BMC Cancer

RESEARCH

Open Access

Thyroid nodule classification in ultrasound imaging using deep transfer learning

Yan Xu[1], Mingmin Xu[1], Zhe Geng[1], Jie Liu[2,3*] and Bin Meng[1*]

Abstract

Background The accurate diagnosis of thyroid nodules represents a critical and frequently encountered challenge in clinical practice, necessitating enhanced precision in diagnostic methodologies. In this study, we investigate the predictive efficacy of distinguishing between benign and malignant thyroid nodules by employing traditional machine learning algorithms and a deep transfer learning model, aiming to advance the diagnostic paradigm in this field.

Methods In this retrospective study, ITK-Snap software was utilized for image preprocessing and feature extraction from thyroid nodules. Feature screening and dimensionality reduction were conducted using the least absolute shrinkage and selection operator (LASSO) regression method. To identify the optimal model, both traditional machine learning and transfer learning approaches were employed, followed by model fusion using post-fusion techniques. The performance of the model was rigorously evaluated through the area under the curve (AUC), calibration curve analysis, and decision curve analysis (DCA).

Results A total of 1134 images from 630 cases of thyroid nodules were included in this study, comprising 589 benign nodules and 545 malignant nodules. Through comparative analysis, the support vector machine (SVM), which demonstrated the best diagnostic performance among traditional machine learning models, and the Inception V3 convolutional neural network model, based on transfer learning, were selected for model construction. The SVM model achieved an AUC of 0.748 (95% CI: 0.684–0.811) for diagnosing malignant thyroid nodules, while the Inception V3 transfer learning model yielded an AUC of 0.763 (95% CI: 0.702–0.825). Following model fusion, the AUC improved to 0.783 (95% CI: 0.724–0.841). The difference in performance between the fusion model and the traditional machine learning model was statistically significant ($p = 0.036$). Decision curve analysis (DCA) further confirmed that the fusion model exhibits superior clinical utility, highlighting its potential for practical application in thyroid nodule diagnosis.

Conclusion Our findings demonstrate that the fusion model, which integrates a convolutional neural network (CNN) with traditional machine learning and deep transfer learning techniques, can effectively differentiate between benign and malignant thyroid nodules through the analysis of ultrasound images. This model fusion approach significantly optimizes and enhances diagnostic performance, offering a robust and intelligent tool for the clinical detection of thyroid diseases.

Keywords Machine learning, Deep learning, Transfer learning, Thyroid, Classification, Ultrasound image

*Correspondence:
Jie Liu
liujie_w@sina.com
Bin Meng
netboysky@qq.com
Full list of author information is available at the end of the article

Introduction

The prevalence of thyroid nodules in the general population is as high as 65%. Although only about 10% prove to be malignant [1]. But thyroid cancer rates are increasing every year among all genders, races, and age groups [2]. Ultrasound is the primary tool used to stratify cancer risk of thyroid nodules and decide whether to perform fine-needle aspiration (FNA) [1, 3]. However, ultrasound diagnostic results are somewhat subjective. The combination of ultrasound and artificial intelligence (AI) makes up for the subjectivity and operator dependence of thyroid ultrasound diagnosis [4]. Using artificial intelligence algorithms to analyze thyroid ultrasound imaging data could help distinguish patients at different risks and avoid unnecessary fine-needle aspiration biopsies or thyroidectomies in lower-risk patients [5].

Computer-aided diagnosis (CAD) systems can provide objective and reliable diagnostic evidence for thyroid diseases [6]. The use of CAD improves diagnostic performance [7, 8] and its diagnostic performance is comparable to that of experienced endocrinologists [9]. Artificial intelligence CAD systems are based on two technologies: machine learning (ML) and deep learning (DL). ML techniques rely on extracting and selecting the most obvious features from a region of interest (ROI) to apply to the ML classifier. DL technology uses deep learning, taking the original image pixels and corresponding category labels as input, and simultaneously performs feature extraction, selection, and final classification during the training process [10]. DL is the most mainstream technology in the field of artificial intelligence medical imaging. Convolutional neural network (CNN) is a common deep learning architecture [11]. Transfer learning has emerged as a key strategy to overcome the challenges of limited annotated medical data, enabling CNNs to achieve high performance even with small datasets [12, 13]. Its diagnostic algorithm can effectively classify benign and malignant thyroid nodules, and its diagnostic performance is equivalent to the results obtained by experienced ultrasound doctors based on TI-RADS reports.

Both machine learning and deep learning-based CAD systems have good diagnostic results for malignant thyroid nodules. Recent studies have further advanced this field, with research proposing a hybrid optimization algorithm for improved feature selection and other studies demonstrating the effectiveness of deep transfer learning [14, 15]. However, many current studies rely solely on either traditional machine learning methods or deep learning and transfer learning methods to diagnose thyroid nodules. This limits the diagnostic performance as traditional machine learning excels in interpretability and handling structured data, while deep learning captures complex patterns in high-dimensional data. Fusion modeling that integrates the strengths of both approaches offers a novel solution to bridge this gap. To clarify the diagnostic performance of traditional machine learning and deep transfer learning, our study uses a post-fusion method that integrates traditional machine learning with deep transfer learning to diagnose thyroid nodules. This combined approach is used to jointly diagnose benign and malignant thyroid nodules, and their diagnostic performance was compared.

Methods

Research objects

Patients and data were obtained from Zhejiang Rongjun Hospital. Since this study was retrospective, the requirement for patient informed consent was exempted, and the research protocol was approved by the Ethics Committee of Zhejiang Rongjun Hospital. This study included thyroid nodules that underwent fine needle aspiration or inpatient surgery from January 2019 to December 2023 and had pathological results.

The inclusion criteria were as below: patients who underwent ultrasound examination at the hospital within 2 weeks before surgery and with complete ultrasound data; all nodules were confirmed by pathological results obtained through either fine-needle aspiration or surgical resection.

The exclusion criteria were as below: (1) patients with a history of thyroid surgery, radiotherapy, chemotherapy, or radiofrequency ablation; (2) poor image quality, making it difficult to identify tumor boundaries and unable to perform image segmentation. (3) Images with measurement markers exceeding the image range or incomplete nodule sections; (4) Images with both color Doppler and measurement markers; a total of 630 patients were screened, including 345 benign nodules and 285 malignant nodules. For example, take one to four images from each patient's ultrasound images.

