Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6147037, 13 pages
https://doi.org/10.1155/2020/6147037

Research Article

Discriminating the Nature of Thyroid Nodules Using the Hybrid Method

Hongjun Sun,[1] **Feihong Yu**,[2] **and Haiyan Xu**[1]

[1]*College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*
[2]*Department of Ultrasound, The First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China*

Correspondence should be addressed to Hongjun Sun; shj_hust@126.com

Received 17 February 2020; Revised 21 May 2020; Accepted 7 July 2020; Published 7 August 2020

Academic Editor: Bogdan Smolka

Copyright © 2020 Hongjun Sun 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.

Prompt and correct diagnosis of benign and malignant thyroid nodules has always been a core issue in the clinical practice of thyroid nodules. Ultrasound imaging is one of the most common visualizing tools used by radiologists to identify the nature of thyroid nodules. However, visual assessment of nodules is difficult and often affected by inter- and intraobserver variabilities. This paper proposes a novel hybrid approach based on machine learning and information fusion to discriminate the nature of thyroid nodules. Statistical features are extracted from the B-mode ultrasound image while deep features are extracted from the shear-wave elastography image. Classifiers including logistic regression, Naive Bayes, and support vector machine are adopted to train classification models with statistical features and deep features, respectively, for comparison. A voting system with certain criteria is used to combine two classification results to obtain a better performance. Experimental and comparison results demonstrate that the proposed method classifies the thyroid nodules correctly and efficiently.

1. Introduction

Thyroid nodules are exceedingly common, with a reported prevalence of more than twenty percent of men and fifty percent of women over 50 years old on high-resolution ultrasound [1]. Some of these nodules are benign and others are malignant. Pathological analysis shows that most nodules are benign. Studies have reported that the prevalence of thyroid cancer is increasing at a rate of 3% per year [2]. Therefore, doctors must distinguish the nature of thyroid nodules in order to make correct clinical decisions. Benign nodules can usually be followed up and followed, and malignant nodules may require immediate surgical treatment to achieve a definitive diagnosis. Currently, fine-needle aspiration (FNA) is the most effective and practical test to determine whether a nodule is malignant or may require surgery [3]. However, most nodules are benign, and even malignant nodules, especially nodules smaller than 1 cm, frequently exhibit indolent or nonaggressive behavior [4]. Therefore, not all detected nodules require FNA or surgery.

Several imaging techniques, including CT, magnetic resonance imaging (MRI), and ultrasound, have been used to discriminate the nature of thyroid nodules in the clinic. Ultrasound is the most commonly used because it is expedient, efficient, inexpensive, noninvasive, and nonradioactive [5]. As shown in Figure 1, ultrasound has two kinds of modalities: B-mode ultrasound (B-US) and shear-wave elastography ultrasound (SWE-US). Many studies have proved that B-US can describe the size, shape, location, and texture of nodules so as to distinguish malignant nodules from benign nodules [6]. Compared with B-US, SWE-US is a novel ultrasound technique that quantitatively represents the stiffness of tissues by assessing the deformability of thyroid nodule, which can assist diagnosis [7, 8]. Nevertheless, perception and understanding of ultrasound images by expert clinicians is usually subjective. They interpret and diagnose ultrasound images based on their experience and knowledge, which may make the diagnosis of the same case inconsistent due to different clinicians. This seriously affects the accuracy of sonography.

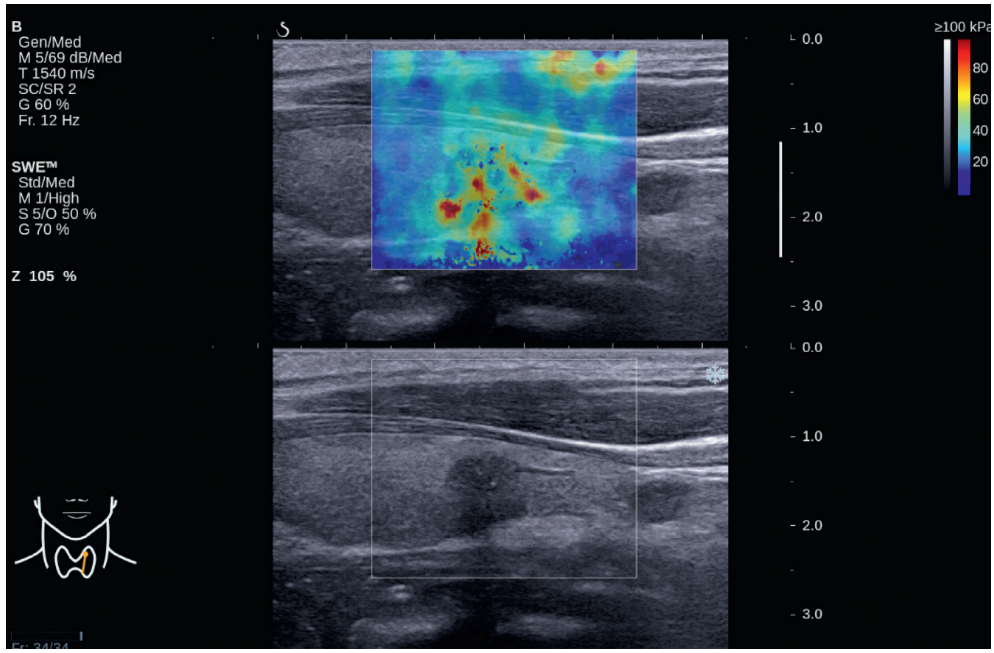

Figure 1: Examples of visualization about B-US image (bottom, grayscale) and SWE-US image (top, color). The regions of interest are marked with rectangles. The color bar on the right indicates the elastic modulus of nodules, which decreases from red to blue.

Tremendous advances have been made in medical imaging and artificial intelligence technologies, which make computer-assisted diagnostics (CAD) become increasingly widespread. CAD can help us solve subjective diagnostic problems based on objective criteria, which traditionally depends on the experience of radiologists. CAD on ultrasound imaging commonly uses the statistical features (SFs) which are also called radiomics data of medical images, including morphological parameters, intensity statistics, and texture features quantifying heterogeneity [9, 10]. However, the disadvantages of ultrasound images, such as image noise and low contrast, and changes in the shape, size, and traces of thyroid nodules limit these statistical features to work well in classification tasks.

Recently, deep learning models, especially convolution neural networks (CNNs), have received great attention in image classification and target recognition [11]. CNNs can also extract image features, which can be regarded as deep hierarchical representations of the inputs so that implicit features within the image can be well captured. These advanced deep features (DFs) are exactly what we need to complement the primary features because CNNs are designed to capture the intrinsic features [12].

In our study, we propose a hybrid approach combining models trained with traditional features extracted from B-US images and deep features extracted from SWE-US images for thyroid nodule classification task. Firstly, we employ a pretrained CNN model, which is transfer learned from ImageNet, as a feature extractor to draw deep features from SWE-US image dataset. To obtain better performance, we compare the classifiers trained with features extracted from each layers of CNNs to find the most discriminative classifier for the classification task. Then, traditional features are extracted from the corresponding B-US image dataset. Different classifiers are used to train with SFs and DFs for comparison. A voting system including pessimistic, optimistic, and compromise criteria is designed and conducted to combine predictive results from different classifiers together to obtain a better classification performance.

The main contributions of this work are as follows:

(1) We propose a novel hybrid framework combining multimodality features for thyroid nodule classification.

(2) The classifiers trained with features extracted from each layers of CNNs are compared to find the most discriminative classifier for the nodule classification task.

(3) The performance of different decision-making strategies on the classification results is compared and analyzed, and reasonable suggestions are put forward.

