precision, and F1 among the utilized methods ($p < 0.001$). It can be seen the models of RF, LSTM, and eLSTM have much larger AUROC values (RF: 0.8282 ± 0.0151 , LSTM: 0.8382 ± 0.0158 , and eLSTM: 0.8451 ± 0.0136) than scoring systems SAPS II, SOFA, and APACHE II (SAPS II: 0.7788 ± 0.0166 , SOFA: 0.7354 ± 0.0184 , and APACHE II: 0.7467 ± 0.0173). The proposed eLSTM model has the largest mean AUROC value of 0.8451, LSTM approach the second largest mean AUROC value of 0.8382, and the RF method the third largest of 0.8282. The eLSTM model outperforms other models in terms of AUPRC with the largest value of

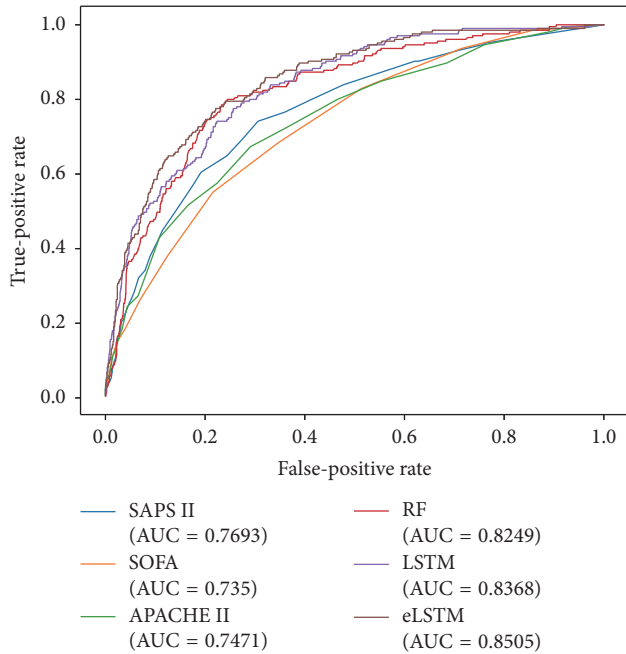


FIGURE 5: The ROC curves of all systems.

0.4862 ± 0.0345 . Also, the eLSTM algorithm has the largest sensitivity/recall of 0.7758 and the RF model and LSTM model have the medium value of 0.7197 and 0.7384, while the three scoring systems get the least value of 0.5418–0.6922. Post hoc analysis by Dunnett test shows the differences in AUROC, AUPRC, and sensitivity between eLSTM and other methods are significant ($p < 0.05$). Totally, the eLSTM model obtains the significant largest value of AUROC, AUPRC, and sensitivity. It is noticed that all methods have low precision and F1 score. It is mainly due to the imbalanced distribution of class label, that is, the number of negative instances is much larger than that of positive ones.

3.2. Dynamic Prediction. Figure 7 shows the time course of mortality prediction during one to ten days after the admission. It is seen that, with the available data updated daily, although the AUROC values of the various systems keep rising, through the whole procedure, the AUROC values of eLSTM, LSTM, and RF go higher than the three scoring systems. And from the third day, the eLSTM holds the highest value till the ending of the records. ANOVA followed by Dunnett test shows the AUROC value of the eLSTM model is significantly higher than that of LSTM and RF models (eLSTM vs. LSTM: $p = 0.011$; eLSTM vs. RF: $p = 0.000$). The charts also clearly reveal that while RF, LSTM, and the three scoring systems reach their highest performance on the last day, eLSTM achieves the corresponding levels at least 6 days earlier than the scoring systems and 2 and 1 days earlier than RF and LSTM, respectively. These facts demonstrate that eLSTM has stronger ability of dynamic prediction as well as early prognosis than the others.