Ultrasound examination and image acquisition

All selected patients underwent preoperative cervical ultrasonography. Ultrasound machines include Mylab90 (Esaote, Italy), Philips EPIQ5, and EPIQ7 ultrasound systems (Netherlands). Ultrasound examinations were performed by radiologists with 10–15 years of experience in thyroid ultrasound evaluation using a 5–12 MHz transducer. After placing the patient in the supine position, serial longitudinal and transverse scans were performed to obtain longitudinal and transverse images of the thyroid nodules. According to the 2020 Chinese Ultrasound Malignant Tumor Risk Stratification Guidelines for Thyroid Nodules: C-TIRADS evaluates the following ultrasound characteristics of thyroid nodules: composition

(cystic or almost completely cystic, spongy, mixed cystic and solid, almost completely solid) nature, solid), echo (anechoic, isoechoic or hyperechoic, hypoechoic, extremely hypoechoic,), edge (smooth, blurred, irregular, toward the outside of the thyroid capsule), morphology (vertical and horizontal diameter ratio <1 or ≥ 1), and calcification (comet tail artifact, peripheral calcification, coarse calcification, microcalcification) [16].

Image preprocessing

Image segmentation

For image preprocessing, regions of interest (ROI) were manually segmented using ITK-SNAP software (version 3.6.0, USA). First, two sonographers with 10–15 years of experience in thyroid imaging diagnosis used a blind method to complete manual cropping. Because the outline of the nodule edge is another factor that affects the classification diagnosis, it is necessary to ensure the accuracy of manual outline. They were then reviewed by a senior sonographer with more than 15 years of experience in thyroid imaging. Image annotation and any disagreements will be resolved through negotiation.

ROI extraction and data splitting

Transfer learning selects the image with the largest nodule area to represent each nodule. Each cropped sub-region image is then resized to 224×224 to generate a minimum area rectangular bounding box for each lesion, and the image within the bounding box is called the ROI image. The lesion ROI images are then used as input images for CNN model training and testing. In the experiment, we randomly select 80% of the images from the total sample as the training data set and 20% of the images as the testing data set.

Feature extraction

Traditional radiomic features

Fifty cases of data were randomly selected for secondary annotation, and the inter-observer consistency in nodules' manual outline delineation was evaluated using the inter-class correlation coefficient (ICC). A correlation coefficient greater than 0.75 was considered satisfactory agreement. A total of 108 traditional radiomic features were extracted from ROIs, including texture features (gray-level co-occurrence matrix), shape features (area, perimeter, circularity), and intensity-based features (mean intensity, standard deviation). These features were calculated using the PyRadiomics library (version 3.1.0) in Python.

Deep transfer learning features

For deep transfer learning, the images are input into a CNN pre-trained on the ImageNet dataset [17], which extracts deep transfer learning features from each ultrasound image modality and performs feature dimensionality reduction.

Feature filtering and model building for deep transfer learning

For the deep transfer learning features with 2048 dimensions, in order to ensure the balance between features, we use principal component analysis (PCA) to reduce the dimensionality of deep transfer learning features to improve the generalization ability of the model and reduce the risk of overfitting. This study uses VGG16, ResNet50, and Inceptionv3 deep convolutional neural networks pre-trained on the ImageNet data set to implement transfer learning, that is, using the pre-trained model to freeze all convolutional layers and fine-tuning the fully connected transfer learning method. We use the Adam optimization algorithm in the model training process. Each of our experiments runs for 30 epochs, with a batch size of 32 and Dropout set to 0.5.

Feature filtering and model building for machine learning

All 108 traditional radiomic features are normalized in the training and test sets, and the correlation between features is calculated using the Spearman correlation coefficient. For features with a correlation coefficient greater than 0.9, one of the two features is retained. The traditional features are then concatenated with the dimensionality-reduced deep transfer learning features obtained through PCA. The least absolute shrinkage and selection operator (LASSO) algorithm is used to screen out the features with the most significant predictive potential, which are then fed into the machine learning models for further analysis. We used logistic regression (LR), SVM, Naive Bayes, k-Nearest Neighbors (KNN), Decision Tree, Random Forest (RF), Extra Trees, eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Gradient Boosting (GBM), Adaptive Boosting (AdaBoost), and Multilayer Perceptron (MLP) neural network to train each classification model and used tenfold cross-validation to obtain the final radiation group learning characteristics. According to the model's predictions on the test set, the experimental results were analyzed to assess the overall classification accuracy and performance.

Model development and evaluation

We used Python 3.70 for model construction, evaluation, and repeated cross-validation, with the following libraries: numpy (1.21.0), scikit-learn (1.0.2), PyTorch (1.10.0), PyRadiomics (3.1.0) and OpenCV (4.5.4). Hyperparameters were optimized using manual tuning. The code was executed on a workstation equipped with

a 12 GiB NVIDIA Titan V GPU and 64 GiB of memory. To improve the accuracy of predicting thyroid nodules in ultrasound images, post-fusion of models was studied using the average method. That is, the results of machine learning and deep transfer learning are combined to predict benign and malignant thyroid nodules, and then the malignant probabilities are averaged and classified.

To quantitatively evaluate and compare the performance of different models, we calculated an overall score based on accuracy, AUC, sensitivity, specificity, PPV, NPV, precision, recall, F1-Score, and overfitting degree. Each metric was assigned an equal weight of 10%. The overfitting degree was quantified as the difference between training set accuracy and test set accuracy, and normalized to a range of [0, 1] across all models. Specifically, the normalized overfitting value was calculated as 1-(training accuracy)/(testing accuracy), ensuring that a higher value indicates less overfitting. The overall score was computed as the weighted sum of the 10 metrics.

Statistical Analysis

ICCs, Spearman rank correlation test, z-score normalization and LASSO regression analysis were performed using Python 3.7. Descriptive statistics for continuous variables are expressed as mean ± standard deviation; categorical variables are expressed as percentages (%). The independent samples t test was used for continuous variables with normal distribution; the Mann–Whitney U test was used for continuous variables without normal distribution. Categorical variables were analyzed using the chi-square test. Use the Receiver Operating Characteristic (ROC) curve to show the diagnostic performance of each model. Delong validation was used to compare the AUC values of different models. A P value less than 0.05 was considered statistically significant. The calibration curve was drawn to evaluate the goodness of fit of the model. Decision Curve Analysis(DCA)was applied to evaluate the clinical application value of the model. The confidence intervals (CIs) for the AUC values were calculated using the DeLong test. DeLong test is a robust non-parametric method for comparing the AUC of ROC curves and estimating the variability.