The remainder of this paper is organized as follows. Background knowledge is summarized and related literatures are reviewed in Section 2. Section 3 describes the framework of proposed method and introduces the materials and method used in our model. The experimental results and model parameters are presented in Section 4. Discussions are drawn in Section 5. Finally, the paper concludes with some comments in Section 6.

2. Related Works

2.1. Ultrasound Application in Thyroid Nodule Diagnosis. Ultrasound is a combination of acoustics, medicine, optics, and electronics. It covers a wide range of applications, including ultrasound diagnosis, ultrasound therapy, and biomedical ultrasound engineering, and is of great value in the prevention, diagnosis, and treatment of diseases.

Ultrasound imaging uses an ultrasound beam to scan the human body and receives and processes the reflected signal to obtain an image of the internal organs. Ultrasound commonly used in medical imaging diagnosis includes B-mode ultrasound (B-US, Figure 1, bottom) and shear-wave elastography ultrasound (SWE-US, Figure 1, top).

B-US can describe the size, shape, location, and texture of nodules so as to distinguish malignant nodules from benign nodules. In the literature [10], texture features of the B-US image are extracted, and the classification model is trained using SVM to divide the ultrasound image into two categories, benign and malignant. On the basis of extracting the texture features, Raghavendra et al. extracted higher order spectral (HOS) entropy features as supplements, and the particle swarm algorithm and SVM are combined in the model training process to improve the accuracy of classification [13]. Statistical features such as texture and contour are fundamental features, which are greatly affected by image quality and noise. These defects limit the development of intelligent diagnostic algorithms based on these statistical features.

Buda et al. have trained a multitask deep convolutional neural network and compared it with a consensus of three ACR TI-RADS committee experts and nine other radiologists, and the results show the performance of deep learning algorithm is similar to the diagnosis of experts [14]. Mei et al. extracted deep features of convolutional autoencoders and fundamental features including local binary patterns (LBP) as well as histogram of oriented gradients (HOG) descriptors in association with medical professional thyroid image characterization from B-US and trained the classifiers using these features to improve negative predictive value of thyroid nodule evaluation [15]. Comparison has been done between radiomics-based and deep learning-based approaches, and the results demonstrate that the deep learning-based method achieves a better performance [16, 17]. Deep learning in conjunction with B-US image characterization could improve nodule characterization and reduce benign biopsies. These advanced semantic features extracted with deep models are exactly a complement to fundamental features, since deep models are not intended to carry out classification task, but to learn to capture the intrinsic characteristics of ultrasound images.

SWE-US is a new technology based on the basic properties of biological tissue with elasticity or hardness, with the advantages of measurement results not being affected by the operator and excellent repeatability. Elastography can significantly improve the differential diagnosis of benign and malignant nodules of the thyroid. [18, 19]. Zhang et al. built a two-layer deep learning architecture for automated extraction of features from the shear-wave elastography and evaluated the deep learning architecture in differentiation between benign and malignant breast tumors [20]. Liu et al. extracted features from multimodality images including B-US and SWE-US and trained a SVM model to discriminate the thyroid tumor with LN metastasis [21]. Image features from SWE-US, including fundamental features and advanced features, can also be extracted to train the classification model for intelligent diagnosis. Comprehensive use of B-US and SWE-US multimodality images, as is used in this paper, can effectively improve the accuracy of model results.

2.2. CNNs and Transfer Learning. CNNs are the most commonly used deep learning model, from which the high-level characteristics can be extracted. CNNs can be used for tasks such as object detection, classification, and also feature extraction. It is a kind of feed-forward neural network, which is a multilayered perceptron inspired by biological thinking. CNN has different layers, and the working methods and functions of each layer are also different [12]. It has been proved to be a method with numerous uses, from segmentation of symptom to tumor diagnosis in medicine [22]. But their use may be unfeasible in many situations since they require very large training sets (from thousands up to several million images). To overcome this difficulty, the common approach in the literature consists of applying transfer learning.

Transfer learning means the ability of a system to recognize and apply knowledge learned in previous tasks to a novel task [23]. The definition of transfer learning is given in terms of domain and task. The domain D consists of a feature space X and a marginal probability distribution $P(X)$, where $X = \{x_1, \ldots, x_n\}$. Given a specific domain, $D = \{X, P(X)\}$, a task consists of two components: a label space Y and an objective predictive function $f(\cdots)$ (denoted by $T = \{Y, f(\cdot)\}$, which is learned from the training data consisting of pairs, which consist of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$. The function $f(\cdot)$ can be used to predict the corresponding label, $f(x)$, of a new instance x. It is natural to use the transfer learning method to apply the knowledge gained while solving the problem of natural image recognition to solve a different problem of medical images classification.

Consider the following two facts: firstly, the scope of the ultrasound image dataset (hundreds or thousands) is much smaller than the natural image dataset (more than millions); secondly, two datasets consist of images from completely different regions. That is, the data distribution of these two datasets is inconsistent. There are two typical uses of transfer learning in the field of medical image classification: one is to remove the last fully connected layer on the top of the pretrained deep model and treat the rest of the network as a fixed feature extractor for the current dataset; the other one is that we adjust the transfer learning method by fixing most earlier layers to reserve generic information and only retraining from scratch the last fully connected layer of the pretrained deep model to capture domain-specific features.

3. Material and Methods

3.1. Overview. In this paper, we propose to evaluate the hybrid approach with multimodalities of ultrasound imaging in discriminating the nature of thyroid nodules. The algorithm deals with the two modalities separately. For deep features, we compare the classifiers trained with features extracted from each layers of CNNs to find the most

discriminative one for the task. For statistical features, the process generally contains 3 steps: image preprocessing, feature extraction, and feature selection. Then, different classifiers are adopted to train classification models with statistical features and deep features for comparison. In the end, two classification models hybridize together with a voting system, employing three kinds of decision criteria. The hybrid model is observed to obtain a better performance. Overall framework of this research is shown in Figure 2. Details of each process are as follows.

3.2. B-US Processing and Statistical Features. B-US image is gray scale image (Figure 1, bottom) which can display the position and shape of the thyroid nodule. The clinically acquired ultrasound thyroid images have low quality, which are mainly reflected in the problems of severe speckle noise, blurred nodule edges, discontinuous boundaries, and low contrast. This paper uses a nonlinear filter for noise reduction of the original ultrasound image. A nonlinear filter combines the spatial proximity and pixel value similarity of the image while comprehensively considering the spatial information and gray similarity. It has the following advantages: first, it can keep the output signal sequence unchanged; then, it can reduce the interference of random noise and impulse noise; in addition, it can well retain the edge information. Therefore, a nonlinear filter is very suitable for ultrasonic image preprocessing, which has been proven in previous research [10]. It can not only eliminate image noise and blur caused by uneven ultrasonic echo but also retain the edge information of nodules.

After ultrasound images are denoised by the median filter, the regions of interest (ROIs) are manually segmented along the nodule contour on each transverse section using an open-source imaging platform named ITK-SNAP. In order to eliminate the difference, the segmentation is carried out by the radiologist with more than 5 years of experience in continuous time. Figures 3(a) and 3(b) show denoised nodule images, whereas Figures 3(c) and 3(d) show the ROIs segmented on sonograms.

A python radiomics package named "Pyradiomics" is used to automatically extract the statistical features from the nodule region, which is outlined by radiologists. A total of 104 dimensional statistics features including first-order statistics features, shape features, gray level co-occurrence matrix- (GLCM-) based features, gray level run length matrix- (GLRLM-) based features, gray level size zone matrix- (GLSZM-) based features, neighboring gray tone difference matrix- (NGTDM-) based features, and gray level dependence matrix- (GLDM-) based features are obtained. The dimensions of all kinds of features are shown in Table 1. Refer to the supplementary material (available here) for a detailed description of segmentation and extraction tools and features.