Figure 8 shows that AUPRC has the similar trend with the data updating as AUROC. The eLSTM model harvests

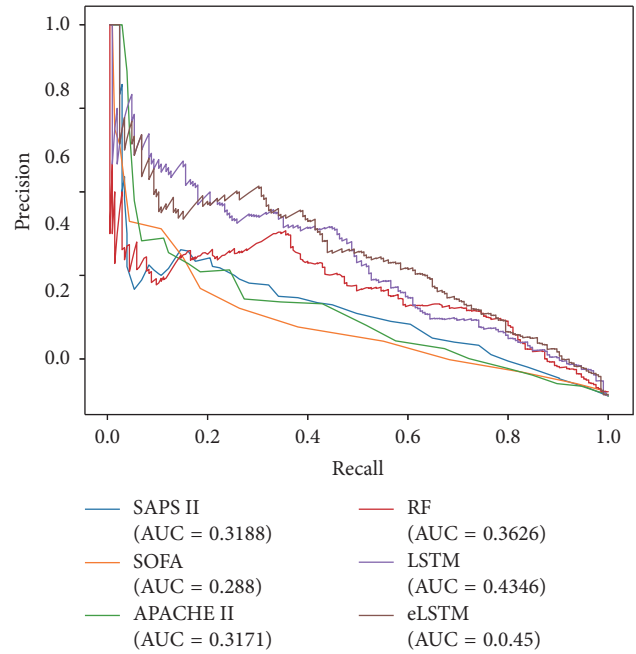


FIGURE 6: The precision-recall curves of all systems.

the largest AUPRC of 0.5 among all methods. ANOVA followed by Dunnett test exhibits that the AUROC value of eLSTM model is significantly higher than that of LSTM and RF (eLSTM vs. LSTM: $p = 0.043$; eLSTM vs. RF: $p = 0.000$).

3.3. Influence of the Number of LSTM Classifiers in eLSTM. The AUROC value of eLSTM goes up with the increase of the number of base LSTM classifiers (Figure 9). It has a steep ascent when less than 40 LSTM classifiers are integrated, then keeps a moderate rising, and finally stays at a plateau after 100 classifiers are involved. Similar situation is also observed in the AUPRC (Figure 10).

3.4. Influence of the Size of Variable Subset in eLSTM. ANOVA test indicates the size of variable subset in eLSTM models leads to significant difference in AUROC and as well as in AUPRC (AUROC: $F = 45.932$, $p = 0.000$; AUPRC: $F = 7.079$, $p = 0.002$). The AUROC values are similarly high for eLSTM with multiple sets of 16, 25, or 32 variables (Figure 11). And eLSTM achieves the largest AUPRC when the size of variable subset is 16, 25, or 32 (Figure 12). Pairwise comparison by Tukey test shows the AUROC and AUPRC values of eLSTM models trained by sets of 16, 25, and 32 variables are significantly higher than those of 8 and 50 variables ($p < 0.05$), while there are no significant differences among the models with sets of 16, 25, and 32 variables. In this work, the size of variable subset was set as the median value of 25, which is in agreement with the recommendation of literature [35].

4. Discussion

It is worth noticing that the algorithms of RF, LSTM, and eLSTM exhibit much better performance than the SAPS II,

TABLE 2: Evaluations of all mortality prediction systems (mean \pm std).