Results

Patient characteristics

A total of 630 patients and 1134 pictures were selected, including 589 pictures of benign nodules and 545 pictures of malignant nodules. The average age is 52.19 ± 13.298 years old; the age range is 15–86 years old. According to 8:2 random stratified sampling, it is divided into a training set of 907 images and a verification set of 227 images. The baseline characteristics of patient pictures in the training and test groups (Table 1). There was no statistically significant difference between the training set and the test set in terms of patient age, gender, tumor maximum diameter, location, composition, margin, morphology, calcification and C-TIRADS. Through single-factor and multi-factor logistic analysis, it was concluded that age, C-TIRADS, composition and echo are independent predictors of malignant thyroid nodules (Table 2).

Image preprocessing

After obtaining the original data set and then outline and label the data. The input ROI image contains the entire tumor area and its boundary area. The image generated after the outline is completed (Fig. 1). When performing deep transfer learning, we selected the largest section image of the thyroid nodule.

Feature screening and modeling

A total of 108 traditional radiomic features were extracted for each lesion, and 42 features were retained after screening. The features were dimensionally reduced and screened through the ten-fold intersection algorithm and LASSO regression, and 8 significant features were finally retained. For convolutional neural network technology, this study uses VGG16, ResNet50, and Inceptionv3 deep convolutional neural networks pre-trained on the ImageNet data set to implement transfer learning, replacing its fully connected layer, SoftMax layer and classification output layer, and Set the fully connected layer to be the same as the number of classes in the new data, thereby generating a new network model. Finally, the optimal model Inceptionv3 was selected for deep transfer learning. There are 32 features after feature compression. The Adam optimizer was used with a learning rate of 0.001, beta1 of 0.9, beta2 of 0.999, and epsilon of 10^{-8}. L2 regularization with a lambda value of 0.01 and early stopping are used to prevent overfitting, and loss rate is used to evaluate model performance.

Model evaluation analysis

Using radiomic features, the optimal model SVM is obtained compared with LR, NaiveBayes, KNN, Decision Tree, RF, Extra Trees, XGBoost, LightGBM, GradientBoosting, AdaBoost and MLP classifiers. It achieved the best overall score in the classification diagnosis of thyroid nodules training set and test set (Table 3, Fig. 2). To find the best deep transfer learning model for evaluating thyroid nodules, we compared the performance of pretrained VGG16, Resnet50, and Inception v3. The results show that Inception v3 has the best performance, with an accuracy of 72.2% and an AUC of 0.763 (95%CI: 0.702–0.825) (Fig. 3). Therefore, we use Inception V3 to further evaluate the model effect.

Table 1 Comparison of general data and ultrasound characteristics of benign and malignant thyroid nodules in training set and test set

Characteristics	All ($n=1134$)	Training sample ($n=907$)	Validation sample ($n=227$)	*P* value
Age(y) +	52.19 ± 13.298	52.19 ± 13.385	52.18 ± 12.97	0.445
Max Size (mm)	14.03 ± 14.447	14.029 ± 14.482	14.01 ± 14.337	0.851
Sex				0.173
Female	424(0.374)	348 (0.384)	76(0.335)	
Male	710(0.626)	559(0.616)	151(0.665)	
Position				0.869
Left thyroid	522(0.46)	419(0.462)	103(0.454)	
Right thyroid	550(0.485)	440(0.485)	110(0.485)	
Thyroid isthmus	62(0.055)	48(0.053)	14(0.061)	
Composition				0.423
Cystic or almost completely cystic	34(0.03)	23(0.025)	11(0.048)	
Spongiform	15(0.013)	13(0.014)	2(0.009)	
Mixed cystic and solid	126(0.111)	104(0.115)	22(0.097)	
Almost completely solid	103(0.091)	83(0.092)	20(0.088)	
Solid	856(0.755)	684(0.754)	172(0.758)	
Echogenicity				0.024*
Anechoic	63(0.056)	44(0.049)	19(0.084)	
Hyperechoic or isoechoic	130(0.115)	109(0.12)	21(0.093)	
Hypoechoic	893(0.787)	710(0.782)	183(0.805)	
Very hypoechoic	48(0.042)	44(0.049)	4(0.018)	
Shape				0.099
Wider-than-tall	783(0.69)	616(0.679)	167(0.736)	
Taller-than-wide	351(0.31)	291(0.321)	60(0.264)	
Margin				0.183
Smooth	901(0.795)	722(0.796)	179(0.789)	
Ill-defined	107(0.094)	85(0.094)	22 (0.097)	
Lobulated or irregular	106(0.093)	88(0.097)	18(0.079)	
Extra-thyroidal extension	20(0.018)	12(0.013)	8(0.035)	
Echogenic Foci				0.173
None	875(0.772)	686(0.756)	189(0.833)	
Comet-tail artifacts	6(0.005)	5(0.006)	1(0.004)	
Macrocal cifications	81(0.071)	69(0.076)	12(0.053)	
Peripheral calcifications	53(0.047)	45(0.05)	8(0.035)	
Punctate echogenic foci	119(0.105)	102(0.112)	17(0.075)	
C TI-RADS risk Level				0.579
TR2	20(0.018)	14(0.015)	6(0.026)	
TR3	219(0.193)	172(0.19)	47(0.207)	
TR4A	330(0.291)	260(0.287)	70(0.308)	
TR4B	379(0.334)	305(0.336)	74(0.327)	
TR4C	177(0.156)	148(0.163)	29(0.128)	
TR5	9(0.008)	8(0.009)	1(0.004)	

Data are numbers of nodules, with percentages in parentheses. *C,TI-RADS* China Thyroid Imaging Reporting and Data System. Numbers in parentheses represent percentage within a given group (benign, malignant). + Data are means ± standard deviation. * $P<0.05$

Model fusion is an effective means to improve the overall prediction ability. In multi-classifier system and ensemble learning, fusion is a very important step. This study uses a post-fusion method to fuse the SVM model results and the InceptionV3 model results in traditional machine learning. That is, the images of the test set are