The purpose of feature selection is to select a subset of the smallest features based on the original value of the dataset instead of removing the irrelevant and redundant attributes, which may increase the complexity of classification and even cause the performance of the classifier to decrease [24]. Therefore, feature with good discrimination can be selected as the results of feature selection. We conduct two methods for feature selection of SFs extracted from ultrasound images and compare their results: one is the principal component analysis (PCA) and the other is the t-test method.

When using PCA, the feature vector is gradually increased at an interval of 10 as an input to determine the optimal number of retained components so that colleagues who retain the data structure information to the greatest extent can reduce the dimension, as elaborated in Algorithm 1.

When using the t-test, set the threshold of the p value and select features with p values smaller than the threshold and input them to the classifier. The feature selection method based on t-test is as follows:

$$t = \frac{\overline{X}_A - \overline{X}_B}{SE\left(\overline{X}_A - \overline{X}_B\right)}, \tag{1}$$

where X_A and X_B represent the malignant and benign samples of the same feature, respectively. $\overline{X}_A$ and $\overline{X}_B$ are the mean of the samples, and $SE(\overline{X}_A - \overline{X}_B)$ is standard error of the difference.

3.3. SWE-US Processing and Deep Features. SWE-US image is considered to be a composite color image (Figure 1, top) superimposed on the corresponding ultrasound grayscale image (Figure 1, bottom). By subtracting the ultrasound image from the synthetic color image, a pure SWE-US image can be obtained. A SWE-US dataset on full-field digital ultrasound image is composed of labeled elastic ultrasound images. A region of interest (ROI) about each nodule has been extracted within each image. In order to extract the ROI from the elastic ultrasound image, the radiologist labeled the center of each nodule, and then a $448 * 448$ box was automatically cropped around the nodule center with a pixel size of 0.1 mm. The ROIs were marked as either benign or malignant according to the pathological analysis reports. As a result, ROIs were converted from the image matrix to a pixel vector, which could be directly used as the inputs of the CNNs.

As mentioned above, the application of transfer learning method for medical image classification is in our research is feature extraction, which removes the last fully connected layer of the pretrained deep model because the output of this layer is for the class score of multiclass classification tasks like ImageNet. The rest of the network acts as a feature extractor for the given dataset. A variety of pretrained models such as resNet-50, Inception-V3, and VGGNet-16 have been used for transfer learning [25, 26]. In this work, we employ the VGGNet-16 model trained from ImageNet to extract features, which are used to train a two-class classifier (e.g., SVM). VGGNet-16 is a model proposed by Oxford University in 2014, and the network structure is shown in Figure 4 [27]. It consists of 5 max pooling layers, 3 fully connected layers, and 1 softmax function. After removing the network part after the last fully connected layer, the output of each remaining layer can be regarded as the input for training model.

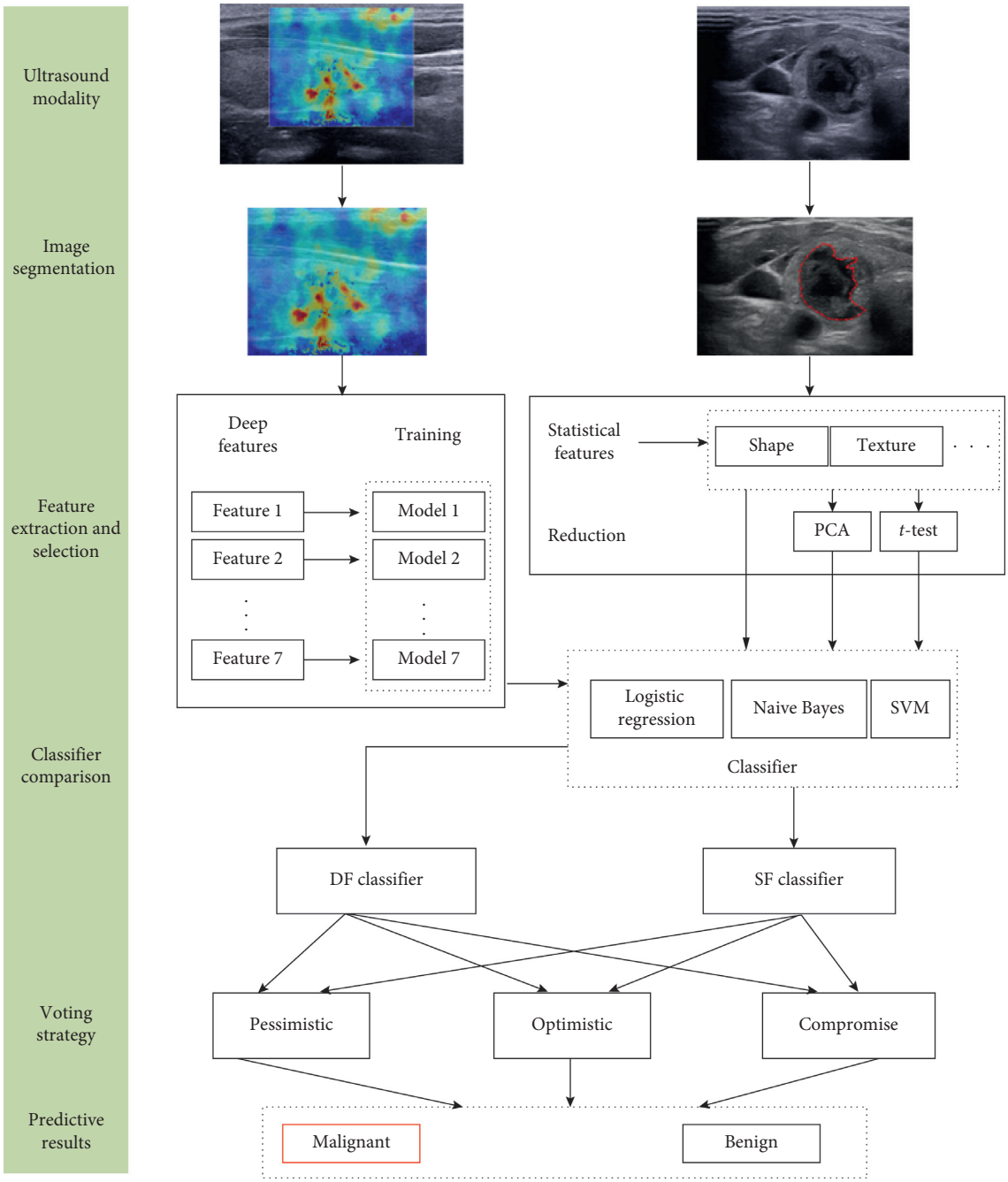

Figure 2: Schematic diagram of our model architecture for thyroid nodule classification.

However, for the features output from the second fully connected layer, it may be the result of various functional combinations. Therefore, in order to obtain the best performance of the nodule classification task, we compare the features extracted from Pool1–Pool5 and FC6 and FC7 layers of the VGG network. Features from each layer are used as sets of input for a classifier after zero-variance removal. It is worth noting that the ROI of SWE-US has three channels to accommodate the original architecture of VGGNet-16, which is designed for colorful images. In addition, the sampling rate of ROIs is reduced by half to $224 * 224$ to accommodate the architecture of the network.

The results are shown in Figure 5 to explain how an image responds to a certain convolution layer. It can be seen from the figure that the results derived from the lower layers can better extract shallow features, including edge, direction, and intensity features. However, in the images output from the last few layers, various features will appear mixed together.