	SAPS II	SOFA	APACHE II	RF	LSTM	eLSTM	ANOVA test
AUROC	0.7788 \pm 0.0166*	0.7354 \pm 0.0184*	0.7467 \pm 0.0173*	0.8282 \pm 0.0151*	0.8382 \pm 0.0158*	0.8451 \pm 0.0136	$F = 926.328,$ $p = 0.000$
AUPRC	0.3800 \pm 0.0334*	0.3381 \pm 0.0307*	0.3515 \pm 0.0306*	0.4197 \pm 0.0393*	0.4751 \pm 0.0351*	0.4862 \pm 0.0345	$F = 426.683,$ $p = 0.000$
Sensitivity/ recall	0.6922 \pm 0.0267*	0.5418 \pm 0.0394*	0.6478 \pm 0.0303*	0.7197 \pm 0.0395*	0.7384 \pm 0.0401*	0.7758 \pm 0.0321	$F = 438.869,$ $p = 0.000$
Specificity	0.7404 \pm 0.0102*	0.7958 \pm 0.0101*	0.7256 \pm 0.0119*	0.7807 \pm 0.0218*	0.7746 \pm 0.0182*	0.7503 \pm 0.0136	$F = 229.707,$ $p = 0.000$
Accuracy	0.7347 \pm 0.0096*	0.7658 \pm 0.0106*	0.7164 \pm 0.0113*	0.7734 \pm 0.0174*	0.7703 \pm 0.0148*	0.7533 \pm 0.0112	$F = 234.492,$ $p = 0.000$
Precision	0.2633 \pm 0.0145*	0.2622 \pm 0.0179*	0.2404 \pm 0.0149*	0.3063 \pm 0.0211*	0.3056 \pm 0.0208*	0.2941 \pm 0.0158	$F = 271.132,$ $p = 0.000$
F1	0.3813 \pm 0.0180*	0.3532 \pm 0.0227*	0.3505 \pm 0.0187*	0.4290 \pm 0.0216	0.4317 \pm 0.0230	0.4262 \pm 0.0181	$F = 363.817,$ $p = 0.000$

*The difference with the eLSTM model is significant at the 0.05 level. Bold indicates the highest mean value.

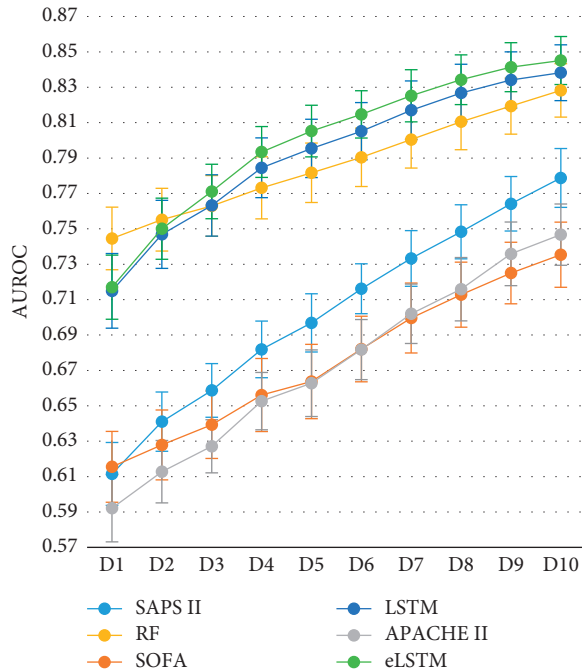


FIGURE 7: The AUROC values of all systems with data updating.

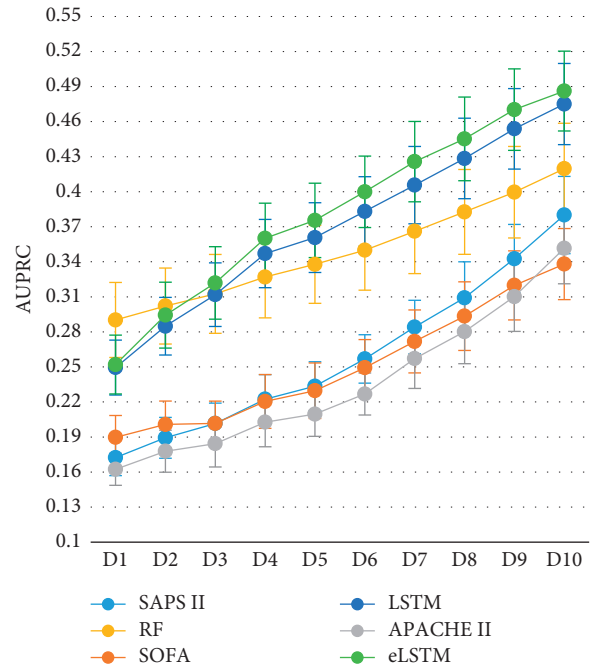


FIGURE 8: The AUPRC values of all systems with data updating.