Table 2 Logistic regression analysis of predictors of thyroid malignant nodules

Factor	Univariable analysis			Multivariable analysis		
characteristics	OR	95% CI	*P* value	OR	95% CI	*P* value
age	0.96	0.95–0.97	$p<0.001$	0.95	0.94–0.97	$p<0.001$
C_TIRADS	2.22	1.91–2.58	$p<0.001$	2.14	1.78–2.57	$p<0.001$
composition	4.85	3.51–6.72	$p<0.001$	2.43	1.67–3.53	$p<0.001$
echo	8.19	5.16–13	$p<0.001$	2.88	1.64–5.06	$p<0.001$
echogenic_Foci	1.27	1.12–1.44	$p<0.001$			
margin	2.27	1.81–2.83	$p<0.001$			
Position	0.87	0.7–1.09	0.228			
Sex	1.07	0.81–1.39	0.645			
shape	3.23	2.41–4.32	$p<0.001$			
size	0.95	0.94–0.96	$p<0.001$	1.01	1–1.03	0.123

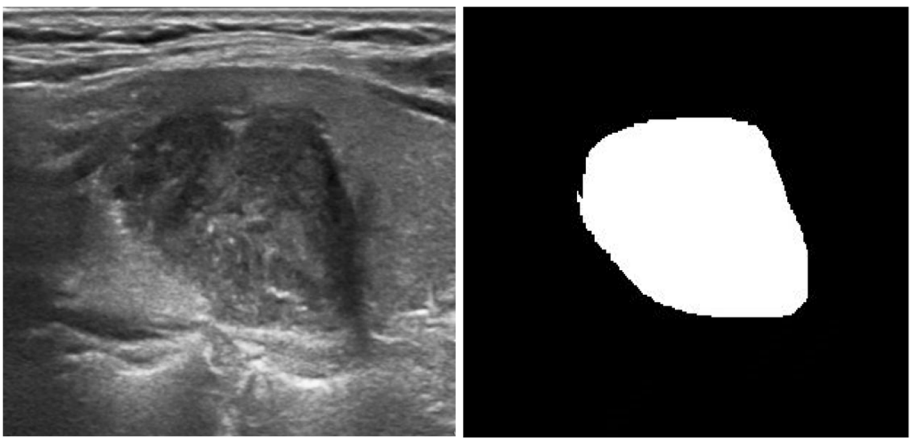

Fig. 1 Manually outline the ROI along the edge of the lesion on the image and its illustration

generated through the trained network to generate a probability value, which represents the malignancy predicted by the model. The results show that in the test set, the performance of the traditional machine learning SVM prediction model based on an AUC of 0.748 (95% CI: 0.684–0.811) is lower than that of the deep transfer learning InceptionV3 model of 0.763 (95% CI: 0:702–0.825) (Fig. 4).

After model fusion, the AUC of the test set is 0.783 (95% CI: 0.724–0.841). The Delong test was used to compare the AUC of the three models. There was no statistically significant difference between the SVM prediction model and the InceptionV3 model in the test cohort ($p>0.05$). The difference between the fusion model and the traditional machine learning model is statistically significant ($p=0.036$). There is no statistically significant difference between the fusion model and deep transfer learning ($P=0.263$). The evaluation effect of traditional machine models can be optimized and improved through model fusion. The calibration curve has good agreement between predictions and observations (Fig. 5). In this study, we evaluated each model through DCA and determined the clinical benefit of the model. The fusion model provided higher net benefit across a wide range of threshold probabilities (Fig. 6), indicating its superior clinical utility. These results suggest that the fusion model can effectively support clinicians in distinguishing benign and malignant thyroid nodules, potentially reducing unnecessary biopsies.

Discussion

Artificial intelligence may increase diagnostic consistency and accuracy, reducing workload for health care professionals [18]. It demonstrates diagnostic performance comparable to or even better than that of sonographers in the ultrasound diagnosis of thyroid nodules [19–22]. There are studies supporting that the CAD method is more objective and the assessment of ultrasound features is relatively accurate [23]. A retrospective study developed an optimized ensemble of artificial

Table 3 Diagnostic performance of machine learning model in predicting benign and malignant thyroid nodules

Model_name	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Precision	Recall	F1	Overall Score	Task
LR	0.653	0.711	0.6779—0.7440	0.78	0.538	0.604	0.729	0.604	0.78	0.681	0.709	label-train
LR	0.67	0.726	0.6608—0.7909	0.719	0.619	0.656	0.686	0.656	0.719	0.686		label-test
NaiveBayes	0.642	0.678	0.6440—0.7129	0.831	0.471	0.587	0.754	0.587	0.831	0.688	0.709	label-train
NaiveBayes	0.674	0.712	0.6451—0.7784	0.658	0.69	0.682	0.667	0.682	0.658	0.67		label-test
SVM	0.721	0.79	0.7606—0.8188	0.766	0.681	0.685	0.762	0.685	0.766	0.723	0.729	label-train
SVM	0.705	0.748	0.6842—0.8116	0.746	0.664	0.691	0.721	0.691	0.746	0.717		label-test
KNN	0.777	0.86	0.8374—0.8819	0.819	0.739	0.74	0.819	0.74	0.819	0.778	0.712	label-train
KNN	0.683	0.746	0.6848—0.8077	0.886	0.486	0.631	0.806	0.631	0.886	0.737		label-test
RandomForest	0.985	0.999	0.9979—0.9997	0.991	0.979	0.977	0.991	0.977	0.991	0.984	0.635	label-train
RandomForest	0.678	0.721	0.6555—0.7871	0.833	0.527	0.638	0.756	0.638	0.833	0.722		label-test
ExtraTrees	1	1	1.0000—1.0000	1	1	1	1	1	1	1	0.684	label-train
ExtraTrees	0.736	0.778	0.7186—0.8375	0.789	0.681	0.714	0.762	0.714	0.789	0.75		label-test
XGBoost	0.939	0.982	0.9750—0.9883	0.947	0.933	0.927	0.951	0.927	0.947	0.937	0.674	label-train
XGBoost	0.687	0.727	0.6611—0.7924	0.93	0.442	0.627	0.862	0.627	0.93	0.749		label-test
LightGBM	0.901	0.962	0.9509—0.9730	0.912	0.891	0.883	0.918	0.883	0.912	0.897	0.649	label-train
LightGBM	0.67	0.725	0.6594—0.7900	0.816	0.522	0.633	0.737	0.633	0.816	0.713		label-test
GradientBoosting	0.752	0.826	0.7993—0.8519	0.775	0.731	0.723	0.782	0.723	0.775	0.748	0.708	label-train
GradientBoosting	0.696	0.733	0.6672—0.7980	0.728	0.664	0.686	0.708	0.686	0.728	0.706		label-test
AdaBoost	0.686	0.744	0.7139—0.7742	0.633	0.733	0.682	0.688	0.682	0.633	0.657	0.710	label-train
AdaBoost	0.674	0.708	0.6447—0.7721	0.781	0.571	0.645	0.719	0.645	0.781	0.706		label-test
MLP	0.68	0.74	0.7082—0.7716	0.798	0.574	0.629	0.758	0.629	0.798	0.703	0.727	label-train
MLP	0.7	0.748	0.6840—0.8112	0.623	0.779	0.74	0.672	0.74	0.623	0.676		label-test