3.4. Classifiers. The diagnosis of benign and malignant thyroid nodules in this paper is a typical two-class problem. Many classifiers including logistic regression (LR), k-nearest neighbor (KNN), random forest (RF), support vector machine (SVM), and so on have been used to discriminate the

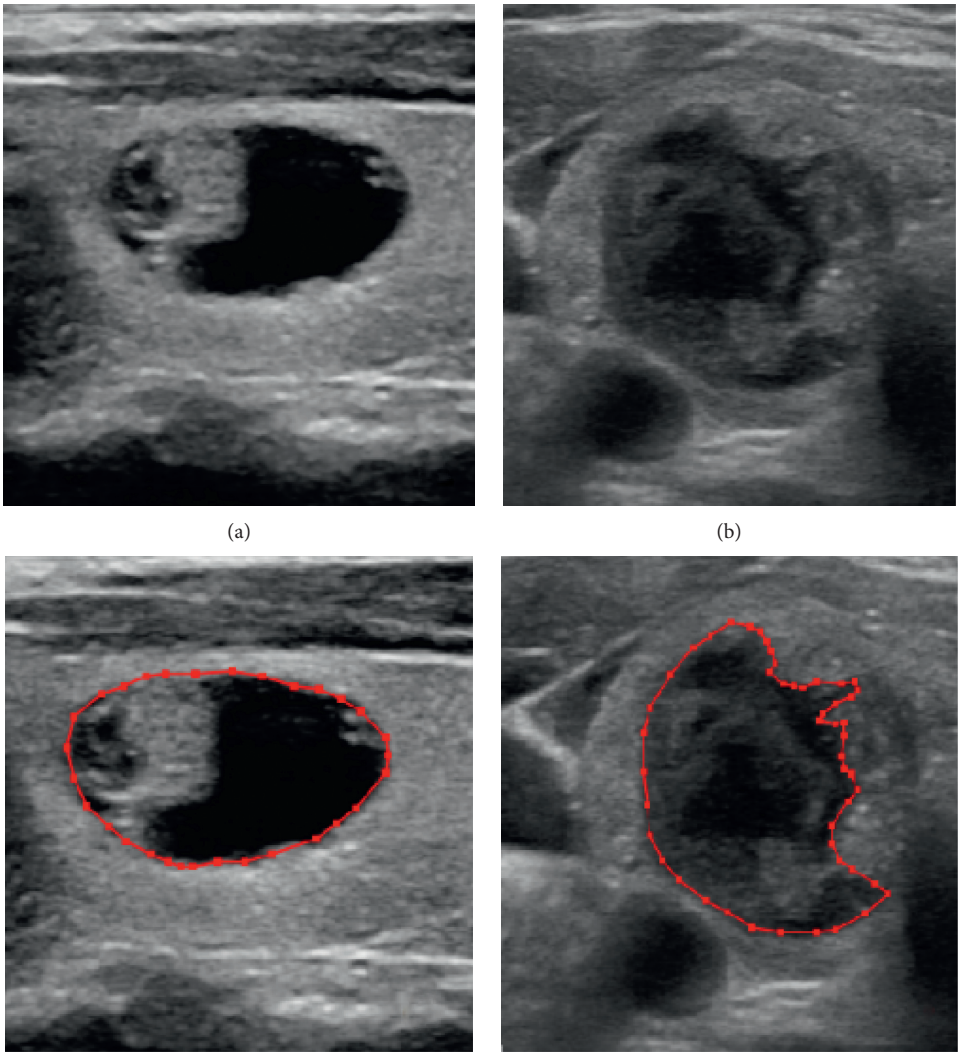

FIGURE 3: Region of interest selection. (a, b) Benign and malignant nodule images. The outline of the nodule is outlined with red lines in (c) and (d).

TABLE 1: Summary of the statistical features.

Id	Feature type	Dimension
1	First-order statistics	19
2	Shape	10
3	Gray level co-occurrence matrix (GLCM)	24
4	Gray level run length matrix (GLRLM)	16
5	Gray level size zone matrix (GLSZM)	16
6	Neighboring gray tone difference matrix (NGTDM)	5
7	Gray level dependence matrix (GLDM)	14
	Total	104

nature of nodules based on features extracted from ultrasound or CT images. We conduct LR, Naive Bayes, and SVM algorithms for comparison in the classification process because they could output the probability value, which were used as the input of voting system for decision fusion. Suppose x is the input vector and y is the label. The general description and mathematical formula of the classification algorithms are as follows.

Logistic regression is a machine learning method used to solve binary classification problems and is used to estimate the probability. The hypothetical function of logistic regression is as follows:

$$y = g\left(\theta^T x\right),$$
$$g(z) = \frac{1}{1 + e^{-z}}, \tag{2}$$

where θ is weight vector which can be chosen through data training.

The Naive Bayes method is a set of supervised learning algorithms based on Bayes' theorem, and it is assumed that each pair of features is independent of each other. The principle is as follows:

$$P\left(y \mid x_1, \ldots, x_n\right) = P(y) \frac{\prod_{i=1}^{n} P\left(x_1 \mid y\right)}{P\left(x_1, \ldots, x_n\right)}. \tag{3}$$

(i) Input: n dimensional dataset $D = (x^{(1)}, x^{(2)}, \ldots, x^{(m)})$
(ii) Output: D' dataset after dimensionality reduction
(iii) Centralize all samples $x^{(i)} = x^{(i)} - \sum_{j-1}^{m} x^{(j)}/m$
(iv) Calculate sample covariance matrix XX^T
(v) Eigenvalue decomposition of the matrix XX^T
(vi) Get the eigenvectors $(\omega_1, \omega_2, \ldots, \omega_k)$ corresponding to the largest k eigenvalues
(vii) Normalize all eigenvectors to form an eigenvector matrix W
(viii) Convert samples $x^{(i)}$ to new samples $z^{(i)} = W^T x^{(i)}$
(ix) Get the output sample set $D\prime = (z^{(1)}, z^{(2)}, \ldots, z^{(m)})$
(x) End

Algorithm 1: Feature selection with PCA.

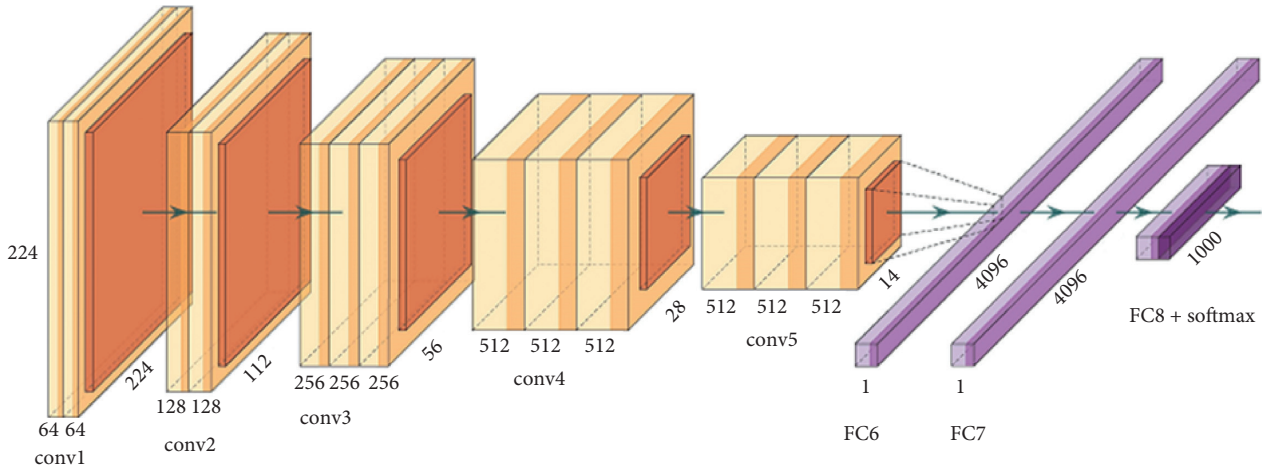

Figure 4: VGGNet-16 network structure diagram. Prior to conv5, the yellow block is the output of the convolutional layer and the orange block is the output of the pooling layer. FC is the output of the fully connected layer.