SOFA, and APACHE II scoring system (Table 2). It indicates that data-driven mathematical model may help improve the mortality prediction in ICU and further other clinical tasks. Different models serve different purposes and situations. The present work demonstrates that, in dynamic prediction, LSTM and eLSTM are superior to the RF algorithm. RF is commonly considered as an easy-to-use algorithm for decision making. However, it is not sensitive to time course, resulting in the weakness in exploiting temporal information in the series data. But in the LSTM block, the values in the previous time steps impose influence on the coming time steps; hence, the LSTM block is capable of capturing temporal trends of the data and suitable for time series modeling. Moreover, with the updating of the input data, the predicting ability of LSTM is continuously improved. In other words, LSTM has the advantage in dynamic

predicting. The results demonstrate that generally, the eLSTM algorithm outperforms a single LSTM classifier. Also, it is seen in Figures 7 and 8 that the eLSTM model has much better achievement in early prediction than LSTM. It can be explained that instead of a single hypothesis space by one LSTM classifier, the eLSTM algorithm generates multiple base learners expanding the hypothesis space, which leads to a better approximation to the true hypothesis.

The proposed eLSTM algorithm successfully handles clinical time series data in ICU and provides a unified model for predicting the mortality of ICU patients. In ICU, patients are suffering from various diseases. Johnson et al. summarized the distribution of primary International Classification of Diseases (ICD) in the entire MIMIC-III database [44], as that the mostly common ones in ICU are infectious and parasitic diseases (ICD-9: 001–139), neoplasms of

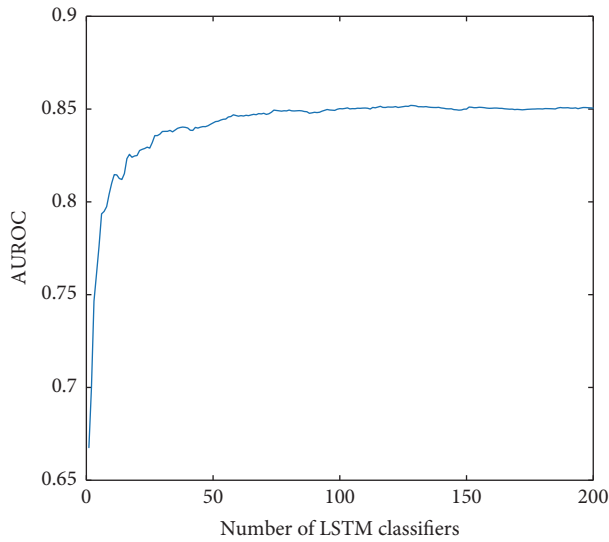


FIGURE 9: The AUROC values of eLSTM with the number of base LSTM classifiers increasing.

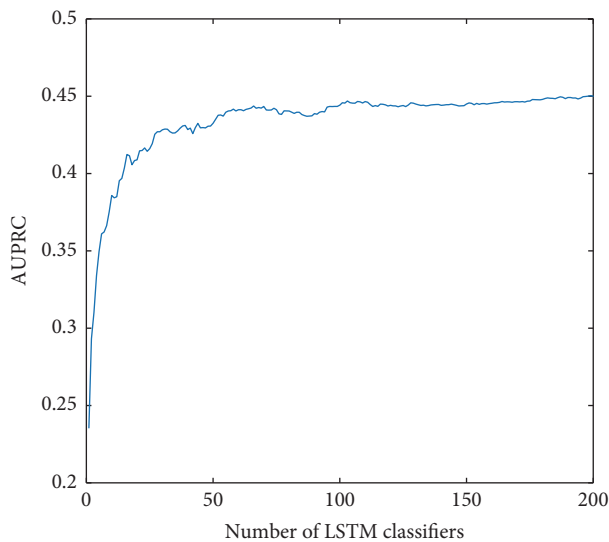


FIGURE 10: The AUPRC values of eLSTM with the number of base LSTM classifiers increasing.