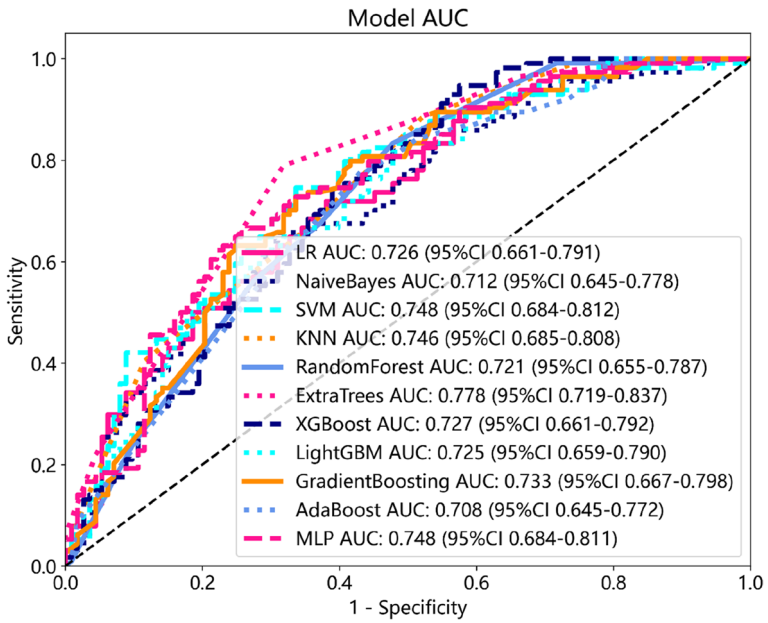

Fig. 2 ROC curves and AUC values of different machine learning models on the test set

intelligence decision aids to assist 10 radiologists with varying levels of expertise. This study analyzed imaging features associated with the effectiveness of artificial intelligence assistance and tested the optimization strategy in a prospective cohort. The results demonstrated that the optimized strategy achieved higher diagnostic

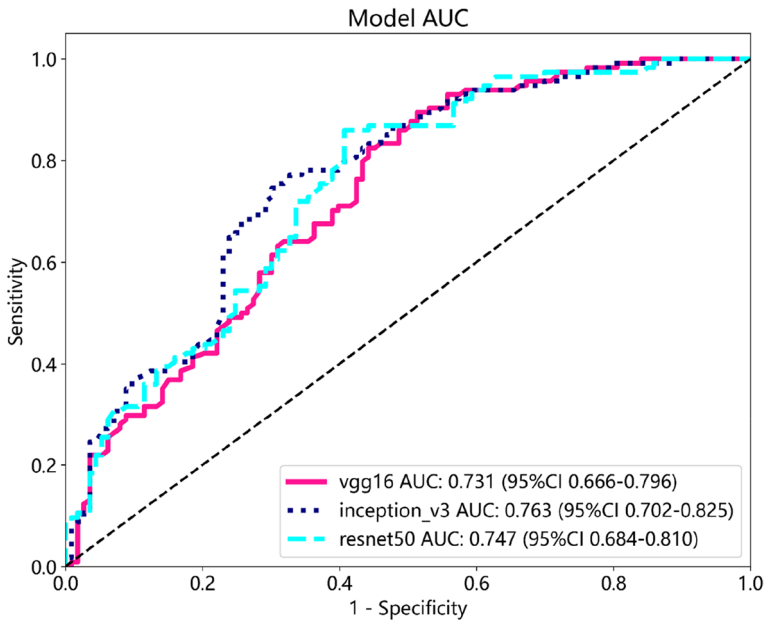

Fig. 3 ROC curves and AUC values of three deep learning models on the test set

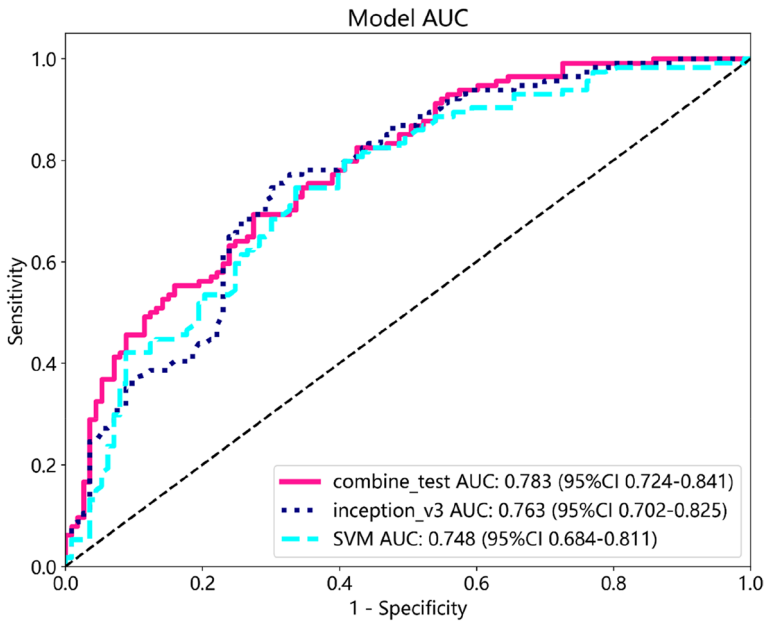

Fig. 4 ROC curves and AUC values of fusion model and single model

performance compared to traditional full artificial intelligence strategies [24]. Another study retrospectively included 10,023 patients with thyroid nodules from 208 hospitals and developed detection, segmentation, and classification models. It was found that a thyroid ultrasound artificial intelligence model developed based on different data sets has high diagnostic performance in the Chinese population and can improve the performance of radiologists in thyroid cancer diagnosis [25].