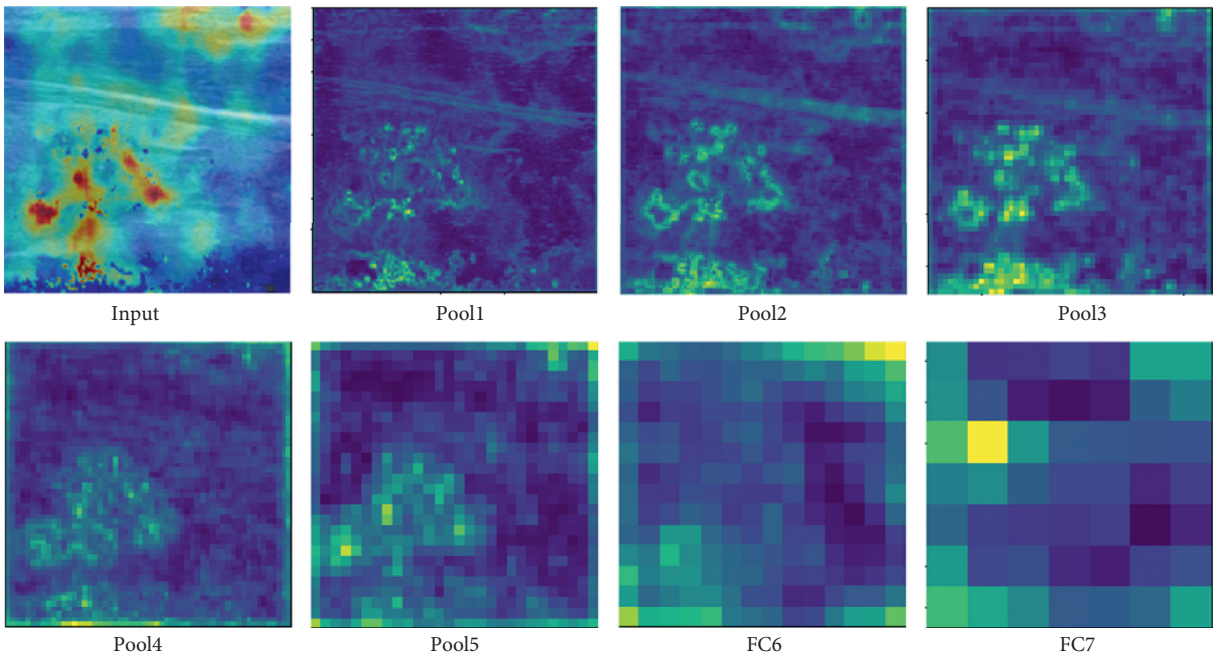

Figure 5: Response of certain convolution layers.

SVM is a convex quadratic programming problem, which can be expressed by the following mathematical formula:

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\xi_i, \\ \text{s.t.} \quad & y_i\left(\omega x_i + b\right) \geq 1 - \zeta_i, \\ & \xi_i \geq 0, \\ & i = 1, 2, \ldots, N, \end{aligned} \tag{4}$$

where ω is the weight vector, b is the bias vector, ξ is the slack variable, and C is the penalty parameter. Besides, x_i is the ith feature and y_i is the label of x_i.

3.5. Voting System. The voting system receives the probability of benign and malignant computed by two classifiers. The combination of the two outputs is responsible for increasing the final accuracy for each modality. The voting system proposed in our research is an adaptation of the uncertainty decision theory [28], where we use a series of combination rules called pessimistic criteria, optimistic criteria, and compromise criteria. It can be represented by the following mathematical expressions.

For a thyroid nodule, (θ_1, θ_2) represent the probability of benign and malignant computed by the transform learning method, and the probability of the model trained by statistical features is (ι_1, ι_2). Define the label of benign and malignant as r. Note that $0 < \theta_1 < 1$, $0 < \theta_2 < 1$, and $\theta_1 + \theta_2 = 1$. If $\theta_1 > \theta_2$, $r = 1$; otherwise $r = 0$. Similarly, $0 < \iota_1 < 1$, $0 < \iota_2 < 1$, and $\iota_1 + \iota_2 = 1$. If $\iota_1 > \iota_2$, $r = 1$; otherwise, $r = 0$.

(1) Define pessimistic criteria $P(\theta, l)$, $P(\theta, l) = (p_1, p_2) = (\max(\theta_1, l_1), \min(\theta_2, l_2))$. Note that $0 < p_1 < 1$, $0 < p_2 < 1$, and $p_1 + p_2 = 1$. If $p_1 > p_2$, $r = 1$; otherwise, $r = 0$. That is, if the prediction result of at least one classifier is malignant, then the output of the voting system is malignant. Only if the prediction results of both classifiers are benign, the output of the voting system is benign.

(2) Define optimistic criteria $O(\theta, l)$, $O(\theta, l) = (o_1, o_2) = (\min(\theta_1, l_1), \max(\theta_2, l_2))$. Note that $0 < o_1 < 1$, $0 < o_2 < 1$, and $p_1 + p_2 = 1$. If $o_1 > o_2$, $r = 1$; otherwise, $r = 0$. According to the optimistic decision criterion, only two classifiers consider it to be malignant; then, the voting result is malignant; otherwise, it is considered benign.

(3) Define compromise criteria $C(\theta, l)$, $C(\theta, l) = (c_1, c_2) = ((1/2)(\theta_1 + l_1), (1/2)(\theta_2 + l_2))$. Note that $0 < c_1 < 1$, $0 < c_2 < 1$, and $c_1 + c_2 = 1$. If $c_1 > c_2$, $r = 1$; otherwise, $r = 0$. The weighted average of the two classifier predictions is considered, where we consider the weights of the two classifiers to be equal. After the weighted average result, if the benign probability is greater than the malignant probability, the output of voting system is benign; otherwise, it is malignant.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Preparation of Dataset. Herein, the experimental data are obtained from the Department of Ultrasound, First Affiliated Hospital of Nanjing Medical University. The study population is composed of 245 patients. Both B-US and SWE-US examinations are performed by experienced radiologists. Images are acquired and stored in DICOM standards. The type of thyroid nodules is the gold standard for pathological analysis including excisional biopsy, core needle biopsy, or FNA biopsy. When a patient undergoes multiple biopsies, the gold standard for final diagnosis will be determined according to the following priorities: excisional biopsy, core needle biopsy, and FNA biopsy. There are 490 images in total (B-US and SWE-US each account for half), consisting of 145 images of benign nodules and 100 images of malignant nodules. This retrospective study was approved by the institutional review board, and the informed consent was obtained from all patients.

4.1.2. Cross Validation. In order to improve the generalization ability of the model on the dataset, five-fold cross validation is executed during the training and testing process of the model. The original data are evenly divided into 5 groups, and each subset of the data is used as a validation set, and the remaining 4 sets of subset data are used as the training set. Repeating 4 times will get 4 models. The average of the classification accuracy of the four models is used as the performance index of this classifier. As listed in Table 2, each set of data includes the number of cases, the number of malignant nodules, and the nodule radius.