digestive organs, and intrathoracic organs, etc. (ICD-9: 140–239), endocrine, nutritional, metabolic, and immunity (ICD-9: 240–279), diseases of the circulatory system (ICD-9: 390–459), pulmonary diseases (ICD-9: 460–519), diseases of the digestive system (ICD-9: 520–579), diseases of the genitourinary system (ICD-9: 580–629), trauma (ICD-9: 800–959), and poisoning by drugs and biological substances (ICD-9: 960–979). Patients admitted to ICU are usually diagnosed with more than one kind of disease, i.e., syndrome. The physiological statuses of the patients are complex, and thus, it is difficult for a single learner to discover the patterns of the patients represented by recorded parameters. Thus, in previous relevant studies, the mathematical models in ICU were usually designed for single

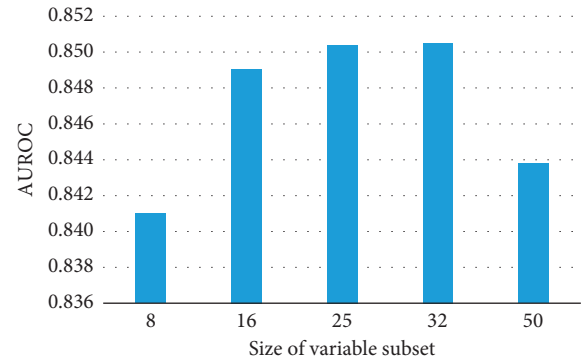


FIGURE 11: The AUROC values of eLSTM with multiple sizes of variable subset.

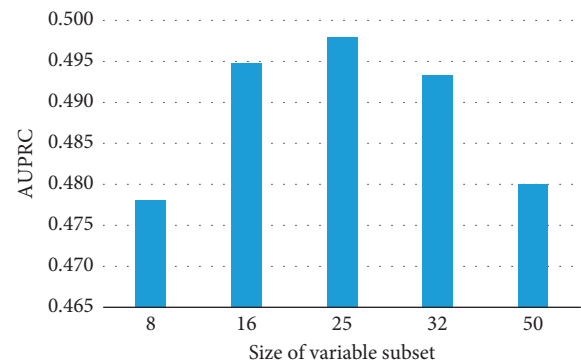


FIGURE 12: The AUPRC values of eLSTM with multiple sizes of variable subset.

specific disease, such as heart failure or sepsis [49–53], and at present, it lacks universal quantitative mortality prediction approach covering all ICU patients. The diversity of the eLSTM is accomplished by employing bagging and RSM algorithm. In the construction of base learners, bootstrap sampling and RSM ensure the learners devoting to various patients and diseases. For model training, bootstrap sampling of ICU data produces divergent datasets of patients with different disease distributions. Meanwhile, RSM assembles different sets of physiological variables for representing patients' status. These procedures in training subsets broaden views at both instance and feature level of the ICU data and therefore yield dissimilar base LSTM classifiers. In this work, the setting of 25 variables in the model brings out the best performance (Figures 11 and 12). While too few variables would greatly decrease the base learner's classifying capacity, redundant variables would damage the learners' diversity. The result is consistent with the previous finding [35]. Moreover, as part of the bagging strategy at the output end of the model, individual base learners are integrated to make the ICU patients' general condition comprehensive and clear. Owing to individual learners' classifying capacity and the ensemble learning ability of the model, the proposed eLSTM algorithm is competent for capturing the complex relationship among the diseases and parameters in ICU data, thus enhancing the outcome prediction.

5. Conclusion

In this paper, we propose a new approach named eLSTM which can deal with the complex and heterogeneous ICU data for mortality prediction. The proposed eLSTM models obtain the prediction result by merging the results of multiple parallel LSTM classifiers. The base LSTM learners are trained on different subsets which are generated using bootstrapped samples and random feature subspace. Experimental results show that the proposed eLSTM algorithm effectively utilizes the ensemble framework in LSTM classifier and achieves excellent performance on the extracted MIMIC-III dataset. Also, it provides an early prognosis of ICU patients. The eLSTM model is promising to offer a universal quantitative tool for assessing risks of all patients in ICU and even for other complex clinical situations. In the future work, other approaches of aggregating component classifiers are worth investigating to optimize the structure as well as the algorithm.