Some findings suggest that automated "second opinions" can be generated using software applications capable of extracting quantitative parameters from Ultrasound images, and that machine learning methods can be equally accurate [26]. Random forest classifier was

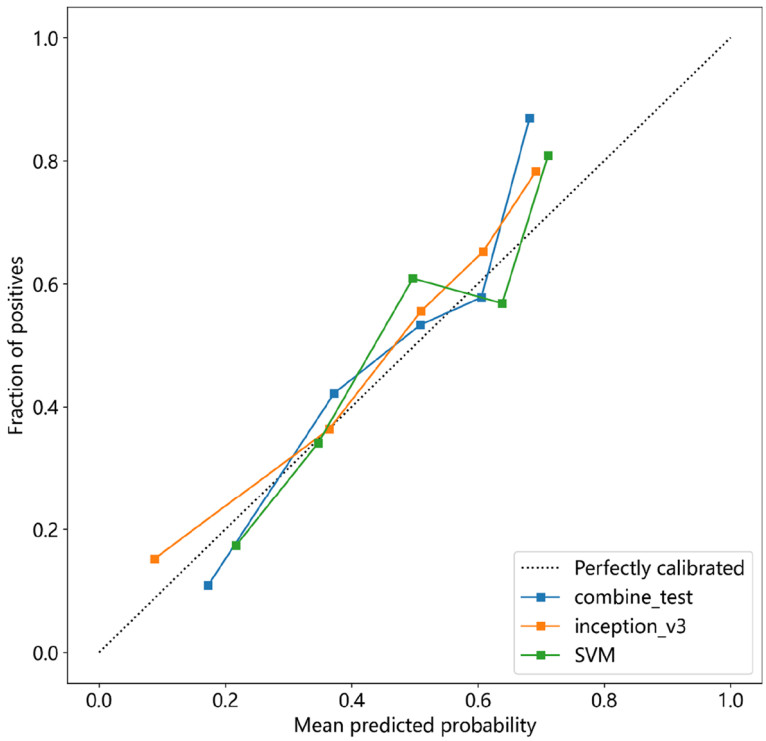

Fig. 5 Calibration curve of the test set

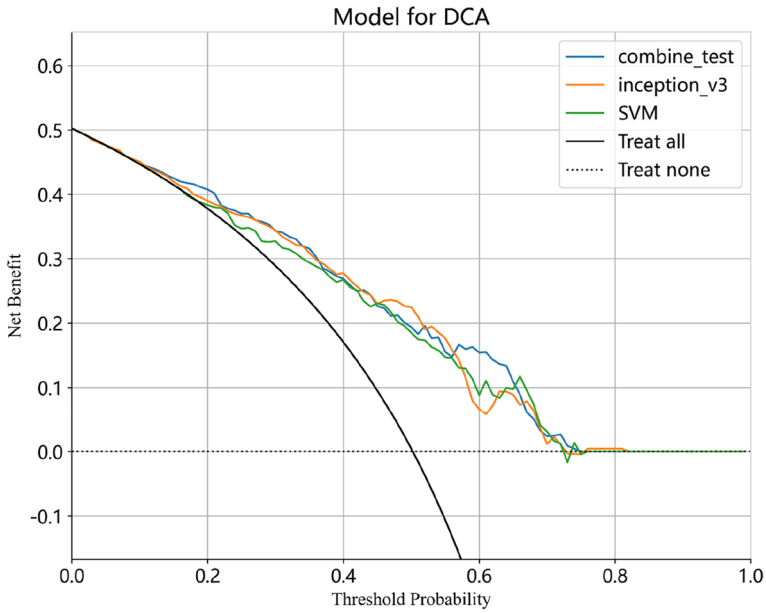

Fig. 6 Decision curve analysis of different models in the test set. The Y-axis represents the net benefit. The fusion model provided higher net benefit across a wide range of threshold probabilities, indicating its superior clinical utility

used to build a final classifier that combined geometric and morphological features to classify thyroid nodules [27], and it was found that the accuracy, sensitivity and specificity of classification could be improved. Zhao et al. showed [28] that RF-based features were used to select the most important features, and it was found that the diagnostic performance and unnecessary biopsy rate of the machine learning-assisted ultrasound vision method

and radiomics method were better than the ACR TI-RADS evaluation method. The study uses 1232 nodules to propose a machine learning framework to predict thyroid nodule malignancy by utilizing GBM, LR, linear discriminant analysis (LDA), radial or linear kernel SVM Training with RF, the RF model has the highest prediction accuracy (0.7931) and the highest AUC (0.8541) [29]. There are studies using four machine learning algorithms including XGBoost, RF, LightGBM and AdaBoost to build a prediction model, incorporating peripheral blood, BRAFV600E gene and Demographic indicators. RF was found to have the highest diagnostic performance, with AUC of 0.874 (95% CI, 0.841–0.906) [30]. Another study used RF and SVM classifiers to classify benign and malignant thyroid nodules and found that both methods have high performance in practice [31, 32]. Our study found that the diagnostic performance of RF in the test set is not as high as the accuracy obtained by the SVM model (accuracy rate is 0.705, AUC is 0.748). It should be noted that our dataset includes both surgical pathology and FNA results. While surgical pathology provides definitive diagnoses, FNA has inherent limitations which may affect the diagnostic accuracy of our study. Additionally, the retrospective design and relatively small sample size may introduce selection bias.

In recent years, image analysis based on convolutional neural networks has been commonly used for lesion detection and classification, which achieved satisfactory results in identifying and classifying thyroid tumors [33]. Previous research on the impact of database size on transfer learning using Deep CNN has shown that for smaller target sets (less than 1000 instances per class) used for transfer learning, freezing the first two to three layers of features significantly improves performance [34, 35]. In the context of transfer learning, better image classification performance can also be achieved by adjusting the optimal number of layers for the target task. Basha et al. proposed to automatically adjust the convolutional neural network to use the knowledge of the target data set to adjust the pre-trained CNN layer to have better transfer learning capabilities [36]. Shao et al. used ResNet50 and VGG16 and found that transfer learning pretrained on ImageNet performed better than the untrained Deep CNN (ResNet50) model [37]. Zhou et al. used a basic convolutional neural network (CNN) model [38], a transfer learning (TL) model and a newly designed deep learning radiomics of thyroid (DLRT) model to conduct research and found that other deep learning models were not as good as the DLRT model, but in this study, patients with nodule diameter less than 10 mm were not selected. The study used 578 patients to perform tenfold cross-validation on ultrasound images of different pathological types of thyroid using InceptionV3 and found that the average accuracy was 0.85 [39]. A recent study trained and tested a set of models using 11,201 images of 6784 nodules, fused the final prediction probabilities of Inception-ResNet and DenseNet121, and used RF models for classification, finding that the highest AUC was 0.94 [40]. In the current study, we compared three pretraining models and selected InceptionV3 as the optimal model based on its comprehensive performance. The model achieved an accuracy of 72.2% and an AUC of 0.763 in predicting malignant thyroid nodules. While these results are promising, the performance was lower than that reported in some previous studies. This discrepancy may be attributed to differences in sample size, image quality, hyperparameter settings, and fine-tuning strategies.