4.1.3. Evaluation Criteria. Herein, quantitative evaluation indexes such as accuracy, sensitivity, and specificity, which are usually used in medical diagnosis, are adopted to evaluate the classification quality. Accuracy is computed by equation (5), and it is the metric how exactly the given thyroid nodules are classified into the right type. Sensitivity is calculated by equation (6), and this metric is used to retrieve the exact malignant nodules from all the gathered malignant nodules. Specificity is calculated by equation (7). This metric is used to retrieve the exact benign nodules from all the gathered benign nodules.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \tag{5}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \tag{6}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%, \tag{7}$$

Table 2: Partitioning and statistical summary of cross-validated datasets.

Subset	Number	Malignancy	Radius (mm)
Subset 1	49	21	0.51 + 0.26
Subset 2	49	19	0.56 + 0.17
Subset 3	49	17	0.53 + 0.24
Subset 4	49	20	0.58 + 0.21
Subset 5	49	23	0.52 + 0.22
Total	245	100	0.54 + 0.2

where TP means true positive while TN means true negative. TP and TN represent the positive and negative sample numbers of correct classification. Conversely, FP is false positive while FN is false negative. They represent the negative and positive sample numbers of false classification. In this research, positive indicates that the type of nodule is malignant and vice versa. Sensitivity defines the possibility of predicting the malignant nodules while specificity defines the possibility of predicting the benign nodules.

4.2. Comparison Results

4.2.1. Comparison Results of CNNs. Features extracted from different layers of CNNs are compared to find the most discriminable one for the nodule classification task in our research. The results of feature extraction and processing are shown in Figure 6. The left column is outputs of images after convolution and pooling. The center column indicates the dimensionality obtained after flattening the output of each layer. The right column represents the length of the feature vector per ROI used as an input for the classifier after zero-variance removal.

Considering that the features are high-dimensional vectors and training samples are limited, we choose a linear kernel SVM as the classifier. Totally 7 SVM classification models are trained. Table 3 shows the performance of 7 classifiers trained on CNN features, which are extracted from the experimental dataset through all pool layers and fully connected layers of the VGGNet-16 network. Pool$i$ represents deep features extracted from the ith pooling layer. Similarly, FCj represents deep features extracted from the jth fully connected layer. Optimize parameters in SVM by using grid search. Derive a malignancy probability from the SVM and choose a threshold of 0.5 to assign the samples to malignant or benign.

Based on the results in Table 3, it is obvious that the performance increases in the beginning of 5 pooling layers until it reaches the best at Pool5 layer; after that, performance begins to decrease slightly. Sharp drops in performance turn out after the FC6 layer. Besides, the length of features has a greater influence on the training time. The training of high-dimensional features obviously takes more time, and the effect of feature length on the testing time is less. One possible reason is that a linear kernel is selected for SVM, and the time consumption of the optimization solution in the training process is related to the number of samples and feature dimensions, which have a limited impact on the time consumption of linear operations in the testing process.

As a result, considering the balance between high predictive performance and relatively low dimensionality, we choose the FC6 layer as the optimal layer to extract deep features. The dimensionality of the feature vector of the pooling layer is one or two orders of magnitude higher than that of the fully connected layer, which greatly increases the computational cost.

4.2.2. Comparison Results of Various Techniques. Finally, multiple sets of comparisons with various techniques are made in our work. First, different types of features are compared, and the SFs from B-US and the DFs from SWE-US are input to the same classifiers for comparison. Second, different feature selection methods are compared. The results of direct training without feature selection are compared with the results of training after feature selection using the PCA and t-test methods, respectively. Third, different classifiers are compared. SVM is compared with the other two classic classifiers, Naive Bayes and logistic regression. Optimize parameters in SVM by using grid search. Derive a malignancy probability from the classifiers and choose a threshold of 0.5 to assign the samples to malignant or benign. For the Naive Bayes method, it is assumed that each pair of features is independent of each other. In the logistic regression algorithm, we use the sigmoid function. The detailed experimental results are shown in Table 4.

Experimental results show that the best performance of models trained with DFs is better than that of models trained with statistical features. According to Table 4, the best classification result of SFs is from SF-PCA-SVM, whose accuracy is 0.804, while the best classification result of DFs is from FC6-SVM, whose accuracy is 0.812. In addition, the classification accuracy of LR and Naive Bayes is similar, while that of SVM is the highest. Especially for DFs, the performance of SVM is significantly better than LR and Naive Bayes indicating that SVM is better at high-dimensional data classification. Of course, SVM takes the longest training and testing time. Furthermore, experiments show that the results after feature selection are better than those without feature selection. In terms of feature selection, PCA performs better than t-test.

It should be noted that when using PCA or t-test for feature reduction of SFs, the classification results listed in Table 4 were obtained with the optimal numbers of reserved features. We find that 30 principal components are best for LR, 20 principal components are best for Naive Bayes, and 50 principal components are best for SVM. The criterion of $p < 0.01$ is better than $p < 0.05$ for feature selection using t-test.

4.3. Voting Results. As previously mentioned, classification models of different modalities are trained, respectively, to compare the performance. According to the experimental results, we choose the predictive results from FC6-SVM and SF-PCA-SVM as the inputs of the voting system which is explained in Section 3. The voting system has

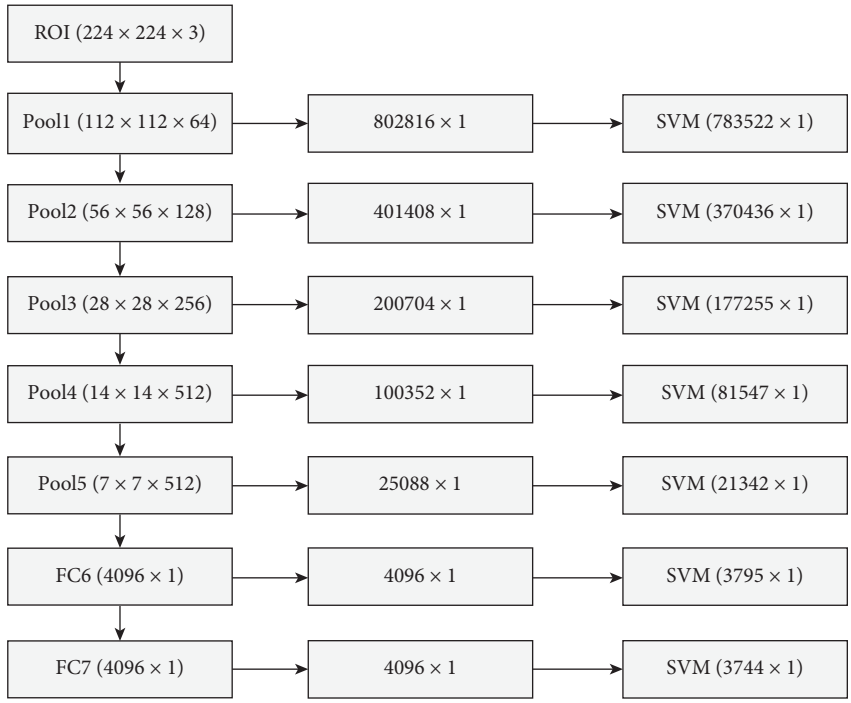

Figure 6: A block diagram of how features are extracted and trained from each layer of a pretrained VGGNet-16.

Table 3: Experimental results with features extracted from different layers of CNNs.