Data Availability

The data used to support the findings of this study are available at MIMIC-III website (<https://physionet.org/physiobank/database/mimic3cdb/>).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant nos. 81271662 and 81871454).

References

- [1] Y. Jin, X. Y. Cai, Y. C. Cai et al., "To build a prognostic score model containing indispensable tumour markers for metastatic nasopharyngeal carcinoma in an epidemic area," *European Journal of Cancer*, vol. 48, no. 6, pp. 882–888, 2012.
- [2] L. Minne, A. Abu-Hanna, and E. de Jonge, "Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review," *Critical Care*, vol. 12, no. 6, pp. 1–13, 2009.
- [3] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *JAMA: The Journal of the American Medical Association*, vol. 270, no. 24, pp. 2957–2963, 1993.
- [4] J.-L. Vincent, R. Moreno, J. Takala et al., "The SOFA (Sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, 1996.
- [5] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: a severity of disease classification system," *Critical Care Medicine*, vol. 13, no. 10, pp. 818–829, 1985.
- [6] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmark of deep learning models on large healthcare MIMIC datasets," 2017, <http://adsabs.harvard.edu/abs/2017arXiv171008531P>.
- [7] A. E. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: a mortality prediction case study," in *Proceedings of the Machine Learning for Healthcare Conference*, pp. 361–376, Boston, MA, USA, August 2017.
- [8] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PLoS One*, vol. 10, no. 5, Article ID e0127428, 2015.
- [9] Y. Luo, Y. Xin, R. Joshi, L. A. Celi, and P. Szolovits, "Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements," in *Proceedings of the 13th AAAI Conference on Artificial Intelligence (AAAI-16)*, pp. 42–50, Phoenix, AR USA, February 2016.
- [10] S. Kim, W. Kim, and R. W. Park, "A comparison of intensive care unit mortality prediction models through the use of data mining techniques," *Healthcare Informatics Research*, vol. 17, no. 4, pp. 232–243, 2011.
- [11] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis," *Journal of the American Medical Informatics Association*, vol. 22, no. 4, pp. 872–880, 2015.
- [12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] K. L. Caballero Barajas and R. Akella, "Dynamically modeling patient's health state from electronic medical records: a time series approach," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 69–78, Sydney, Australia, August 2015.
- [14] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the Eleventh Annual Conference of the International Speech Communication Association*, pp. 1045–1048, Makuhari, Japan, September 2010.
- [15] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE 2015 Conference on Computer Vision and Pattern Recognition*, pp. 1110–1118, Boston, MA, USA, June 2015.
- [16] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," 2016, <http://adsabs.harvard.edu/abs/2016arXiv160601865C>.
- [17] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proceedings of the AMIA Annual Symposium*, pp. 371–380, Chicago, IL, USA, November 2016.
- [18] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: predicting clinical events via recurrent neural networks," in *Proceedings of the Machine Learning for Healthcare*, pp. 301–318, Los Angeles, CA, USA, August 2016.
- [19] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," 2015, <https://arxiv.org/abs/1507.06947>.
- [20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [21] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *Computer Science*, vol. 5, p. 36, 2015.
- [22] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: a review," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 1–13, 2018.
- [23] H. Wang and D. Y. Yeung, "Towards Bayesian deep learning: a survey," 2016, <https://arxiv.org/abs/1604.01662>.
- [24] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the Neural*