Studies have pointed out that machine learning (ML) combined with deep learning (DL) models show the potential to reduce the possibility of misdiagnosis of breast cancer, and this model is significantly better than any single model [41]. Consistent with the results of this study, our study showed that a fusion classification model to distinguish different pathological classifications of malignant thyroid nodules in ultrasound images. It was found that the fused model had higher diagnostic accuracy. Compared with CNNs and transformer-based architectures which often lack interpretability, our fusion approach can leverage the interpretability of traditional methods and the high-dimensional feature extraction capabilities of deep learning. The findings of this study have broader implications beyond thyroid nodule classification. The fusion modeling approach can be adapted to other diagnostic imaging tasks (e. g. lung nodule classification) and other modalities (e. g. CT, MRI, PET) to further validate its utility in clinical practice. This study also has several limitations. The sample size may limit the generalizability of our findings. Larger, multicenter studies are needed to validate the performance of the fusion model across diverse populations and imaging protocols. Besides, the retrospective design introduces potential selection bias as the included patients were those who underwent fine-needle aspiration or surgical resection. The dataset primarily consists of static ultrasound images. Future studies should consider incorporating video-based analysis to further improve diagnostic accuracy. Lastly, despite their advantages, fusion models may be limited by computational demands and the potential for overfitting.

Conclusion

This study demonstrates that a CNN-based fusion model integrating traditional machine learning and deep transfer learning techniques can effectively differentiate between benign and malignant thyroid nodules using

ultrasound images. The fusion approach optimizes and enhances the model's diagnostic performance, offering a robust and intelligent solution for thyroid disease detection. This method provides a clinically practical tool for thyroid nodule classification, with the potential to improve diagnostic accuracy, reduce the need for unnecessary biopsies, and support more informed clinical decision-making in the evaluation of thyroid nodules.

Abbreviations

AUC	The area under the curve
DCA	Decision curve analysis
FNA	Fine-needle aspiration
CNN	Convolutional neural network
TL	Transfer learning

Acknowledgements
This research was supported by Science and Technology Project of Jiaxing (2023AY11045, 2022AD30123, 2022AD30118, 2019AY32014). We apologize to the many researchers whose work could not be cited due to space limitations.

Authors' contributions
Conceptualization: BM, JL, ZG, YX Methodology: BM, JL, ZG, YX Investigation: BM, JL, ZG, YX Visualization: BM, JL, ZG, YX Writing—original draft: YX Writing—review & editing: BM, JL, YX Supervision: BM, JL.

Funding
This research was supported by Science and Technology Project of Jiaxing (2023AY11045, 2022AD30123, 2022AD30118, 2019AY32014).

Data availability
The datasets used in this study are available within the manuscript and any request can be made from the corresponding author.

Declarations

Ethics approval and consent to participate
This study was conducted in accordance with Declarations of Helsinki and approved by the Medical Ethics Committee of Zhejiang Rongjun Hospital (approval number: 2023 Lun Shen Yan No. 15, 27). Informed consent to participate was waived by Dr Xingen Zhang, the director of Medical Ethics Committee of Zhejiang Rongjun Hospital (No. F-A307-014, 2023–27) because of the retrospective nature of the study and the analysis of anonymous clinical data.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Ultrasound, Zhejiang Rongjun Hospital, No.309 Shuangyuan Road, Jiaxing 314001, China. [2]Interventional Cancer Institute of Chinese Integrative Medicine, Putuo Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai 200062, China. [3]Department of Clinical Center, Jiaxing Hospital of Traditional Chinese Medicine, Jiaxing 314001, China.