State of CNNs	Dimension	Accuracy	Sensitivity	Specificity	Training time (s)	Testing time (s)
Pool1	783522	0.735	0.72	0.745	4219.3	63.7
Pool2	370436	0.755	0.74	0.766	2395.6	53.3
Pool3	177255	0.784	0.76	0.793	1500.4	41.4
Pool4	81547	0.804	0.78	0.821	957.6	35.1
Pool5	21342	0.823	0.8	0.848	510.9	27.3
FC6	3795	0.812	0.79	0.828	109.1	12.7
FC7	3740	0.653	0.59	0.697	107.3	11.5

three decision criteria: pessimistic, optimistic, and compromise, corresponding to three decision results. In this paper, the voting system is used to predict the predictive nature of thyroid nodules (probability of benign or malignant). The performances before and after the voting system are shown in Table 5. The receiver operating characteristic curve (ROC) and area under curve (AUC) are shown in Figure 7.

It is easy to know that the performances of hybrid approach based on multimodalities are better than that of single modality such as FC6-SVM and SF-PCA-SVM. The performances after voting system have been greatly improved, achieving an accuracy of 0.865 (compromise criteria), sensitivity of 0.89 (pessimistic criteria), specificity of 0.931 (optimistic criteria), and AUC of 0.921 (compromise criteria). The accuracy rates of pessimistic criteria and optimistic criteria are relatively close, slightly better than the results of FC6-SVM and SF-PCA-SVM. However, the accuracy rate of compromise criteria has increased by more than 5 percentage points and the AUC is nearly improved by 6 percentage points.

5. Discussion

The difference in visual imaging can be quantified by algorithms such as statistical machine learning and deep learning to train a CAD system, which can automatically differentiate the nature of thyroid nodules. A lot of research studies use ultrasound images as a dataset to train the CAD system for predicting the benign and malignant nodules [14–19, 22, 25, 26]. These research studies confirmed the usefulness of discriminating the nature of thyroid nodules by learning from B-US and SWE-US images. Knowing this, we have shown the detailed comparison results of different CAD models based on features of different ultrasound modalities.

From the experiment results, we found that it is obvious that the scores of specificity are higher than sensitivity for each model. There may be two reasons for this result. One is that in our dataset, the number of benign samples is more than the number of malignant samples, which is nearly 1.5 times; on the other hand, as shown in the ROIs of Figure 3, the imaging performance of malignant nodules is more

Table 4: Experimental results of various techniques.

Features	Methods	Accuracy	Sensitivity	Specificity	Training time (s)	Testing time (s)
SF-TTEST	LR	0.78	0.76	0.793	6.5	1.8
	Naive Bayes	0.764	0.75	0.772	6.2	1.5
	SVM	0.784	0.77	0.793	17.8	5.2
SF-PCA	LR	0.795	0.77	0.807	8.6	1.9
	Naive Bayes	0.784	0.76	0.793	8.5	1.8
	SVM	0.804	0.79	0.814	18.7	5.3
SF	LR	0.755	0.75	0.758	9.4	2.1
	Naive Bayes	0.733	0.71	0.751	8.9	1.9
	SVM	0.776	0.76	0.786	19.8	5.7
FC6	LR	0.784	0.77	0.793	57.9	11.2
	Naive Bayes	0.755	0.74	0.766	48.2	10.6
	SVM	0.812	0.79	0.828	109.1	12.7

Table 5: Experimental results of thyroid classification with voting system.

Methods	Accuracy	Sensitivity	Specificity
FC6-SVM	0.812	0.79	0.828
SF-PCA-SVM	0.804	0.79	0.814
Pessimistic	0.824	**0.89**	0.779
Optimistic	0.832	0.69	**0.931**
Compromise	**0.865**	0.82	0.897

The best performing results are shown in bold.

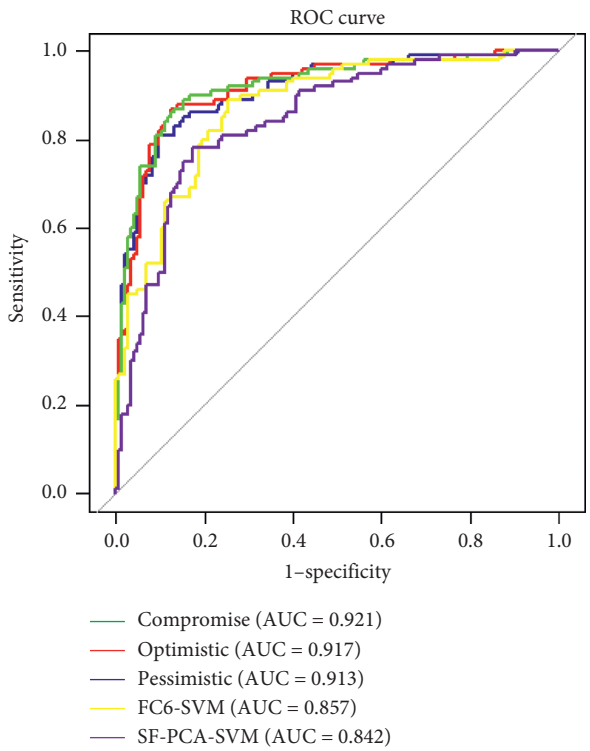

Figure 7: ROCs and AUCs.

diverse while the benign nodules have more typical characteristics and are easier to identify.

The deep features generated by CNNs can better represent the inherent features of the image, regardless of the field and type of the image. Therefore, it is feasible to transfer the pretrained VGGNet-16 model to the ultrasound domain, which has been proven in experiments. In addition, models trained with features extracted from different CNN layers have different performance. According to experiments, we find that the higher the number of layers, the better the performance before fully connected layer, and the effect of the fully connected layer is weaker than the pooling layer. One possible explanation is that operations such as convolutional layer and pooling layer are to map the original data to the hidden layer feature space. Higher layers can capture different kinds of common features, while lower layers only calculate low-level features and cannot represent high-level semantic functions. The main role of the fully connected layer is to map the hidden layer features to the sample label space, but the target domain and the source domain are quite different, which results in the performance of fully connected layer being relatively weaker.

To the best of our knowledge, this is the first attempt to train models with statistical features from B-US and deep features from SWE-US, respectively, and fuse them together by a voting system, which can improve the accuracy of thyroid nodule diagnosis. Sun et al. [22] extracted the deep and statistical features of B-US and fused them to train a SVM model but did not extract the features from SWE-US. Liu et al. [21] extracted statistical features from B-US and SWE-US, respectively, and merged them for feature selection and modeling. However, they did not consider the deep features to better represent the image characteristics. In fact, B-US and SWE-US have different characteristics. They can help diagnose thyroid nodules from different perspectives, and they are complementary. B-US clearly characterizes the nodule shape, contour, texture, and other features, which can be well consistent with the diagnosis criteria of

ACR TI-RADS [29], so we extract statistical features from B-US. SWE-US is a colorful image, which indicates the elasticity or hardness change between adjacent tissues, and it has high specificity and sensitivity for the differential diagnosis of malignant lesions. So, we use pretrained CNNs to extract deep features from SWE-US. Finally, we train two different classifier models using these two kinds of features separately and combine the models together with a voting system to predict the nature of thyroid nodule. The AUC of our model is improved about 6% from 0.857 and 0.842 to 0.921. Experimental results indicate that the proposed hybrid approach based on multimodalities has better performance than the model trained with single modality (B-US or SWE-US) separately.

It needs to be emphasized that sensitivity is more important in clinical practice. Sensitivity indicates the correct proportion of malignant while specificity indicates the correct proportion of benign. The higher the sensitivity, the lower the rate of missed diagnosis. In our research, a voting system is employed to combine different classifier results to improve the accuracy of prediction. Although the accuracy of compromise criteria is better than that of pessimistic criteria, the sensitivity of the pessimistic criteria is better, reaching 89%, which is 7% higher the compromise criteria. We strongly recommend using pessimistic criteria in clinical practice because high sensitivity can reduce misdiagnosis.