- Information Processing Systems 2014*, pp. 3104–3112, Montreal, Canada, December 2014.
- [25] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, <https://ui.adsabs.harvard.edu/abs/2014arXiv1409.0473B>.
- [26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: a neural image caption generator,” 2014, <http://adsabs.harvard.edu/abs/2014arXiv1411.4555V>.
- [27] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.
- [28] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, “Learning to diagnose with LSTM recurrent neural networks,” 2015, <http://adsabs.harvard.edu/abs/2015arXiv151103677L>.
- [29] Y. Jo, L. Lee, and S. Palaskar, “Combining LSTM and latent topic modeling for mortality prediction,” 2017, <http://adsabs.harvard.edu/abs/2017arXiv170902842J>.
- [30] H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan, “Multitask learning and benchmarking with clinical time series data,” 2017, <http://adsabs.harvard.edu/abs/2017arXiv170307771H>.
- [31] T. Pham, T. Tran, D. Phung, and S. Venkatesh, “DeepCare: a deep dynamic memory model for predictive medicine,” in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 30–41, Auckland, New Zealand, April 2016.
- [32] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, “Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach,” *International Journal of Medical Informatics*, vol. 108, pp. 185–195, 2017.
- [33] T. Ma, C. Xiao, and F. Wang, “Health-ATM: a deep architecture for multifaceted patient health record representation and risk prediction,” in *Proceedings of the SIAM International Conference on Data Mining, 2018*, pp. 261–269, San Diego, CA, USA, May 2018.
- [34] T. G. Dietterich, “Ensemble methods in machine learning,” in *Proceedings of the International Workshop on Multiple Classifier Systems*, pp. 1–15, Cagliari, Italy, June 2000.
- [35] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
- [36] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, Boca Raton, FL, USA, 2012.
- [37] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [38] L. Nanni, A. Lumini, and S. Brahnam, “A classifier ensemble approach for the missing feature problem,” *Artificial Intelligence in Medicine*, vol. 55, no. 1, pp. 37–50, 2012.
- [39] A. Ozcift and A. Gulten, “Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms,” *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. 443–451, 2011.
- [40] A. Özçift, “Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis,” *Computers in Biology and Medicine*, vol. 41, no. 5, pp. 265–271, 2011.
- [41] H. Chen, S. Yuan, and K. Jiang, “Wrapper approach for learning neural network ensemble by feature selection,” in *Proceedings of the International Symposium on Neural Networks*, pp. 526–531, Chongqing, China, June 2005.
- [42] P. H. Abreu, H. Amaro, D. C. Silva et al., “Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data,” in *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013*, pp. 1366–1369, Springer, Cham, Switzerland, 2014.
- [43] H. Kim, H. Kim, H. Moon, and H. Ahn, “A weight-adjusted voting algorithm for ensembles of classifiers,” *Journal of the Korean Statistical Society*, vol. 40, no. 4, pp. 437–449, 2011.
- [44] A. E. W. Johnson, T. J. Pollard, L. Shen et al., “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, no. 1, p. 160035, 2016.
- [45] E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: bagging, boosting, and variants,” *Machine Learning*, vol. 36, pp. 105–139, 1999.
- [46] L. Breiman, “Using iterated bagging to debias regressions,” *Machine Learning*, vol. 45, no. 3, pp. 261–277, 2001.
- [47] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 249–256, Chia Laguna Resort, Italy, May 2010.
- [48] J. Twisk and W. de Vente, “Attrition in longitudinal studies: how to deal with missing data,” *Journal of Clinical Epidemiology*, vol. 55, no. 4, pp. 329–337, 2002.
- [49] T. Lagu, P. K. Lindenauer, M. B. Rothberg et al., “Development and validation of a model that uses enhanced administrative data to predict mortality in patients with sepsis,” *Critical Care Medicine*, vol. 39, no. 11, pp. 2425–2430, 2011.
- [50] C. Vorwerk, B. Loryman, T. J. Coats et al., “Prediction of mortality in adult emergency department patients with sepsis,” *Emergency Medicine Journal*, vol. 26, no. 4, pp. 254–258, 2009.
- [51] B. Steinhart, K. E. Thorpe, A. M. Bayoumi, G. Moe, J. L. Januzzi, and C. D. Mazer, “Improving the diagnosis of acute heart failure using a validated prediction model,” *Journal of the American College of Cardiology*, vol. 54, no. 16, pp. 1515–1521, 2009.
- [52] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, “Using recurrent neural network models for early detection of heart failure onset,” *Journal of the American Medical Informatics Association*, vol. 24, pp. 361–370, 2016.
- [53] X. Fu, Y. Ren, G. Yang et al., “A computational model for heart failure stratification,” in *Proceedings of the 2011 Computing in Cardiology*, pp. 385–388, Hangzhou, China, September 2011.