Received: 5 June 2024 Accepted: 11 March 2025
Published online: 25 March 2025

References

1. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The diagnosis and management of thyroid nodules: a review. JAMA. 2018;319(9):914–24.
2. Lim H, Devesa SS, Sosa JA, Check D, Kitahara CM. Trends in thyroid cancer incidence and mortality in the United States, 1974–2013. JAMA. 2017;317(13):1338–48.
3. Floridi C, Cellina M, Buccimazza G, Arrichiello A, Sacrini A, Arrigoni F, Pompili G, Barile A, Carrafiello G. Ultrasound imaging classifications of thyroid nodules for malignancy risk stratification and clinical management: state of the art. Gland Surg. 2019;8(Suppl 3):S233.
4. Liang X, Yu J, Liao J, Chen Z. Convolutional neural network for breast and thyroid nodules diagnosis in ultrasound imaging. Biomed Res Int. 2020;2020: 1763803.
5. Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, Xin X, Qin C, Wang X, Li J, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. 2019;20(2):193–201.
6. Zhang Y. Classification and diagnosis of thyroid carcinoma using reinforcement residual network with visual attention mechanisms in ultrasound images. J Med Syst. 2019;43(11):323.
7. Wu MH, Chen KY, Shih SR, Ho MC, Tai HC, Chang KJ, Chen A, Chen CN. Multi-reader multi-case study for performance evaluation of high-risk thyroid ultrasound with computer-aided detection. Cancers. 2020;12(2):373.
8. Göreke V. A novel deep-learning-based CADx architecture for classification of thyroid nodules using ultrasound images. Interdisciplinary Sciences: Computational Life Sciences. 2023;15(3):360–73.
9. Chambara N, Ying M. The diagnostic efficiency of ultrasound computer-aided diagnosis in differentiating thyroid nodules: a systematic review and narrative synthesis. Cancers. 2019;11(11):1759.
10. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–10.
11. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54.
12. Tang F, Ding J, Wang L, Ning C. A novel distant domain transfer learning framework for thyroid image classification. Neural Process Lett. 2022;55(3):2175–91.
13. Chaouchi L, Gaceb D, Touazi F, Djani D, Yakoub A. Application of deep transfer learning in medical imaging for thyroid lesion diagnostic assistance. In: 2024 8th International Conference on Image and Signal Processing and their Applications (ISPA). 2024. p. 1–7.
14. Sureshkumar V, Balasubramaniam S, Ravi V, Arunachalam A. A hybrid optimization algorithm-based feature selection for thyroid disease classifier with rough type-2 fuzzy support vector machine. Exp Syst. 2021;39(1):1–14.
15. Sureshkumar V, Jaganathan D, Ravi V, Velleangiri V, Ravi P. A comparative study on thyroid nodule classification using transfer learning methods. Open Bioinform J. 2024;17(1):1–10.
16. Zhou J, Yin L, Wei X, Zhang S, Song Y, Luo B, Li J, Qian L, Cui L, Chen W, et al. 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. Endocrine. 2020;70(2):256–79.
17. Jia D, Wei D, Socher R, Li LJ, Kai L, Li FF: ImageNet: A large-scale hierarchical image database. In: 2009; 2009: 248–255.
18. Toro-Tobon D, Loor-Torres R, Duran M, Fan JW, Singh Ospina N, Wu Y, Brito JP. Artificial intelligence in thyroidology: a narrative review of the current applications, associated challenges, and future directions. Thyroid. 2023;33(8):903–17.
19. Koh J, Lee E, Han K, Kim EK, Son EJ, Sohn YM, Seo M, Kwon MR, Yoon JH, Lee JH, et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. Sci Rep. 2020;10(1):15245.
20. Wei X, Gao M, Yu R, Liu Z, Gu Q, Liu X, Zheng Z, Zheng X, Zhu J, Zhang S. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. Med Sci Monit. 2020;26:e926096.
21. Yoon J, Lee E, Koo JS, Yoon JH, Nam KH, Lee J, Jo YS, Moon HJ, Park VY, Kwak JY. Artificial intelligence to predict the BRAFV600E mutation in patients with thyroid cancer. PLoS One. 2020;15(11): e0242806.
22. Ni C, Feng B, Yao J, Zhou X, Shen J, Ou D, Peng C, Xu D. Value of deep learning models based on ultrasonic dynamic videos for distinguishing thyroid nodules. Front Oncol. 2023;12:1066508.
23. Chambara N, Liu SYW, Lo X, Ying M. Diagnostic performance evaluation of different TI-RADS using ultrasound computer-aided diagnosis of thyroid nodules: an experience with adjusted settings. PLoS One. 2021;16(1): e0245617.
24. Tong W-J, Wu S-H, Cheng M-Q, Huang H, Liang J-Y, Li C-Q, Guo H-L, He D-N, Liu Y-H, Xiao H, et al. Integration of artificial intelligence decision aids

to reduce workload and enhance efficiency in thyroid nodule management. JAMA Netw Open. 2023;6(5): e2313674.
25. Xu W, Jia X, Mei Z, Gu X, Lu Y, Fu CC, Zhang R, Gu Y, Chen X, Luo X, et al. Generalizability and diagnostic performance of AI models for thyroid US. Radiology. 2023;307(5):e221157.
26. Buda M, Wildman-Tobriner B, Hoang JK, Thayer D, Tessler FN, Middleton WD, Mazurowski MA. Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. Radiology. 2019;292(3):695–701.
27. Gomes Ataide EJ, Ponugoti N, Illanes A, Schenke S, Kreissl M, Friebe M. Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. Sensors (Basel, Switzerland). 2020;20(21):6110.
28. Zhao CK, Ren TT, Yin YF, Shi H, Wang HX, Zhou BY, Wang XR, Li X, Zhang YF, Liu C, et al. A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate. Thyroid. 2021;31(3):470–81.
29. Xi NM, Wang L, Yang C. Improving the diagnosis of thyroid cancer by machine learning and clinical data. Sci Rep. 2022;12(1):11143.
30. Guo YY, Li ZJ, Du C, Gong J, Liao P, Zhang JX, Shao C. Machine learning for identifying benign and malignant of thyroid tumors: a retrospective study of 2,423 patients. Front Public Health. 2022;10:960740.
31. Prochazka A, Gulati S, Holinka S, Smutek D. Patch-based classification of thyroid nodules in ultrasound images using direction independent features extracted by two-threshold binary decomposition. Comput Med Imaging Graph. 2019;71:9–18.
32. Prochazka A, Gulati S, Holinka S, Smutek D. Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition. Technol Cancer Res Treat. 2019;18: 1533033819830748.
33. Deng C, Li D, Feng M, Han D, Huang Q. The value of deep neural networks in the pathological classification of thyroid tumors. Diagnostic Pathol. 2023;18(1):95.
34. Dar SUH, Ozbey M, Catli AB, Cukur T. A transfer-learning approach for accelerated MRI using deep neural networks. Magn Reson Med. 2020;84(2):663–85.
35. Gupta P, Malhotra P, Narwariya J, Vig L, Shroff G. Transfer learning for clinical time series analysis using deep neural networks. J Healthc Inform Res. 2020;4(2):112–37.
36. Basha SHS, Vinakota SK, Pulabaigari V, Mukherjee S, Dubey SR. AutoTune: automatically tuning convolutional neural networks for improved transfer learning. Neural networks : the official journal of the International Neural Network Society. 2021;133:112–22.
37. Shao J, Zheng J, Zhang B. Deep convolutional neural networks for thyroid tumor grading using ultrasound B-mode images. The Journal of the Acoustical Society of America. 2020;148(3):1529.
38. Zhou H, Jin Y, Dai L, Zhang M, Qiu Y, Wang K, Tian J, Zheng J. Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images. Eur J Radiol. 2020;127: 108992.
39. Zhang X, Lee VCS, Rong J, Lee JC, Liu F. Deep convolutional neural networks in thyroid disease detection: a multi-classification comparison by ultrasonography and computed tomography. Comput Methods Programs Biomed. 2022;220: 106823.
40. Chen C, Jiang Y, Yao J, Lai M, Liu Y, Jiang X, Ou D, Feng B, Zhou L, Xu J, et al. Deep learning to assist composition classification and thyroid solid nodule diagnosis: a multicenter diagnostic study. Eur Radiol. 2024;34(4):2323–33.
41. Akselrod-Ballin A, Chorev M, Shoshan Y, Spiro A, Hazan A, Melamed R, Barkan E, Herzel E, Naor S, Karavani E, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology. 2019;292(2):331–42.

Publisher's Note