6. Conclusion

A large percentage (about 70%) of FNA results of thyroid nodules turn out to be benign [30]. So, it is significant to predict the nature of thyroid nodules before FNA. In this paper, we have proposed a novel hybrid approach by using preoperative multimodality images. Multimodality ultrasound images can mine different kinds of information about the nodule. Our research assumes that a fusion model based on the two modalities of B-US and SWE-US may provide better performance than the model trained on each single ultrasound imaging mode. The results have confirmed this hypothesis.

Further research could be performed in this area to overcome limitations of this study. Deep models in medical intelligent diagnosis require large datasets, especially multicenter large datasets, to mine the hidden information of predictions to avoid overfitting the model. Additionally, a new hybrid approach including feature hybridization and classification result hybridization needs to be proposed and applied to improve the accuracy of the model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hongjun Sun and Feihong Yu contributed equally to this study.

Acknowledgments

This study was financially supported by the Research and Innovation Program for Graduate Education of Jiangsu Province (KYZZ15_0110) and the National Natural Science Foundation of China (NSFC) (71971115 and 71471087).

Supplementary Materials

ROI segmentation: it mainly introduces the software and the methods to segment the ROI of ultrasound image. Statistical feature extraction: it contains python code to perform the statistical feature extraction described in the article. Statistical feature extraction methodology: it mainly introduces the description and calculation method of statistical characteristics in detail. (*Supplementary Materials*)

References

[1] R. M. Walsh, J. C. Watkinson, and J. Franklyn, "The management of the solitary thyroid nodule: a review," *Clinical Otolaryngology & Allied Sciences*, vol. 24, no. 5, pp. 388–397, 1999.

[2] American Cancer Society, *What are the Key Statistics About Thyroid Cancer?*, American Cancer Society, Atlanta, GA, USA, 2014, http://www.cancer.org/Cancer/ThyroidCancer/DetailedGuide/thyroid-cancer-key-statistics.

[3] N. Singh Ospina, J. P. Brito, S. Maraka et al., "Diagnostic accuracy of ultrasound-guided fine needle aspiration biopsy for thyroid malignancy: systematic review and meta-analysis," *Endocrine*, vol. 3, no. 53, pp. 651–661, 2016.

[4] L. Davies and H. G. Welch, "Current thyroid cancer trends in the United States," *JAMA Otolaryngology—Head & Neck Surgery*, vol. 140, no. 4, p. 317, 2014.

[5] S. J. Mandel, "Diagnostic use of ultrasonography in patients with nodular thyroid disease," *Endocrine Practice*, vol. 10, no. 3, pp. 246–252, 2004.

[6] J. R. Wienke, W. K. Chong, J. R. Fielding, K. H. Zou, and C. A. Mittelstaedt, "Sonographic features of benign thyroid nodules," *Journal of Ultrasound in Medicine*, vol. 22, no. 10, pp. 1027–1031, 2003.

[7] Y. Xiao, J. Zeng, L. Niu et al., "Computer-aided diagnosis based on quantitative elastographic features with supersonic shear wave imaging," *Ultrasound in Medicine & Biology*, vol. 40, no. 2, pp. 275–286, 2014.

[8] M. Mehrmohammadi, P. Song, D. D. Meixner et al., "Comb-push ultrasound shear elastography (cuse) for evaluation of thyroid nodules: preliminary in vivo results," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 97–106, 2015.

[9] H. Köhler, C. Happel, J. Klebe, H. Ackermann, and F. Grünwald, "Diagnostic accuracy of elastography and scintigraphic imaging after thermal microwave ablation of thyroid nodules," *RöFo—Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 187, no. 5, pp. 353–359, 2015.

[10] G. Song, F. Xue, and C. Zhang, "A model using texture features to differentiate the nature of thyroid nodules on sonography," *Journal of Ultrasound in Medicine*, vol. 34, no. 10, pp. 1753–1760, 2015.

[11] N. Ketkar, "Convolutional neural networks," in *Deep Learning with Python*, Apress, Berkeley, CA, USA, 2017.

[12] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on*

Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015.

[13] U. Raghavendra, A. Gudigar, M. Maithri et al., "Optimized multi-level elongated quinary patterns for the assessment of thyroid nodules in ultrasound images," *Computers in Biology and Medicine*, vol. 95, pp. 55–62, 2018.

[14] M. Buda, B. Wildman-Tobriner, J. K. Hoang et al., "Management of thyroid nodules seen on US images: deep learning may match performance of radiologists," *Radiology*, vol. 292, no. 3, pp. 695–701, 2019.

[15] X. Mei, X. Dong, T. Deyer, J. Zeng, T. Trafalis, and Y. Fang, "Thyroid nodule benignty prediction by deep feature extraction," in *Proceedings of the IEEE International Conference on Bioinformatics & Bioengineering. IEEE Computer Society*, Washington, DC, USA, October 2017.

[16] Y. Wang, W. Yue, X. Li et al., "Comparison study of radiomics and deep learning based methods for thyroid nodules classification using ultrasound images," *IEEE Access*, vol. 8, 2020.

[17] X. Yu, H. Wang, and L. Ma, "Detection of thyroid nodules with ultrasound images based on deep learning," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 16, no. 2, pp. 174–180, 2019.

[18] S. Yue, X. Mingxing, J. Wang et al., "Real-time shear wave elastography in differential diagnosis of benign and malignant thyroid nodules," *Journal of Chinese Physician*, vol. 12, pp. 324–330, 2019.

[19] S. Y. Nam, J. H. Shin, B.-K. Han et al., "Preoperative ultrasonographic features of papillary thyroid carcinoma predict biological behavior," *The Journal of Clinical Endocrinology & Metabolism*, vol. 98, no. 4, pp. 1476–1482, 2013.

[20] Q. Zhang, Y. Xiao, W. Dai et al., "Deep learning based classification of breast tumors with shear-wave elastography," *Ultrasonics*, vol. 72, pp. 150–157, 2016.

[21] T. Liu, X. Ge, J. Yu et al., "Comparison of the application of B-mode and strain elastography ultrasound in the estimation of lymph node metastasis of papillary thyroid carcinoma based on a radiomics approach," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 10, pp. 1617–1627, 2018.

[22] W. Sun, T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images, using deep model based transfer learning and hybrid features," in *Proceedongs of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, New Orleans, LA, USA.

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" vol. 27, pp. 3320–3328, 2014, https://arxiv.org/abs/1411.1792.

[24] S. Wang and J. Wei, "Feature selection based on measurement of ability to classify subproblems," *Neurocomputing*, vol. 224, pp. 155–165, 2017.

[25] O. Moussa, H. Khachnaoui, R. Guetari, and N. Khlifa, "Thyroid nodules classification and diagnosis in ultrasound images using fine-tuning deep convolutional neural network," *International Journal of Imaging Systems and Technology*, vol. 30, no. 1, pp. 185–195, 2020.

[26] J. Song, Y. J. Chai, H. Masuoka et al., "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules," *Medicine (Baltimore)*, vol. 98, no. 15, Article ID e15133, 2019.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, https://arxiv.org/abs/1409.1556, 2014.

[28] J. Branke, S. E. Chick, and C. Schmidt, "Selecting a selection procedure," *Management Science*, vol. 53, no. 12, pp. 1916–1932, 2007.

[29] F. N. Tessler, W. D. Middleton, E. G. Grant et al., "ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee," *Journal of the American College of Radiology*, vol. 14, no. 5, pp. 587–595, 2017.

[30] J. Ma, S. Luo, M. Dighe, D. J. Lim, and Y. Kim, "Differential diagnosis of thyroid nodules with ultrasound elastography based on support vector machines," in *Proceedings of the 2010 IEEE International Ultrasonics Symposium*, San Diego, CA, USA, October 2011.